Detection of DDoS Attacks Using SHAP-Based Feature Reduction

C Cynthia*, Debayani Ghosh, and Gopal Krishna Kamath

***Abstract*—Machine learning techniques are widely used to protect cyberspace against malicious attacks. In this paper, we propose a machine learning-based intrusion detection system to alleviate Distributed Denial-of-Service (DDoS) attacks, which is one of the most prevalent attacks that disrupt the normal traffic of the targeted network. The model prediction is interpreted using the SHapley Additive exPlanations (SHAP) technique, which also provides the most essential features with the highest Shapley values. For the proposed model, the CICIDS2017 dataset from Kaggle is used for training the classification algorithms. The top features selected by the SHAP technique are used for training a Conditional Tabular Generative Adversarial Networks (CTGAN) for synthetic data generation. The CTGAN-generated data are then used to train prediction models such as Support Vector Classifier (SVC), Random Forest (RF), and Naïve Bayes (NB). The performance of the model is characterized using a confusion matrix. The experiment results prove that the attack detection rate is significantly improved after applying the SHAP feature selection technique.**

***Index Terms*—DDoS, SHAP, IDS, machine learning, CTGAN**

I. Introduction

The Internet is becoming increasingly pervasive since it is the primary mode for data sharing. This has led to its usage increasing rapidly. However, the sharing of files through the Internet is vulnerable to various malicious attacks. An organization or a disgruntled individual could create a system to perpetrate these cyber-attacks. Owing to such ease of perpetrating cyber-attacks, there are several cyber-attacks in existence today, including Malware, Phishing, Man-in-the-Middle attacks, Denial of Service (DoS) attacks, Distributed Denial of Service (DDoS) attacks and SQL Injection, to name a few [1, 2]. The DoS attack is exclusively meant to shut down a system or a network, thereby making it unavailable for legitimate users [3–5]. The DDoS attack is a form of DoS attack wherein different sources target a specific network/server to deprive legitimate users its access. Primarily, the DDoS attack is carried out at the application layer of a network protocol stack; however, certain types of DDoS attacks can also utilise transport layer protocols such as the Transmission Control Protocol (TCP) or the User Datagram Protocol (UDP).

Machine Learning is the burgeoning technology used for data analysis and making predictions without human intervention. It is the task of imparting knowledge and intelligence to the machine to find insightful information from the data. There are a few steps involved in training the machine before deploying it to an environment: Data collection, data pre-processing (feature reduction/feature selection), choosing a model, training the model, model evaluation, parameter tuning, and making predictions.

Machine learning techniques are used in various fields, and one among them is cyber-security. After training the model with adequate data, it can be used to detect network traffic abnormalities and prevent various cyber-attacks. For instance, to predict the spikes in network usage, the authors in [6] have used machine learning techniques. It helps to improve cyber-security by training the machine learning model with different cyber-attacks.

Researchers have identified different machine learning algorithms for classifying and predicting cyber-attacks. To start with the data pre-processing, the following steps are involved: encoding, standardization, and finally, dimensionality reduction using Principle Component Analysis (PCA) [7]. In the paper [8], the researchers have used feature transformation-based dimensionality reduction for feature clustering using a Gaussian traffic attribute pattern. After data pre-processing, the classification algorithms like Random Forest (RF), XGBoost, and Keras Sequential algorithms explain the results of any machine learning model model [9].

There are a few other machine learning algorithms such as K_Nearest_Neighbours (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR) algorithms used for the Intrusion Detection Systems (IDS) for IoT devices [10–12].

Deep learning is derived from machine learning, and it consists of different levels of algorithms which are based on very complicated neural networks that mimic the human brain. Deep Learning approaches such as Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN), and Fully Connected Layers can also be used to extract features and detect DDoS attacks [13].

In [14], hybrid detection-based methods, Stacked Auto Encoder and CNN, and feature selection, are used to detect DDoS attacks better. In [15], the authors find that the network intrusion detection systems use Sparse Auto-Encoders and Auto-Encoders modules along with GINI feature selection.

DDoS attacks in Software Defined Networks can be detected using other feature selection methods, such as Information Gain (IG) and RF to analyze the most relevant features of these attacks [16].

Manuscript received February 10, 2023; revised March 13, 2023; accepted April 8, 2023.

C Cynthia and Gopal Krishna Kamath are with the Department of Electrical and Electronics Engineering at BITS-Pilani Hyderabad Campus, Hyderabad, Telangana, India. Email: gopal.kamath@hyderabad.bits-pilani.ac.in (G.K.K.)

Debayani Ghosh is with the Department of Electronics and Communication Engineering at Thapar Institute of Engineering and Technology, Patiala, Punjab, India. Email: debayani.ghosh@thapar.edu (D.G.)

*Correspondence: p20210415@hyderabad.bits-pilani.ac.in (C.C.)

For these machine learning and deep learning models to be deployed successfully, the network traffic data must be appropriately captured, and important features are extracted. These features assist in classifying whether the network is safe or under attack. The most impactful features are sufficient in analyzing the network, thereby also resulting in feature reduction. Feature selection or feature extraction techniques can aid in identifying the most influencing attributes.

SHapley Additive exPlanations (SHAP) [17] is one of the feature selection methods which provides feature importance. It is a coalition game-theoretic method to distribute the defray among the features justly. The SHAP explanation model gives the significance of the features, which supports interpretation and accuracy for model prediction [18]. Integration of machine learning methods and SHAP gives a better prediction using fewer number of features [19]. The dimensionality reduction techniques convert the high dimensional dataset to a low dimensional dataset by preserving the structures [20]. The operationalizations of SHAP values can be studied using different axiomatic approaches, and there are various techniques such as Baseline SHAP, Integrated Gradients, and Conditional Expectation Shapley [21]. The selected features can be used to train any generative model, which will then be used to generate synthetic data.

Generative adversarial networks (GANs) are generative machine learning modeling techniques that learn from the original data to generate plausible synthetic data, with the same distribution as that of the natural data. When natural data pertaining to the requisite process is insufficient, we can use this model to generate synthetic data, which would aid in better analysis of the data. GANs operate by training both the Generator and the Discriminator, which are its main components, to generate synthetic data. Further, two different data are used for training the networks: one is real, and the other is noise. The Discriminator should learn the real data, and the Generator should process the noise. The main objective is that the Discriminator should classify the real from fake data, whereas the Generator tries to fool the discriminator [22]. This, in turn, forms a feedback loop, thereby finally obtaining the synthetic or adversarial data via the Generator. In GAN, the primary task is to identify the distribution between the real and synthetic data; if it does not overlap, it results in a vanishing gradient problem. Wasserstein Generative Adversarial Networks (WGAN) overcomes this major issue by replacing the discriminator with a critic and calculating the loss function based on the difference between real and synthetic data. This Wasserstein loss function is based on the Earth Mover's Distance. The WGAN can be used to detect the mutation of attacks which is known as a polymorphic attack [23].

In the context of network security, the data is typically in the form of tables, as opposed to traditional time-series data in signal processing fields. Conditional Tabular Generative Adversarial Networks (CTGAN) is used to generate the tabular data without any vanishing gradient problem [24]. The CTGAN can effectively process both categorical and continuous data. The statistical properties between real and synthetic data are measured quantitatively using WGAN loss with gradient penalty [25].

The following provides a summary of the crucial steps involved in our work:

1) Feature Selection - There are two conventional methods for reducing the features: Feature Extraction and Feature Selection. Feature Extraction is the process of reducing the number of features without affecting the original information. The redundant data are removed using different combinations and transformations of the original dataset. In contrast, feature selection ranks features' importance and helps discard features with least importance. It better explains how our model performs accurate predictions with those selected features. We use SHAP for feature selection because of its prominent properties like local accuracy, missingness, and consistency. It assures a fair distribution of contribution among each of the features. SHAP gives feature importance at a global level by adding the absolute value of the SHAP for each data point. This provides us with the flexibility to select features with high importance for further processing.
2) Synthetic Data Generation - Features with significant importance are selected by SHAP and given as input to Conditional Tabular Generative Adversarial Networks (CTGAN) for synthetic data generation. For better prediction of DDoS attacks, we use a reduced number of features. We can infer that the model can generate the data with the most significant features. The output of this phase is the synthetic data with the same distribution as the original data, which is very efficient in predicting the abnormality of the network. The final stage is a prediction which uses this synthetic data for classification.
3) Prediction - Our model's final phase predicts whether the network is benign or under a DDoS attack. The trained model has been saved and used for further classifications. We have used three different classifiers; namely, SVC, RF, and NB.

The remnant part of this paper has been organized as follows. Section II is an elaborate discussion of the related works. Section III elucidates the proposed methodology. In addition, the conduction of experiments with the existing CICIDS2017 dataset and the synthetic data generated by CTGAN. The later part includes the performance comparison with the current baselines in Section IV. Finally, the broad conclusion gives the proper scope for future research in Section V.

II. Related Work

In this section, we briefly review pertinent research findings on different cyber-attacks in which the researchers used various machine learning algorithms.

Ismail *et al.* [2] used the Australian Centre for Cyber Security (ACCS) dataset containing DDoS attacks features. Therein, they used machine learning approaches with data pre-processing, including standardization and normalization. The researchers used RF and XGBoost Classifier algorithms and the evaluation of the result using a confusion matrix.

Parvinder Singh Saini *et al.* [6] focus on the detection of DDoS attacks by the machine learning tool WEKA. Validation of this approach on various types of attacks like

Normal, UDP-Flood, Smurf, SID DOS, and HTTP-Flood DDoS attacks. The J48 classifier is shown to outperform other classifying algorithms like MLP, RF, and NB. Accuracy is the metric for identifying the best performance of different algorithms.

Fryer *et al*. [17] explain the SHapley Additive exPlanations (SHAP) and SHapley Additive Global importance (SAGE). They elaborate on the feature selection using SHAP. Wilson E.Marcilio-Jr and Danilo M.Eler. [20] proposed a method for dimensionality reduction using SHAP.

Lundberg and Lee [18] proposed a unified approach for interpreting the prediction for complex models using SHAP. The model has improved the computational performance and gives better consistency. Benedek Rozem-berczki *et al*. [19] designed a framework for the explainable machine-learning model using SHAP.

Alenezi and Ludwig [9] proposed different classification models to detect various types of cyber-attacks by using ensemble techniques. Researchers worked on two other cyber-security datasets: malicious URLs and Android malware. Explainable AI uses SHAP to identify the most significant features. TreeShap, KernelShap, and DeepShap are three SHAP methods used for extracting better feature contributions. The Machine Learning algorithms used for further analysis are Random Forest Classifier (RFC), XGBoost Classification, and the Keras Sequential algorithm. The criteria used by RFC is Gini which measures the quality of a split. The Random Forest with the TreeShap Explanation and XGBoost classification with the KernelShap Explanation classifies the attacks from the normal traffic.

Sambangi *et al*. [8] use feature transformation based dimensionality reduction to detect low- and high-rate network attacks. They proposed a model to find the traffic similarity function between two networks to classify and detect the attacks.

Chen *et al*. [22] proposed an attack-agnostic defense model against poisoning attacks. A poisoning attack is an attack on the training data that leads to misclassification. The learning of data augmentation on the models and the predictions are shown to be better using GAN. This paper proposed an architecture that constitutes Synthetic Data Generation, Mimic Model Construction, and Poisoned Data Recognition. The training of the model using De-Pois to detect four different attacks; the attacks are Targeted Clean Label Poisoning Attack (TCL- Attack), Poisoning Attack with GAN (PGAN-attack), Label Flipping Attack (LF-attack), and Regression Attack (R-attack).

Chauhan *et al*. [23] proposed a model for detecting polymorphic attacks. Polymorphic malware is an attack that consistently changes its identifiable feature to evade the detection system. Since the attack feature profile continuously changes, the defense system should adopt Incremental Learning. This paper focuses on generating DDoS attacks using adversarial network techniques such as WGAN. SHAP identifies the function feature of an attack. There are three IDS techniques to identify the various attacks: Signature-Based Detection (Knowledge-Based), Anomaly Based Detection (Behaviour Based), and Stateful Protocol Analysis (Specification Based). The classifiers for the Signature - Based detection systems are SVM, NB, KNN, DT, and RF.

Xu *et al*. [24] designed a model for tabular data using Conditional GAN, which outperforms the Bayesian methods on most real datasets. Zilong Zhao *et al*. [25] have designed a model for encoding mixed categorical and continuous variables along with the missing values. The metrics used for measuring statistical similarity between the real and synthetic data are Jensen-Shannon divergence, Wasserstein distance, and Difference in pair-wise correlation.

Meenakshi *et al*. [13] proposed a model for DDoS attack detection, and the researchers used deep learning approaches like Recurrent Neural Networks and Convolutional Neural Networks.

Sudugala *et al*. [12] experimented on three different datasets: CIS-DoS, CISIDS2017, and CSE-CIC-IDS2018. The authors used four machine learning algorithms: SVM, NB, Simple Linear Regression, and DT to detect DDoS attacks.

III. Proposed Framework

In this section, we provide a detailed description of our proposed model. The steps followed to detect a DDoS attack are shown in Fig. 1. Captured network traffic data contains several features such as time stamp, IP address of the source and IP address of the destination, among others. These features can be used to identify abnormality in the network. DDoS attacks are among the cyber-attacks that make the server busy with the bombardment of simultaneous fake requests. The proposed model detects DDoS attacks with a reduced number of features.

Our work proposes a DDoS attack detection framework that includes SHAP for Feature Selection and CTGAN model for synthetic data generation. The SHAP provides the most significant features that help the prediction. The remaining features are discarded. These features are used to train a CTGAN, which will then generate synthetic data. This synthetic data is used to obtain better performance for predictions of DDoS attacks via feedback of CTGAN. Finally, the classifiers are trained with the labeled original data, and then the model can be used for testing the synthetic data. Thus, the prediction accuracy has been improved with the reduced features. The following sub-sections provide comprehensive explanations of each block in our proposed model.

Fig. 1. Proposed model.

A. SHAP for Feature Selection

Shapley value-based explanations of machine learning models are used for finding the contribution of each feature to the model prediction and providing the feature importance for the entire feature in the dataset.

Machine Learning and Game theory work together in such a way that the model's input features are similar to players in a game while the model function is similar to the rules of the game. Transferable Utility is one of the assumptions in many cooperative games, where the individual players do not

receive any payoff. Instead, the coalitions get standard payoffs. We integrate the other features using conditional expected value formulation to evaluate an existing model when only a subset of features is a part of the model.

The partial dependence plot is a global method showing the marginal effect of one or two features on the predicted outcome. It portrays the relationship between a target and a feature and helps verify whether they are dependent linearly. One of the main properties of SHAP values is that they sum up the difference between the game outcome in presence of all features and the game outcome when no features are present. In our machine learning model, this means that SHAP values of all the input features will always sum up the difference between the expected output and the predicted output of the current model.

SHAP assigns an importance value to each feature that represents the effect on the model prediction of including that feature. The model has been trained with the presence $f_{s\cup i}$ and the absence of that feature f_s. Predictions from the two models are compared on the current input $f_{s\cup i}(x_{s\cup i}) - f_s x_s$, where x_s is the values of input features in the set S. The additive feature attribution methods are explained in the later part of this section [18].

Let f be the original prediction model and g be the explanation model. The prediction $f(x)$ based on every single input x focuses on local methods.

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i', \tag{1}$$

where $x' \in 0, 1^M, M$ is the number of input features, and $\phi_i \in \mathbb{R}$.

Local Accuracy is given by Lundberg as in (1). The explanation model $g(x')$ matches the original model $f(x)$ when $x = h_x(x')$, where $?_0 = f(h_x(0))$ represents the model output.

$$x_i' = 0,\ \phi_i = 0. \tag{2}$$

Missingness constraints are represented as in (2), where $x_i' = 0$ has no attributed impact.

Furthermore, the next property is consistency. Let $f_x(z') = f\big(h_x(z')\big)$ and $z'?i$ denote setting the i^{th} coordinate to zero, $z_i' = 0$. For any two models f and f', if $f_x'(z') - f_x'(z' \setminus i) \geq f_x(z') - f_x(z'?i)$ for all inputs $z' \in \{0,1\}^M$, then $?_i(f_i'x) = ?_i(f_ix)$.

The only possible explanation model g that satisfies all three properties (missingness, consistency, and local accuracy), is given by [18] as represented in (3).

Algorithm 1 Feature Selection

INPUT : Labelled dataset by x with i features.
Consider a cooperative game with M features (numbered from 1 to M), and let F be the set of features. Assume $S \subseteq F$ represents a coalition.
FEATURE SELECTION : Set of Features F ={1,....,M}
COALITION, $S \subseteq F$, where F is grand coalition
Train and Test Feature Vector Sets are X_s^{train} *and* X_s^{test} where $X_s^{train} = \{x_i^{train} | i \in S\}$ and $X_s^{test} = \{x_i^{test} | i \in S\}$
Assumptions: Let P be the permutation and C be the characteristic function, $C : 2^F \rightarrow \mathbb{R}$

```
for i = {1,....,n}
    Perform permutations for all the features P(i)
    for each P(i) compute coalition φ_i
        φ_i = (1/F!) Σ_P (v(s ∪ i) − v(s))
        where v(s) is function matches every coalition
    s
        if Σ_{i∈F} φ_i = C(F) then
            C(F) is lossless among the features
        elseif φ_i = 0 then
            Feature i has no contribution to the model
        elseif φ_i1 = φ_i2 then
            Two features has equal role on the model
        elseif φ_i1(C,K) = φ_i1(C) + φ_i1(K) then
            Two subgames (C,i_1),(K,i_1) have equal
    contributions with i_1
        end if
    end for
end for
```

$$\phi_i(f, x) = \sum_{z' \subset x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]. \tag{3}$$

B. Synthetic Data Generation Using CTGAN

Conditional Tabular GAN is used for synthetic data generation with the selected number of features. The architecture of the model is shown in Fig. 2. CTGAN is a GAN-based method to model tabular data distribution and mode-specific normalization to account for the non-Gaussian and multimodal nature of distribution. Additional information, such as class labels, can be added for training the model. CTGANs model is trained by a zero-sum minimax game, where the critic (known as discriminator for regular GANs) tries to maximize the objective, and the generator tries to minimize it. The generator in a CTGAN is trained with a vector sampled from a standard multivariate normal distribution. The critic eventually obtains a deterministic transformation that maps the standard multivariate normal distribution to the data distribution.

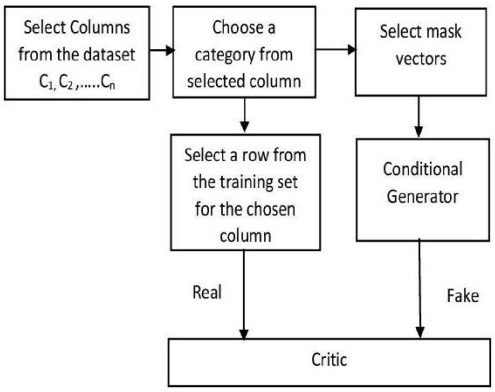

Fig. 2. Architecture of CTGAN.

The generator learns the re-sampled distribution, which is different from the real data distribution. The main goal is to

re-sample efficiently such that all categories from discrete attributes are evenly sampled during the training process and to recover the real data during the test. Conditional GAN has been widely used to generate a particular class of data. The critic maximizes the objective function, whereas the generator minimizes the objection function.

C. Predictive Classifiers

The prediction phase uses the synthetic data with selected features and different machine learning algorithms to classify benign and DDoS attacks. SVC is one of the most efficient supervised learning algorithms for classification and regression problems. The main objective of this algorithm is to create the best-fit line that divides x-dimensional spaces into different classes. This best fit is also known as a hyperplane. The number of dimensions depends on the total number of features in the dataset. For linear SVM, one straight line is sufficient to divide the space into classes, whereas for non-linear data, new dimensions have to be added to the existing one. The data points closer to hyperplanes influence the plane's orientation.

RF is a classifier that combines multiple decision trees to create and classify the data. Merging numerous decision trees gives a better prediction. The classification type prefers voting where maximum voting wins. This classifier provides better regression accuracy and considers the average of all the outputs. RF is a type of bootstrap aggregation. It is an ensemble machine learning algorithm, which is also known as the bagging method. Bootstrap aggregation is the best method to reduce the variance, which causes the model to overfit.

NB is a classifier based on Bayes' Theorem. From the given Bayes' equation (4), A is referred to as a hypothesis and B as evidence. It is named as Na ïve because the presence of one particular feature does not affect the other.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{4}$$

Algorithm 2 Synthetic Data Generation CTGAN

INPUT : Labelled data with SHAP selected features

Assumptions : D_{train} for Training Data, *G* for Conditional Generator, *C* for Critic, *M* for Mask Vector, *Condnj* for Condition Vector

Critic - Maximize the objective function to classify real and fake.

Generator - Minimize the objective function to fool the Critic.

STEP 1: Generate mask vector $m_1, m_2, \ldots, m_n$, for $1 \le j \le n$

STEP 2: Create a condition vector *Condnj*, for $1 \le j \le m$

STEP 3: Generate fake data $\hat{x}_j = Z_j | Condnj, \forall 1 \le j \le m$

STEP 4: Extract real data sample

$x_j \sim uniform(D_{train}|Condnj)$, for $1 \le j \le m$

STEP 5: Cost Function = $\frac{1}{m}\sum_{j=1}^{m} C\left(x_j\middle|Condnj\right) - \frac{1}{m}\sum_{j=1}^{m} C\left(\hat{x}_j\middle|Condnj\right)$

STEP 6: Gradient Penalty = $\frac{1}{m}\sum_{j=1}^{m} \lambda(\| \nabla_C\left(\hat{x}_j\middle|Condnj\right) \|_2 - 1)^2$

IV. Experimental Results

In our proposed method, we used the CICIDS2017 Kaggle dataset. This dataset captures the network traffic fluctuations for five days a week. The dataset comprises the following attacks: DoS, BruteForce, Web Attack, Bot, Portscan, Infiltration, and DDoS. We have used the data generated for Friday afternoon working hours in which 97,718 are benign, and 1,28,027 are DDoS. There is a total of 79 features captured in this dataset.

There are four significant stages in our method as shown in Fig. 3.

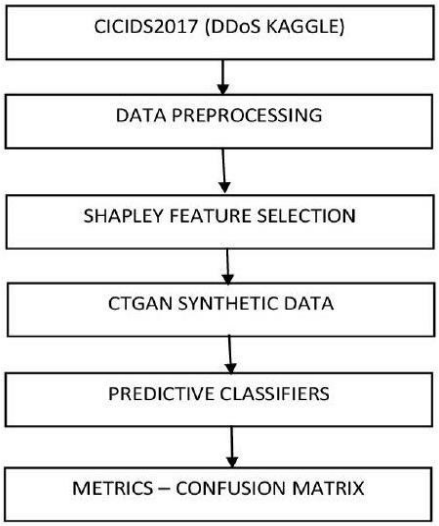

Fig. 3. Proposed workflow.

Step 1: Data Preprocessing is the process of removing the duplicated data and null spaces. The dataset should be balanced for the training to not overfit the model. The label encoding is to code benign as 0 and DDoS as 1.

Step 2: After data cleaning, the most important features must be considered for synthetic data generation. The feature selection has been made using SHAP, where the model has been given with a train-test spit of $80\% - 20\%$, respectively. SHAP performs permutations to find the marginal contributions of each feature.

The SHAP summary plot combines the feature importance and its effects on the model prediction. The SHAP values explain each feature's contribution toward the model's predicted output.

The feature importance measures are classified as local feature importance and global feature importance. Individual feature contribution to the predicted output for a specific data point is known as Local feature importance. In contrast, the average SHAP value for each feature across entire data points is given by Global feature importance. In the plot, the vertical line indicates global importance, and the horizontal line represents local importance.

The CICIDS 2017 dataset contains 60 features, of which the SHAP selects the 20 most impactful features. As shown in Fig. 4, features towards the top end (represented in darker shades of red) contribute more towards the output, while the features towards the bottom end (represented by darker shades of blue) contribute less. Hence, for the CICIDS 2017 dataset that we make use of, feature 37 (mean packet length) is the most significant, whereas feature 5 (total length of the forward packet) has the least significance.

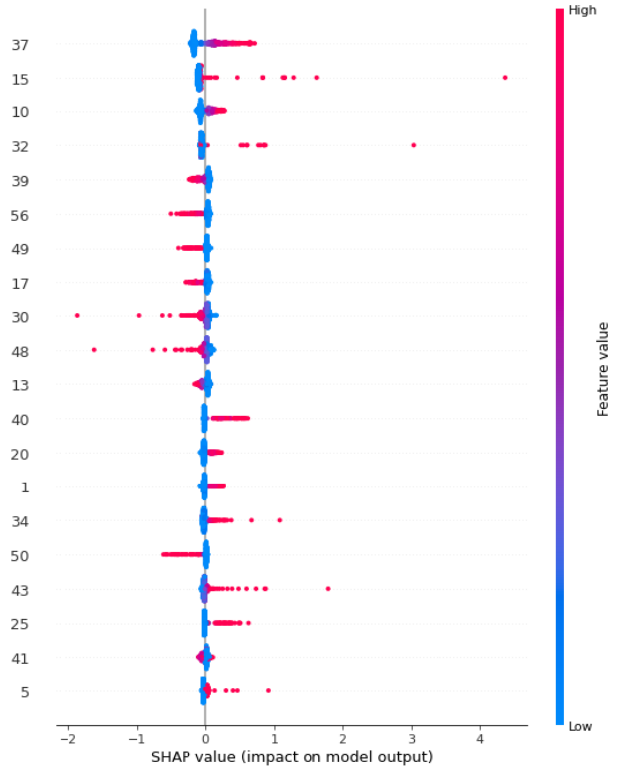

Fig. 4. SHAP feature Importance.

Step 3: The dataset with reduced features is given for CTGAN to generate synthetic data. CTGAN has been used to create the tabular data for further analysis. There are no mode collapse and vanishing gradient issues in CTGAN. The synthetic data generation by CTGAN before and after feature selection is computed and the results are compared in the Table I.

TABLE I: COMPARISON OF ERROR FOR DIFFERENT MODELS

MODEL	PREDICTION (RMSE)
100 PERCENT FEATURE UTILIZATION (no Shapley)	0.07260
50 PERCENT FEATURE UTILIZATION (Shapley)	0.04321
75 PERCENT FEATURE UTILIZATION (Shapley)	0.0484

Step 4: The SHAP selected feature dataset has been used for training the predictive classifiers such as the SVC, RF, and NB. The confusion matrix gives the performance metric for all the classifiers.

The Confusion Matrix is the map between actual and predicted values with the following evaluation parameters: Accuracy, F1 Score, Precision, and Recall. The abbreviations used to define these quantities, along with their definitions, are provided in Table II.

Table III, Table IV, Table V, and Table VI portray the various performance metrics for our proposed model. Accuracy, F1 Score, Precision, and Recall are computed with SHAP (using reduced number of features) and without SHAP (using all 79 features) datasets. The Accuracy measure before and after applying SHAP is 99.9% for Random Forest Classifier. Naïve Bayes performs well after feature reduction using SHAP, where the measure is 98%.

TABLE II: LIST OF USED ABBREVIATIONS AND IT'S DEFINITION

SYMBOL	DEFINITION
True Positive (TP)	It is the number of true DDoS attacks.
True Negative (TN)	It is the number of true legitimate traffic: the benign is recognized as legitimate.
False Positive (FP)	It is the number of false legitimate traffic: the benign is misidentified as attacks
False Negative (FN)	It is the number of DDoS attacks that cannot be recognized as an attack

Accuracy is the ratio of correctly classified values to the total number of values. It captures the model's overall performance, and is given by (5).

$$ACCURACY = \frac{TN + TP}{TN + FP + TP + FN} \quad (5)$$

Precision is the ratio of true positives to the total number of values identified as positive, and is written as (6).

$$PRECISION = \frac{TP}{TP + FP}. \quad (6)$$

Recall is the ratio of correctly classified values to the total number of elements that belong to the positive class. It is also referred to as Detection Rate. True Positive Rate is the measure of true samples which are identified correctly, and is also referred to as Sensitivity. Mathematically, this is captured as (7).

$$RECALL = \frac{TP}{TP + FN} \quad (7)$$

And lastly, F1 Score is the harmonic mean of precision and recall as given by (8).

$$F1\_SCORE = 2 * \frac{PRECISION * RECALL}{PRECISION + RECALL} \quad (8)$$

TABLE III: PREDICTIVE CLASSIFIERS COMPARISON BETWEEN WITH AND WITHOUT SHAPLEY FOR ACCURACY

CLASSIFIER	ACCURACY		
	WITHOUT SHAP	WITH SHAP (50 percent)	WITH SHAP (75 percent)
SVC	0.9933	0.9884	0.9947
RF	0.999	0.9996	0.9997
NB	0.9474	0.9759	0.9803

TABLE IV: PREDICTIVE CLASSIFIERS COMPARISON BETWEEN WITH AND WITHOUT SHAPLEY FOR F1-SCORE

CLASSIFIER	F1 SCORE		
	WITHOUT SHAP	WITH SHAP (50 percent)	WITH SHAP (75 percent)
SVC	0.9941	0.9899	0.9953
RF	0.9991	0.9996	0.9997
NB	0.9555	0.9791	0.9829

TABLE V: PREDICTIVE CLASSIFIERS COMPARISON BETWEEN WITH AND WITHOUT SHAPLEY FOR PRECISION

CLASSIFIER	PRECISION		
	WITHOUT SHAP	WITH SHAP (50 percent)	WITH SHAP (75 percent)
SVC	0.9904	0.9818	0.9919
RF	0.9998	1.0	1.0
NB	0.9161	0.9598	0.967

TABLE VI: PREDICTIVE CLASSIFIERS COMPARISON BETWEEN WITH AND WITHOUT SHAPLEY FOR RECALL

CLASSIFIER	RECALL		
	WITHOUT SHAP	WITH SHAP (50 percent)	WITH SHAP (75 percent)
SVC	0.9978	0.99803	0.9986
RF	0.9984	0.9992	0.9994
NB	0.9984	0.9992	0.9992

TABLE VII: PERFORMANCE COMPARISON OF PROPOSED MODEL WITH OTHER LITERATURES

Reference Papers	Feature Selection	Classifier	Accuracy
1. A Feature Similarity Machine Learning Model for DDoS Attack Detection in Modern Network Environments for Industry 4.0 [8]	Swathi	Naïve Bayes	0.9091
2. Distributed Denial of Service Attack Detection using Deep Learning Approaches [13]	CNN and RNN	Stacked LSTM and CNN	0.9955 and 0.96
3.A Comparison of Various Machine Learning Algorithms in a Distributed Denial of Service Intrusion [11]	max,min,mean and std deviation	SVM Linear, RF, NB	0.6325, 0.9997, 0.9977
4. A Deep Learning Approach for DDoS Attack Detection Using Supervised Learning [14]	Knowledge Graph	Stacked Auto Encoder (SAE) and CNN	0.9997
5. A Flow Based Anomaly Detection Approach with Feature Selection Method Against DDoS Attacks in SDN [16]	RFR and IG	SDN	0.9995
6. Proposed Model	SHAP	SVM, RFC, NB	0.9947,0.9997,0.9803

We now provide results of related works. In [8], the feature transformation-based dimensionality reduction is performed with their novel approach, named Swasthi, and with the classifier Naïve Bayes, the accuracy percentage is 90.91%. In addition, [13] uses RNN and CNN models with an accuracy measure of 99.55% and 96%, respectively. Furthermore, the authors in [11] use a few feature selection methods using min, max, mean, and standard deviation with the predictive classifiers SVM Linear, Random Forest, and Naïve Bayes, resulting in accuracy percentages 63.25%, 99.97%, and 99.77%, respectively. In [14], authors compute a knowledge graph for better feature extraction, and the classifier integrates Stacked Auto-Encoder and CNN, which gives a better accuracy measure of 99.97%. The Random Forest Regressor and Information Gain used in [16] to extract the best features from the existing dataset provides an accuracy of 99.95%. Table VII summarizes the comparison of the result of our proposed model with similar works in the literature.

The proposed model demonstrates a significant improvement in accuracy, as indicated by the experimental results. The following lines compare the state of the art of proposed model with other references' state of the art:

1) Compared to the results presented in [8], our proposed model shows a substantial increase in accuracy by 10%.
2) The performance of our proposed model surpasses that of the CNN model in [13] by 4%.
3) In comparison to the results presented in [11], our proposed model exhibits comparable accuracy percentages, as shown in Table V for the RF classifier.
4) Our proposed model also shows similar accuracy percentage as that presented in [14], based on the results presented in the table.
5) The method presented in [16] shows a negligible improvement in accuracy by 0.02%, in comparison to our proposed model.

Overall, the results suggest that our proposed model is quite effective and performs better than some existing models while being comparable to others.

Furthermore, the implementation of SHAP feature selection in the proposed model has resulted in a noteworthy reduction in computational cost. Specifically, the training time for the CTGAN model using all features was observed to be 11.06 s/iterations, while the execution time was reduced to 5.9 s/iterations by utilizing the SHAP-selected features. These results indicate that the proposed DDoS attack detection framework is efficient and performs well with a reduced set of features obtained through SHAP feature selection.

V. CONCLUSION

In this paper, we have proposed a framework to detect DDoS attacks with a reduced set of features. We have selected the CICIDS2017 dataset from the Kaggle repository, which contains information about DDoS attacks. In the proposed detection model, we have used the SHAP feature selection technique to reduce the number of features to work with. This reduces the computational cost and improves the accuracy. A feature's contribution towards model prediction and its importance are identified using the SHAP values. Synthetic data with the selected (reduced) features are generated using a CTGAN. Predictive classifiers are then used to detect the DDoS attack from the synthetic data. Our model has focused on the pattern of the specific attack, which is otherwise known as signature-based detection. This work can be extended in two ways: (i) To train the model for different cyber-attacks and (ii) To build anomaly-based intrusion detection systems. It is imperative to develop new techniques for better detection systems and thus also protect against any malicious attacks.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Ms. C Cynthia formulated the problem, proposed the methodology presented in this manuscript and implemented the proposed methodology via coding. Dr. Debayani Ghosh and Dr. Gopal Krishna Kamath helped formalize the problem and helped Ms. C Cynthia, in equal measure, solve the problem via the use of SHAP technique. All authors were involved in writing and revision of the manuscript, with Ms. C Cynthia taking the lead by preparing the preliminary draft and generating all the figures/tables presented in this manuscript. All authors have approved the final version of the manuscript.

FUNDING

This research is funded, in part, by BITS-Pilani Hyderabad Campus under the RIG scheme with grant number BITS/GAU/RIG/2022/H0753, and, in part, by Thapar Institute of Engineering and Technology under the grant number TU/DORSP/07/02/2022. C Cynthia's doctoral research is supported by BITS-Pilani Hyderabad Campus under the "Institute Fellowship" scheme.

REFERENCES

[1] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, and M. Xu, "A survey on machine learning techniques for cyber security in the last decade," *IEEE Access*, vol. 8, pp. 222310–222354, 2020. doi: 10.1109/ACCESS.2020.3041951

[2] Ismail, M. I. Mohamand, H. Hussain, A. A. Khan, U. Ullah, M. Zakarya, A. Ahmed, M. Raza, I. U. Rahman, and M. Haleem, "A machine learning-based classification and prediction technique for DDoS attacks," *IEEE Access*, vol. 10, pp. 21443–21454, 2022. doi: 10.1109/ACCESS.2022.3152577

[3] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: an overview from machine learning perspective," *J Big Data*, vol. 7, no. 1, Dec. 2020. doi: 10.1186/s40537020-00318-5

[4] M. Wazid, A. K. Das, V. Chamola, and Y. Park, "Uniting cyber security and machine learning: Advantages, challenges and future research," *ICT Express*, vol. 8, no. 3. pp. 313-321, Sep. 01, 2022. doi: 10.1016/j.icte.2022.04.007

[5] B. B. Gupta, R. C. Joshi, and M. Misra, "Prediction of number of zombies in a DDoS attack using polynomial regression model," *Journal of Advances in Information Technology*, vol. 2, no. 1, Feb. 2011. doi: 10.4304/jait.2.1.57-62

[6] P. S. Saini, S. Behal, and S. Bhatia, "Detection of DDoS attacks using machine learning algorithms," presented at 7th International Conference on Computing for Sustainable Global Development (INDIACom), 12-14 March 2020, New Delhi, India.

[7] S. A. Abbas and M. S. Almhanna, "Distributed denial of service attacks detection system by machine learning based on dimensionality reduction," *Journal of Physics: Conference Series*, Mar. 2021, vol. 1804, no. 1. doi: 10.1088/1742-6596/1804/1/012136

[8] S. Sambangi, L. Gondi, and S. Aljawarneh, "A feature similarity machine learning model for DDoS attack detection in modern network environments for Industry 4.0," *Computers and Electrical Engineering*, vol. 100, May 2022. doi: 10.1016/j.compeleceng.2022.107955

[9] R. Alenezi and S. A. Ludwig, "Explainability of cybersecurity threats data using SHAP," in *Proc. 2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2021, pp. 01–10.

[10] R. J. Alzahrani and A. Alzahrani, "Security analysis of ddos attacks using machine learning algorithms in networks traffic," *Electronics (Switzerland)*, vol. 10, no. 23, Dec. 2021. doi: 10.3390/electronics10232919

[11] S. H. Kok, A. Abdullah, M. Supramaniam, T. R. Pillai, I. Abaker, and T. Hashem, "A comparison of various machine learning algorithms in a distributed denial of service intrusion," 2019.

[12] A. U. Sudugala, W. H. Chanuka, A. M. N. Eshan, U. C. S. Bandara, and K. Y. Abeywardena, "WANHEDA: A machine learning based DDoS detection system," in *Proc. ICAC 2020 - 2nd International Conference on Advancements in Computing*, Dec. 2020, pp. 380–385. doi: 10.1109/ICAC51239.2020.9357130

[13] Meenakshi, K. Kumar, and S. Behal, "Distributed denial of service attack detection using deep learning approaches," in *Proc. 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, 2021, pp. 491–495.

[14] H. Tekleselassie, "A deep learning approach for DDoS attack detection using supervised learning," in *Proc. MATEC Web of Conferences*, vol. 348, p. 01012, 2021. doi: 10.1051/matecconf/202134801012

[15] C. Zhang, Y. Chen, Y. Meng, F. Ruan, R. Chen, Y. Li, and Y. Yang, "A novel framework design of network intrusion detection based on machine learning techniques," *Security and Communication Networks*, vol. 2021, Article ID 6610675. https://doi.org/10.1155/2021/6610675

[16] M. S. El Sayed, N.-A. Le-Khac, M. A. Azer, and A. D. Jurcut, "A flow based anomaly detection approach with feature selection method against ddos attacks in sdns," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 4, pp. 1862–1880, 2022.

[17] D. Fryer, I. Strümke, and H. Nguyen, "Shapley values for feature selection: The good, the bad, and the axioms," Feb. 2021.

[18] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 2017.

[19] B. Rozemberczki, L. Watson, P. Bayer, H.-T. Yang, O. Kiss, S. Nilsson, and R. Sarkar, "The shapley value in machine learning," arXiv preprint arXiv:2202.05594, 2022.

[20] W. E. M. Júnior and D. M. Eler, "Explaining dimensionality reduction results using Shapley values," Mar. 2021.

[21] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *Proc. International Conference on Machine Learning*, 92699278, 21 Nov. 2020.

[22] J. Chen, X. Zhang, R. Zhang, C. Wang, and L. Liu, "De-Pois: An Attackagnostic defense against data poisoning attacks," *IEEE Transactions on Information Forensics and Security*, 2021. https://doi.org/10.48550/arXiv.2105.03592

[23] R. Chauhan, U. Sabeel, A. Izaddoost, and S. Shah Heydari, "Polymorphic adversarial cyberattacks using WGAN," *Journal of Cybersecurity and Privacy*, vol. 1, no. 4, pp. 767–792, Dec. 2021.

[24] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," *Advances in Neural Information Processing Systems*, 2019.

[25] Z. Zhao, A. Kunar, H. van der Scheer, R. Birke, and L. Y. Chen, "CTAB-GAN: Effective table data synthesizing," in *Proc. Asian Conference on Machine Learning*, pp. 97–112, 2021.

Copyright © 2023 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).