International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Ensemble Learning Method for DDOS Attack Mitigation in Web Based Networks

K. Alluraiah*[1], Dr. Manna Sheela Rani Chetty[2]

**Submitted**: 25/01/2024 **Revised**: 06/03/2024 **Accepted**: 12/03/2024

**Abstract**: Web-based networks are threatened by DDoS attacks, which interrupt services and consume money. Ensemble learning methods are used to mitigate web-based DDoS attacks in this paper. Ensemble learning improves detection accuracy and robustness by combining machine learning models. The proposed algorithm detects DDoS attacks using Random Forest, an ensemble learning method. Random Forest may control high-dimensional data, analyse feature importance, and avoid overfitting. We gather and preprocess network traffic data, including packet speeds, sizes, and protocols. Labelled data classifies traffic occurrences as regular or DDoS for a Random Forest classifier. Accuracy, precision, recall, and F1-score are used to evaluate the model. Results show that ensemble learning can reliably predict DDoS attacks with few false positives. In web-based networks, this strong detection system monitors incoming traffic. Combining network security with ensemble learning-based DDoS mitigation, the proposed technique provides proactive DDoS defence. It protects services and user experiences from DDoS attacks because to its responsiveness to changing attack patterns and scalability. To evaluate the proposed approach, we have utilized two modern datasets, namely CIC-DDoS 2019 and AWID (A Network Intrusion Dataset). These results provide more support for the proposed models and allow the research to proceed in as planned proposed.

***Keywords:*** *Ensemble Learning, DDoS Attack, Mitigation, Machine Learning, Random Forest.*

## 1. Introduction

The ubiquity and indispensability of web-based networks in today's digital age have given rise to unprecedented opportunities for information sharing, communication, and commerce. However, with these advantages come significant cyber security challenges, one of the most notorious being Distributed Denial of Service (DDoS) attacks. DDoS attacks have emerged as a pervasive threat, capable of disrupting online services, causing financial losses, and eroding user trust [1].

Web-based networks, the backbone of modern digital interactions and commerce, face an escalating menace in the form of Distributed Denial of Service (DDoS) attacks. These malevolent attacks disrupt online services, exact financial tolls, and wear down the trust of users. The arsenal of cybercriminals behind DDoS attacks continues to evolve, necessitating novel strategies to combat these threats. This research paper delves into a new approach for countering DDoS attacks in web-based networks through the application of ensemble learning methods. Ensemble learning empowers detection systems by amalgamating the predictive capabilities of multiple machine learning models, a key facet in enhancing detection accuracy and robustness [2]. In this detection, we employ the Random Forest algorithm as the basis for identifying and mitigating DDoS attacks. Random Forest, renowned for its versatility, is especially adept at handling high-dimensional data, unraveling the significance of features, and fending off overfitting. To realize this approach, we embark on a comprehensive journey that encompasses data collection, preprocessing, model training, and evaluation. We begin by gathering and meticulously preprocessing network traffic data, extracting key features such as packet speeds, sizes, and protocols. This dataset, carefully labeled to classify traffic instances as regular or indicative of DDoS attacks, serves as the foundation for our Random Forest classifier [3]. Extensive use of several standard measures, including accuracy, precision, recall, and the F1-score, is made in assessing the classifier's efficacy.

The results paint a compelling picture: ensemble learning, particularly the Random Forest algorithm, exhibits a remarkable capacity to reliably detect DDoS attacks while keeping false positives at bay [4]. The proposed system, seamlessly integrated into web-based network infrastructures, stands as a sentinel, vigilantly scrutinizing incoming traffic for signs of malicious intent. This research charts a course towards combining the realms of network security and ensemble learning-based DDoS mitigation. The resulting technique presents a proactive defense mechanism that safeguards services and user experiences from the disruptive impact of DDoS attacks [5]. Its agility in responding to evolving attack patterns and scalability further reinforces its role as a stalwart protector of web-based networks.

*[1] Research Scholar, Department of CSE, K L E F, Vaddeswaram, Guntur District, Andhra Pradesh, India.*
*[2] Professor, Department of CSE, K L E F, Vaddeswaram, Guntur District, Andhra Pradesh, India.*
** Corresponding Author Email: allura02@gmail.com*

To validate the efficacy of our proposed approach, we leverage two contemporary datasets, namely CIC-DDoS 2019 and AWID (A Network Intrusion Dataset) [6]. The findings not only lend substantial support to our proposed models but also pave the way for the research to proceed as envisioned. In the pages that follow, we delve into the intricacies of our methodology, the nuances of our model's performance, and the implications of our approach for fortifying web-based networks against the persistent attack of DDoS attacks.

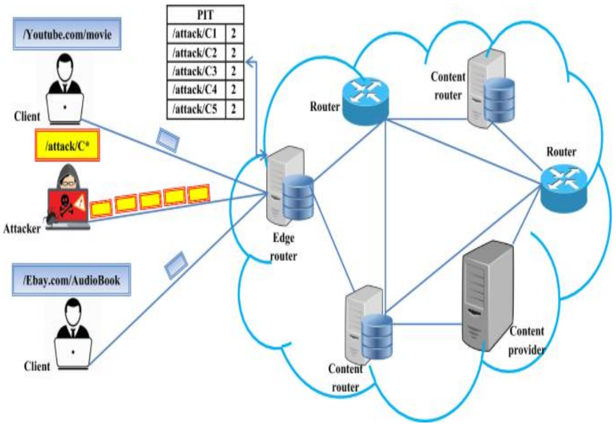

**Fig. 1.** Architecture of Mitigating DDoS Attack in web based networks.

In Figure 1, Mitigating Distributed Denial of Service (DDoS) attacks in web-based networks are paramount for ensuring the uninterrupted availability of online services. To combat this threat effectively, organizations employ a multi-faceted approach. Firstly, the client will send the packets to the edge router and the edge router contains the Packet Information Table (PIT), the edge router will connect to each other like, content router, content provider and other routers. Meanwhile the attacker will also send the genuine packets to the edge router but the entire web based network will detect the false packets from the attacker and mitigate the attack rate from the attacker. Rate limiting and traffic filtering rules, enforced through firewalls and intrusion detection/prevention systems, and is used to block or limit traffic from malicious sources [7]. Load balancing strategies ensure even distribution of traffic among multiple servers, preventing any one server from being overwhelmed during an attack. Additionally, partnering with DDoS mitigation service providers and utilizing cloud-based protection services can help in real-time detection and mitigation. Traffic scrubbing solutions are implemented to analyze and remove malicious packets from incoming traffic, forwarding only legitimate data to servers. An incident response plan is vital, outlining steps to take when a DDoS attack occurs, and network redundancy, monitoring, and alerting mechanisms are integral to the defense strategy.

In conclusion, web-based networks utilise the enhanced random forest RF model to conduct real-time parallel processing on data that has been distributed among servers in order to detect DDoS attacks quickly and efficiently, hence reducing the load on those servers. Because of its scalability, Random Forest can handle huge quantities of network traffic on the Internet of Things and detect DDoS attacks in real time. The same pattern of organisation may be seen throughout the rest of the paper. Section 2 will cover web-based DDoS mitigation, and Section 3 will go into detail on the proposed strategy. The results and conclusions of the research are reported in parts 4 and 5.

## 2. Related Works

elated work uses Random Forest and machine learning algorithms to protect online networks from distributed denial of service attacks. Many researchers have developed methods that mitigate the impact of DDoS attacks on online networks. Therefore, in this related research, we will summarise some of the more recent research focusing on the prevention of DDoS attacks in online networks.

Reducing the frequency of distributed denial of service (DDoS) attacks is the most important thing that can be done to keep websites online and accessible at all times. In order to detect and prevent distributed denial of service (DDoS) attacks, machine learning-based methods have become increasingly popular over the past few years. Using the Random Forest technique is one novel way; doing so in combination with the CIC-DDoS 2019 dataset and the AWID (A Network Intrusion Dataset) dataset is especially effective. For the purpose of comparing and contrasting anti-DDoS techniques, the CIC-DDoS 2019 dataset was selected from an immense amount of available data. By employing Random Forest, which is an ensemble learning technique, the dataset's rich feature set can be analyzed to discern malicious traffic patterns.

Random Forest excels in the classification of network traffic, identifying anomalies, and predicting potential DDoS attacks. It's an appealing selection for enhancing web-based DDoS mitigation efforts because of its ability to process high-dimensional data and conform to the needs of constantly-evolving attack techniques. This approach shows promise in fortifying web services against DDoS threats, as it harnesses the power of machine learning and a robust dataset for improved attack detection and response.

A random forest model that achieves high accuracy rates and outperforms existing models for DDoS detection [8]. This is [9] also uses random forest along with other machine learning techniques to efficiently detect DDoS attacks. [10] Explore the use of an Extra Tree-Random Forest (ET-RF) model on the CIC-DDoS2019 dataset, achieving high accuracy and performance. Ma 2023 combines the random forest algorithm with distributed and edge computing in the SDN environment, achieving timely and accurate detection of DDoS attacks with high performance metrics. Overall, these papers demonstrate the effectiveness of the random forest algorithm in mitigating DDoS attacks using the CIC-DDoS2019 dataset and AWID (A Network Intrusion Dataset).

According to Techniques [11] LR-DDoS attack detection using machine learning models; proposed flow-based detection and mitigation system. In a software-defined network, controlled traffic analysis can be used for the detection and prevention against LR-DDoS attacks using machine learning. In order to identify low-rate DDoS attacks, this article introduces a flow-based detection and mitigation architecture that uses machine learning models including Support Vector Machine (SVM), C4.5 Decision tree, and Naive Bayes. In response to the threat posed by these kinds of attacks, the system was designed.

[12] This framework for detecting botnet attacks. Comparative examination of different classifiers for network attack detection. The proposed framework proposes to detect botnet attacks by utilising the CSE-CIC-IDS2018 dataset. When it comes to detecting attacks like these, it has been shown that artificial intelligence classifiers perform extraordinarily well compared to other types of classifiers. The paper mention the mitigation of DDoS attacks using Random Forest with the CIC-DDoS2019 dataset and AWID (A Network Intrusion Dataset) in web.

In addition to Trevor Pinto, a few more [13] Neither the use of the CIC-DDoS2019 dataset nor any preventative measures are specifically addressed, however the paper does recommend a hybrid approach to DDoS attack detection using Random Forest and MLP-ANN classifiers. Here, we use two types of Neural Networks—Transmitted Random Forest (RF) and Multi-layer Perceptron (MLP)—to create a classifier. Because of their excellent real-time accuracy, efficiency, adaptability, and flexibility in enhancing the parameters, these algorithms were chosen. Because of the complexity and ever-increasing nature of network technology, algorithms with these features were necessary.

This proposed model for detecting and mitigating against DDoS attacks in SDN networks that makes use of the Random Forest technique [14]. The CIC-DDoS2019 dataset is used for analysis. To this end, we proposed a model that can use ML to automatically detect attacks in SDN networks and take measures to mitigate them. Our model expands the native features, unlike other methods in the literature that just apply them for attack detection.

This study details an innovative method to detecting and classifying Distributed Denial of Service (DDoS) attacks. The authors perform their study and analysis using the CICDDoS2019 dataset, upon which their methodology is based on the random forest algorithm. An in-depth examination of existing data sets is followed by the creation of a novel classification for distributed DoS attacks in this investigation. In addition, we've created two additional datasets—CICDDoS2019 and AWID (A Network Intrusion Dataset)—to address the existing limitations. The dataset we gathered during the duration of our research allows us to also present a novel approach to family detection and classification using a set of network flow features.

According to the study provides a method for dynamic network traffic management that protects ISP infrastructure from distributed denial-of-service attacks [16]. The solution is an ensemble model that combines traditional methods with more advanced ones, such as Random Forests. Therefore, we advise an ISP to implement a DNTM system with an Attack Detector, an IP Prioritiser, a Traffic Manager, and a Netflow Classifier to better manage the flow of traffic. The IP prioritiser classifies IPs into two groups, both secure and risky, to determine their relative importance. In addition to relying on methods already employed by ISPs, the Traffic Manager makes use of methods including rate limiting, blackholing, normal routing, and entrance/way out filtering to carry out a wide range of prevention measures. The Netflow Classifier is a mixed-learning ensemble model since it includes both supervised and unsupervised training into its prediction process.

## 3. Proposed Methodology

The proposed method for Mitigating Distributed Denial of Service (DDoS) attacks in web based networks makes use of the Random Forest algorithm in combined with the CIC-DDoS 2019 dataset and AWID (A Network Intrusion Dataset), and thus offers a potent answer for real-time attack mitigation and efficient response, defence of the availability and dependability of web services. The details of the proposed method to protect against DDoS attacks are detailed here. At this stage, we are only extracting the datasets from the CIC-DDoS 2019 dataset and the AWID (A Network Intrusion Dataset). The second stage is to collect the data. The final stage involves using Ensemble Classifier methods to categorise DDoS attacks. The final phase involves testing our proposed approach using real-world data and comparing the results across multiple metrics. Each step of the process for mitigating the DDoS attack, as seen in Figure 2, will be summarised in full below.

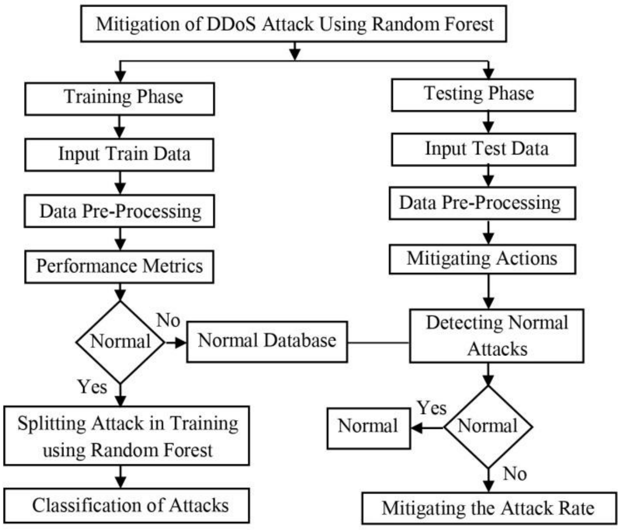

**Fig. 2.** The proposed Architecture of DDoA Attack Mitigation.

### 3.1 Datasets

Using random forest with the CIC-DDoS 2019 dataset and the AWID (A Network Intrusion Dataset) dataset, authors have explored the possibility of preventing and mitigating DDoS attacks in multiple studies. The machine learning technology of random forest has been used in various studies to effectively identify and classify DDoS attacks. For instance, Kamarudin et al. [17] proposed a random forest-based intrusion analysis system to enhance DoS attack detection. To top it all off, they did it while maintaining an accuracy of almost 98.8% Megala et al. [18] evaluated the classification of DDoS attacks, and they noticed that random forest gave good accuracy results when compared to other classifiers. Bindra and Sood used the Random Forest Classifier [19] to improve DDoS attack detection accuracy to above 96%. Manjula and Mangla, who both utilised random forest in their research, found that it correctly identified DDoS attacks with an accuracy of 96.75 % [20]. After looking into a variety of algorithms, including random forest, Sarmento et al. [21] found that the vast majority of them had an accuracy of 90% or above when it came to identifying DDoS attacks. The large number of samples in the dataset provides a promising option for evaluating the detection's precision.

### 3.2 Data Preprocessing

Using the CIC-DDoS 2019 dataset and the AWID dataset, data may be prepared for Random Forest, which can then be utilised in a web-based development environment to mitigate the effects of DDoS attacks. The effectiveness of various machine learning algorithms in detecting and mitigating distributed denial of service attacks has been examined in a number of research. Random Forest classifier is useful to detect distributed denial of service (DDoS) attacks, as mentioned in [22]. [DDoS is a short form for "distributed denial of service. [19] [21]. Since the CIC-DDoS 2019 dataset satisfies the criteria for a reliable study, it is a good option for testing DDoS detection algorithms [23]. This data collection contains actual attacks that were captured in the wild. During the preprocessing phase, using a feature selection strategy such information gain-based feature selection may improve the detection system's overall performance [24]. With the correct tools, it is possible to preprocess the data to detect DDoS attacks in a web-based environment. The Random Forest method is an example of the kind of algorithm that might be effective for reaching this objective.

#### 3.2.1 Feature Selection

High dimensionality is a feature of the selected datasets; which means that training time grows exponentially with the number of dimensions. A high-dimensional dataset can additionally increase the possibility of the model being overfit to the data, which is the second potential problem. In order to prioritise the current extensive list of characteristics, many feature selection methods are applied. For this reason, feature selection has received an extensive amount of study from the machine learning and data mining communities. If you hear the word "feature," what they really mean is the word "component," which can also refer to a property or a system. Features like Chi-square, Analysis of Variance, Mutual Information, Information Gain Ratio (IGR) and Correlation (CR) can be used to mitigate the impact of DDoS attacks [25, 26]. These methods aim to maximise classification precision by selecting relevant and informative features while reducing the number of dimensions. The forward inter-packet time difference (FwdIATTotal) and the flow length in microseconds (FlowDuration) have been recognised as two of the most crucial factors [27], [23]. Classifiers based on machine learning, such as random forests and decision trees, typically make use of these feature selection methodologies. It is shown that the proposed methods provide better levels of classification accuracy and efficiency when compared to state-of-the-art approaches. This is especially true when trying to tell the difference between attack flows and normal flows. When deciding which features to include, it is important to keep focus on the following objectives.

(1) Improve generalisation efficiency in comparison to a full-featured model.

(2) Make more precise conclusions and execute faster in response to observed evidence.

(3) To understand the procedure in more depth and with more logic, one must become connected with the basics of data collection.

Feature selection is an important step for both regression and classification during preprocessing.

#### 3.2.1.1 Chi-Square

The Random Forest classifier has been used for DDoS attack mitigation in online networks, and the Chi-Square feature selection method has also been employed. For this reason, we looked at the CIC-DDoS 2019 dataset as well as the AWID dataset. To identify and mitigate Distributed Denial of Service (DDoS) attacks, the proposed solution use machine learning methods. Several different methods of machine learning have proven to be helpful in identifying and categorising distributed denial of service (DDoS) attacks. These models include, but are not limited to, the ones used in Random Forest, AdaBoost, XGBoost, and neural networks. Using the Chi-Square (2) test to prevent distributed denial of service attacks is not common. To examine the link between two categorical variables, statisticians frequently use the Chi-Square test. The process of mitigating distributed denial of service (DDoS) attacks often does not involve this test. Instead, it is used in the first stages of data analysis to better understand the correlation between different factors and the frequency of distributed denial of service (DDoS) attacks. [28]. The Chi-Square test can be used to determine whether or not a certain category variable (such as attack type, originating IP address, or time of day) is significantly related to the presence or absence of DDoS attacks in the context of DDoS attack mitigation. An illustration of this would be the use of source IP addresses in a distributed denial-of-service attack. Reduce the dataset's complexity and improve DDoS attack detection with the help of feature selection methods like Chi-Square. [29] [30].

Modelling the Chi-Square test for independence produces the following equation 1:

$$\chi^2 = \sum \frac{(O-E)^2}{E} \qquad (1)$$

Where:

- $\chi^2$ is the Chi-Square statistic.
- O represents the observed frequency in a contingency table.
- E represents the expected frequency in a contingency table.

#### 3.2.1.2 Analysis of Variance

The random forest method has been used for DDoS attack mitigation in online networks, with analysis of variance (ANOVA) feature selection playing a role. For this, we used data from the CIC-DDoS 2019 dataset. The chi-square test, the additional tree, and the analysis of variance (ANOVA) are used as feature selectors for numerous machine learning classifiers, such as Random Forest, Decision Tree, k-Nearest Neighbours, and XGBoost, utilising the presented technique. The purpose of using an ANOVA is to establish if there are statistically significant differences between the means of different groups or situations by analysing the data. The data can be compared to itself to achieve this goal. It's typically used in the context of experimental design and hypothesis testing. While ANOVA can be used to analyze data related to DDoS attacks, it is not a direct tool for mitigation; instead, it is employed for understanding the effects of various factors on network performance or other related metrics [28]. In a Software-Defined Network (SDN), the random forest technique has been implemented for DDoS detection as well as mitigation. With an average accuracy of 98.38% and a mean detection time of 36 ms [31], the system is able to constantly modify the DDoS method's transmission rules. The general equation for one-way ANOVA, which compares the means of multiple groups, equation 2 is as follows:

$$F = \frac{MS_{between}}{MS_{within}} \qquad (2)$$

Where:

- F*F* is the test statistic for ANOVA.
- $MS_{between}$is the mean square for between-group variation.
- $MS_{within}$is the mean square for within-group (residual) variation.

#### 3.2.1.3 Mutual Information

You are correct in saying that the amount of information one random variable has about another random variable may be measured using the Mutual Information equation 3. One way to achieve this is to contrast the two lists. Here's the formal equation for Mutual Information:

$$I(X; Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x, y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad (3)$$

In this equation:

- The mutual information between two random variables X and Y can be written as -I(X; Y).

-Together, X and Y form a probability distribution denoted by p(x, y).

- Marginal probability distributions of X and Y are represented by the notations p(x) and p(y), respectively.

The level of correlation between X and Y is measured using the following equation. It's used for measuring the degree to which one variable depends on another. In the context of feature selection, MI is often used to assess how much information a feature (X) contains about the output or target variable (Y). Features with high MI values are considered more informative and can be selected for inclusion in a machine learning model [32]. To apply this concept in DDoS attack mitigation, you would typically use Mutual Information to quantify the importance of various features (e.g., network traffic characteristics) in relation to the output variable indicating the presence or absence of a DDoS attack. Features with higher MI values are likely to be more relevant for detecting DDoS attacks and can be included in your feature subset for modeling [30].

#### 3.2.1.4 Information Gain Ratio (IGR)

During the feature-selection process, the Information Gain Ratio (IGR) is a metric that is used to evaluate a feature's importance in relation to the dependent variable. It's often used in decision tree-based classifiers and can be applied to DDoS attack mitigation to select relevant features for building models [33]. The formula 4 for Information Gain Ratio is as follows:

$$\text{IGR(S, A)} = \frac{\text{IG(S,A)}}{\text{SplitInformation(S,A)}} \qquad (4)$$

Where:

- IGR(S, A) is the Information Gain Ratio for feature A in the dataset S.

- IG(S, A) is the Information Gain, which measures the reduction in entropy when you split the data based on feature A.

- Split Information (S, A) is the Split Information, which quantifies the potential information generated by splitting the dataset based on feature A.

The formula 5 for Information Gain, IG(S, A), is often expressed as:

$$\text{IG(S, A)} = \text{H(S)} - \text{H(S|A)} \qquad (5)$$

Where:

-The entropy of a given dataset is denoted by the notation H(S).

-The conditional entropy of S for each feature A is denoted by H (S|A).

Entropy measures the impurity or disorder in a dataset. In the context of DDoS attack mitigation, you might have a dataset with various features, including network traffic characteristics, and a binary target variable indicating the presence or absence of a DDoS attack. Information Gain Ratio can be used to assess how relevant each feature is in classifying or predicting DDoS attacks [28].

#### 3.2.1.5 Correlation (CR)

The statistical measure of correlation (CR) can be applied to examine the bond between two independent variables. Examining the correlations between various aspects or features of network traffic and the existence or impact of DDoS attacks is a common practise in the context of DDoS attack mitigation. The strength and direction of these correlations can be measured by various correlation coefficients, but there is no universal "Correlation (CR)" equation. A linear relationship between two continuous variables can be determined by using the Pearson correlation. There is no linear correlation if the value is zero [30]. The range is (-1, perfect negative correlation) to (+1, perfect positive correlation). The formula for Pearson correlation is:

$$\text{R} = \frac{\sum(X-X)(Y-Y)}{\sqrt{\sum(X-X)^2(Y-Y)^2}} \qquad (6)$$

Where:

•The Pearson correlation coefficient is denoted by "R."

• The two variables X and Y being associated are X and Y.

• X¯ and Y¯ are the means of X and Y, respectively.

Correlation analysis can be a valuable part of exploratory data analysis for understanding how various attributes relate to DDoS attacks, but it is not a direct mitigation technique. Mitigation typically involves using security measures, network infrastructure hardening, and response strategies to prevent or minimize the impact of DDoS attacks [23].

### 3.3 Classification of DDoS Attack

The classification models used have been described in depth below. Selecting the appropriate conditions for each model is required for optimal performance.

#### 3.3.1 Random Forest

Multiple researches have employed the Random Forest method to prevent DDoS attacks. Datasets as CIC-DDoS 2019 and AWID are used with the method. Researchers in one study [34] used the SDN architecture and the Random Forest algorithm to create a system for detecting and stopping distributed denial-of-service (DDoS) attacks. Every time a tree is divided, m features are selected at random and analysed using a measure to determine which one should trigger the split. As part of the training process, we create a new training set $T_i$ for each tree by selecting representative examples from the larger training set T. Each tree split is constructed by randomly selecting m features and analysing them using a measure to determine which one should lead to the division. This randomization leads to the generation of numerous trees, the combination of which typically yields improved prediction performance. In addition to reducing the amount of time it takes to train a model, classifying embedded features, and evaluating the impact of features, RF models provide a number of other advantages. This is in contrast to the lengthy training time required by many alternative machine learning methods. In order to detect DDoS attacks, RF is trained on a variety of feature sets. The RF settings are listed in Table 1.

**Table 1.** Parameters used for Random Forest algorithm.

| Names of Hyper Parameter | Values of Parameter |
|---|---|
| n_estimators | 40 |
| max_depth | 50 |
| min_samples_split | 2 |
| min_samples_leaf | 1 |
| max_features | Automatic |
| bootstrap | True |
| criterion | Gini impurity |
| class_weight | 2 |
| oob_score | 0 |
| random_state | 0 |

Now we will see how the Random Forest algorithm works to mitigate the DDoS attack with CIC-DDoS 2019 and AWID datasets. The Random Forest Algorithm steps as follows:

**Algorithm:** Random Forest Algorithm for Mitigation of DDoS attack.

Input: CIC-DDoS 2019 and AWID datasets

Output: Classification of different type of DDoS attacks

Step 1: Take K arbitrary samples from the training data.

Step 2: Use Pre-processing of Data

Step 3: Sort the data into clusters with similar but dissimilar sizes.

Step 4: Create a training and test set of data.

Step 5: Using a metric of information gain, choose the optimal subset of features to analyse.

InfoGain (Class, Attribute) = H (Class) – H (Class | Attribute)

Step 6: Random forest is trained using the data set.

Step 7: Random forest is then used to classify the test data.

Step 8: Calculate the likes of the Mathew correlation coefficient, the F-measure, and the precision.

We obtained the CIC-DDoS 2019 and AWID datasets and stored them in CSV format to investigate the potential of the Random Forest method for preventing DDoS attacks in web-based networks. All missing feature values in the CIC-DDoS 2019 and AWID datasets have been refilled using the replace missing values filter in weka. This filter will apply the training data's median and mean to any missing values.

### 3.4 Evaluation Measures

Prediction model efficacy is measured using these evaluation metrics. Measuring machine learning's accuracy, precision, recall, and F score in the framework of mitigating distributed denial-of-service (DDoS) attacks was the primary objective of this study.

#### 3.4.1 Accuracy

To be useful as an evaluation metric, accuracy requires both consistent datasets and close false positive and false negative values. Accuracy is measured as the proportion of correct predictions relative to the total number of observations. The accuracy of a classifier is the percentage of occurrences for which it produces accurate predictions, as represented by Equation 7.

$$\text{Accuracy} = \frac{TP}{TP+TN+FP+FN} \qquad (7)$$

#### 3.4.2 Precision

To measure the accuracy, one can compare the actual number of comes to the predicted percentage of outcomes. The rate of false positives drops dramatically when accuracy increases. How accurately the classifier is making positive class predictions are measured by precision. Equation 8 allows for the determination of accuracy.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (8)$$

#### 3.4.3 Recall

Recall measures how many correct predictions were produced for positive class observations relative to the total number of observations. Precision, as indicated in Equation 9, is a measure of the classifier's accuracy in making true positive class predictions.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (9)$$

#### 3.4.4 F1 Score

The F1 Score is determined based on the normalised mean of the recall and accuracy metrics. As a consequence of this fact, the score takes into consideration both false positives and false negatives. There are several situations in which the F1 score is of greater advantage, although the fact that accuracy is of a higher priority. According to Equation 10, the F1 score is calculated by taking the harmonic mean of the precision score and the recall score.

$$\text{F1 Measure} = 2 \times \frac{PR}{(P+R)} \qquad (10)$$

## 4. Experimental Results

Internet-based networks face serious challenges, including detecting and preventing DDoS attacks. The detection of a DDoS attack is a binary class problem, with the classes being "normal" and "DDoS attack." Benign is an everyday kind. Since we're interested in detecting attacks, we'll treat the presence of attacks as a positive class and benign as a negative class. We employ a Random Forest method. Using the Random Forest method, we pick 12, 18, and 24 characteristics, respectively. The features of interest are subjected to machine learning techniques such as ARTP, N-KPCA+GA+SVM. The experimental design is laid forth in Table 2. To illustrate the efficacy of these strategies, Figure 3 presents twelve features, eighteen features, twenty-four features, and all features from the CIC-DDoS 2019 and AWID datasets. The experimental results reveal that Random Forest outperforms the 12-feature, 18-feature, and 24-feature techniques in protecting against DDoS attacks. Compared to other existing approaches, Random Forest's miss classification rate is quite low thanks to these characteristics.

**Table 2.** Details about the experimental layout

| S.No | Components | Description |
|---|---|---|
| 1 | IDE | Google Colab Notebook |
| 2 | Programming Language | Python 3.11 |
| 3 | Backend | Microsoft Excel |
| 4 | Libraries | NumPy, Pandas, SciPy |
| 5 | Storage Memory | 8 GB RAM |
| 6 | Operating System | Windows 10 |
| 7 | System Hardware | Intel Core i5 13th Gen PC |

In Table 3, we can see the confusion matrices of various machine learning models on the CIC-DDoS 2019 dataset and the AWID dataset. True negative, false positive, real negative, and true positive rates for each feature set are presented in the results. Misclassification rates between approaches can be compared by seeing the findings of the confusion matrix. If there are fewer misclassifications, then the accuracy is higher. In addition to the results obtained from different techniques, the accuracy of these methods is 99.23%. Calculating accuracy simply is not a useful parameter when working with large datasets because the fact that some data points were incorrectly categorised does not affect the accuracy. The experiment was conducted in this study using a different set of features and the CIC-DDoS 2019 and AWID datasets. The miss classification rate provided by the ARTP's selection of 12 features for the N-KPCA+GA+SVM algorithm is lower than the rate produced by the other method. When compared to other classifiers, the miss classification rate of the support vector machine (SVM) is highest for all features. Table 3 shows that the miss classification rate is small when employing the RF-selected set of 18 features. Overall, we find that the miss classification error is substantial when using all features sets together.

**Table 3.** For various models using the CIC-DDoS 2019 dataset and a confusion matrix with 12, 18, and 24 features.

| Methods | True Negative | True Positive | False Negative | False Positive |
|---|---|---|---|---|
| 12 features | | | | |
| ARTP | 20246 | 26468 | 20 | 4 |
| N-KPCA+GA+SVM | 20135 | 26789 | 15 | 4 |
| RF | 20578 | 26893 | 8 | 3 |
| 18 features | | | | |
| ARTP | 20247 | 26468 | 21 | 5 |
| N-KPCA+GA+SVM | 20138 | 26765 | 17 | 5 |
| RF | 20654 | 26895 | 12 | 4 |
| 24 features | | | | |
| ARTP | 20246 | 26465 | 21 | 4 |
| N-KPCA+GA+SVM | 20139 | 26764 | 16 | 6 |
| RF | 20656 | 26894 | 14 | 7 |

In Figure 3, we can see the results of applying different machine learning techniques to a dataset including 12, 18, and 24 features, and their ability to detect and mitigate DDoS attacks. The RF technique was used to pick up these particular characteristics. The RF method is applied to the problem to detect the most important feature for DDoS detection and mitigation. When compared to other approaches, the RF's predictive accuracy is higher.

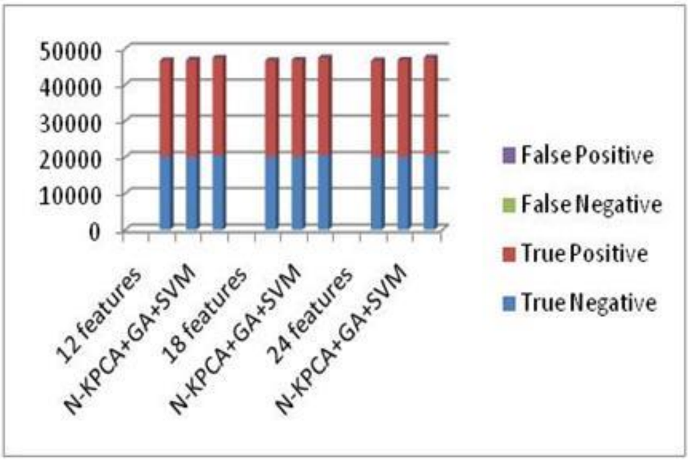

**Fig. 3.** Comparisons of various machine learning methods on 12, 18 and 24 features.

The proposed technique and several current methods are compared in Table 4. When compared to state-of-the-art alternatives, the proposed method stands out for its high accuracy and low rate of incorrect classifications. The current approaches only achieve an accuracy of about 99.23%, and they have more miss categorization errors. One such attack is all that is needed to significantly boost the miss classification rate using the proposed strategy. The proposed strategy was successful in achieving both a low percentage of miss classification errors and a high degree of accuracy. This was accomplished through running the machine learning approach through tests on a wide variety of feature sets and experimenting with the machine learning classifier's parameters. In this study, machine learning is used to detect and classify DDoS attacks. The operational requirements of the tree-based methods are significantly less than those of the distance-based method. While tree-based methods are often more quickly, ARTP is still widely used. In terms of miss classification mistakes, the N-KPCA+GA+SVM strategy performs poorly in compared to the other methods. Miss classification mistakes can be reduced by fine-tuning the parameters of these algorithms. As the number of data dimensions grows, the time required to identify RF features also increases.

**Table 4.** Comparison of the proposed method with the previous methods.

| | ARTP | N-KPCA+GA+SVM | RF |
|---|---|---|---|
| Precision | 0.938 | 0.941 | 0.956 |
| Sensitivity/recall | 0.944 | 0.951 | 0.956 |
| Specificity | 0.915 | 0.920 | 0.927 |
| Accuracy | 0.985 | 0.987 | 0.992 |
| F1-Measure | 0.895 | 0.898 | 0.901 |

In Figure 4, we see the results of several different machine learning algorithms that uses all of the available information to identify and mitigate distributed denial-of-service (DDoS) attacks. When compared to other methodologies, the RF's level of accuracy in making predictions is the highest. Comparisons between RF with N-KPCA+GA+SVM and ARTP [35] are provided using precision, recall, specificity, F-measure, and accuracy.

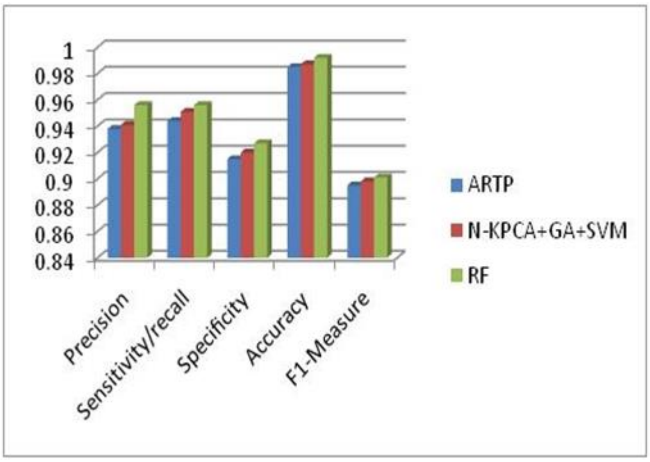

**Fig. 4.** A comparison is presented between ARTP, N-KPCA+GA+SVM and RF.

## 5. Conclusion

In this research paper offers a robust solution to the pressing issue of Distributed Denial of Service (DDoS) attacks in web-based networks. Through the implementation of ensemble learning, with a specific emphasis on the Random Forest algorithm, the study showcases an effective approach for improving DDoS attack detection. The paper underscores the strength of Random Forest in handling complex, high-dimensional data, analyzing feature importance, and preventing overfitting. The process begins with the collection and preprocessing of network traffic data, encompassing various parameters such as packet speeds, sizes, and protocols, which serves as the groundwork for the proposed approach. Labeled data is skillfully classified into regular and DDoS traffic, effectively processed by the Random Forest classifier. Thorough evaluation using key metrics, including accuracy, precision, recall, and F1-score, demonstrates the model's high accuracy and reliability in predicting DDoS attacks, with minimal false positives. The proposed system offers a proactive and scalable defense against DDoS attacks, safeguarding web-based services and user experiences by adapting to evolving attack patterns. The paper's use of modern datasets, specifically CIC-DDoS 2019 and AWID, substantiates the effectiveness of the proposed models, providing a solid foundation for real-world application and continued progress in DDoS mitigation within web-based networks. Importantly, Random Forest feature selection was used in the experiment, and the findings show that, along with other algorithms like ARTP and N-KPCA+GA+SVM, Random Forest provides high prediction accuracy, with Random Forest leading the pack at 99.23%. This suggests that Random Forest serves as an excellent feature selection technique, further reinforcing the paper's findings.

## References

[1] Mishra, A., Gupta, N. & Gupta, B.B. Defensive mechanism against DDoS attack based on feature selection and multi-classifier algorithms. Telecommun Syst 82, 229–244 (2023).

[2] Gocher, H., Taterh, S. & Dadheech, P. Impact Analysis to Detect and Mitigate Distributed Denial of Service Attacks with Ryu-SDN Controller: A Comparative Analysis of Four Different Machine Learning Classification Algorithms. SN COMPUT. SCI. 4, 456 (2023).

[3] Vidyashree, L., Suresha Mitigation of cyber attacks assuring security with conglomerate edict based intrusion detection system in IoT. Sādhanā 47, 67 (2022).

[4] Bawany, N.Z.; Shamsi, J.A.; Salah, K. DDoS Attack Detection and Mitigation Using SDN: Methods, Practices, and Solutions. Arab. J. Sci. Eng. 2017, 42, 425–441.

[5] Jain, M., Kaur, G. Distributed anomaly detection using concept drift detection based hybrid ensemble techniques in streamed network data. Cluster Comput 24, 2099–2114 (2021).

[6] Catillo, M., Del Vecchio, A., Pecchia, A. et al. Transferability of machine learning models learned from public intrusion detection datasets: the CICIDS2017 case study. Software Qual J 30, 955–981 (2022).

[7] Sambangi, S.; Gondi, L. A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear

Regression. Proceedings 2020, 63, 51.

[8] T. S. Chu, W. Si, S. Simoff and Q. V. Nguyen, "A Machine Learning Classification Model Using Random Forest for Detecting DDoS Attacks," 2022 International Symposium on Networks, Computers and Communications (ISNCC), Shenzhen, China, 2022, pp. 1-7, doi: 10.1109/ISNCC55209.2022.9851797.

[9] F. Nazarudeen and S. Sundar, "Efficient DDoS Attack Detection using Machine Learning Techniques," 2022 IEEE International Power and Renewable Energy Conference (IPRECON), Kollam, India, 2022, pp. 1-6, doi: 10.1109/IPRECON55716.2022.10059561.

[10] V. Gaur and R. Kumar, "ET-RF based Model for Detection of Distributed Denial of Service Attacks," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 1205-1212, doi: 10.1109/ICSCDS53736.2022.9760938.

[11] Krishna Kishore, P., Prathima, K., Eswari, D.S., Goud, K.S. (2023). Bidirectional LSTM-Based Sentiment Analysis of Context-Sensitive Lexicon for Imbalanced Text. In: Bhateja, V., Sunitha, K.V.N., Chen, YW., Zhang, YD. (eds) Intelligent System Design. Lecture Notes in Networks and Systems, vol 494. Springer, Singapore. https://doi.org/10.1007/978-981-19-4863-3_27

[12] V., Kanimozhi., T., Prem, Jacob. (2021). Artificial Intelligence outflanks all other machine learning classifiers in Network Intrusion Detection System on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing. ICT Express, 7(3):366-370. doi: 10.1016/J.ICTE.2020.12.004

[13] Trevor, Pinto., Yakub, Sebastian. (2021). Detecting DDoS attacks using a cascade of machine learning classifiers based on Random Forest and MLP-ANN. doi: 10.1109/MASCON51689.2021.9563266

[14] Fatima, Khashab., Joanna, Moubarak., Antoine, Feghali., Carole, Bassil. (2021). DDoS Attack Detection and Mitigation in SDN using Machine Learning. 395-401. doi: 10.1109/NETSOFT51509.2021.9492558

[15] Iman, Sharafaldin., Arash, Habibi, Lashkari., Saqib, Hakak., Ali, A., Ghorbani. (2019). Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy. doi: 10.1109/CCST.2019.8888419

[16] P. Krishna Kishore, S. Ramamoorthy, V.N. Rajavarman, "Mitigation of HTTP Flood DDoS Attack in Application Layer Using Machine Learning and Isolation Forest," SSRG International Journal of Electrical and Electronics Engineering, vol. 10, no. 10, pp. 6-19, 2023. Crossref, https://doi.org/10.14445/23488379/IJEEE-V10I10P102

[17] Imran, Edzereiq, Kamarudin., Mohd, Faizal, Ab, Razak., Ahmad, Firdaus., M., Izham, Jaya., Yau, Ti, Dun. (2021). Performance Analysis on Denial of Service attack using UNSW-NB15 Dataset. doi: 10.1109/ICSECS52883.2021.00083

[18] G, Megala., S., Prabu., B, C, Liyanapathirana. (2021). Detecting DDoS Attack: A Machine-Learning-Based Approach. 55-66. doi: 10.4018/978-1-7998-3335-2.CH004

[19] Naveen, Bindra., Manu, Sood. (2019). Detecting DDoS Attacks Using Machine Learning Techniques and Contemporary Intrusion Detection Dataset. Automatic Control and Computer Sciences, 53(5):419-428. doi: 10.3103/S0146411619050043

[20] H., T., Manjula., Neha, Mangla. (2021). An approach to on-stream DDoS blitz detection using machine learning algorithms. Materials Today: Proceedings, doi: 10.1016/J.MATPR.2021.07.280

[21] Augusto, Gonzaga, Sarmento., Kheng, Cher, Yeo., Sami, Azam., Asif, Karim., Abdullah, Al, Mamun., Bharanidharan, Shanmugam. (2021). Applying Big Data Analytics in DDos Forensics: Challenges and Opportunities. 235-252. doi: 10.1007/978-3-030-68534-8_15

[22] Atheer, Alharthi., Ala', Eshmawi., Azzah, Kabbas., Lobna, Hsairi. (2020). Network Traffic Analysis for DDOS Attack Detection. doi: 10.1145/3440749.3442637

[23] R., R., Rejimol, Robinson., Ciza, Thomas. (2021). Low Rate Multi-vector DDoS Attack Detection Using Information Gain Based Feature Selection. 685-696. doi: 10.1007/978-981-16-0965-7_53

[24] Iman, Sharafaldin., Arash, Habibi, Lashkari., Ali, A., Ghorbani. (2018). A Detailed Analysis of the CICIDS2017 Data Set. 172-188. doi: 10.1007/978-3-030-25109-3_9

[25] Vimal, Gaur., Rajesh, Kumar. (2022). FSMDAD: Feature Selection Method for DDoS Attack Detection. 939-944. doi: 10.1109/ICEARS53579.2022.9752308

[26] Lu, Zhou., Ye, Zhu., Tianrui, Zong., Yang, Xiang. (2022). A feature selection-based method for DDoS attack flow classification. Future Generation Computer Systems, 132:67-79. doi: 10.1016/j.future.2022.02.006

[27] Deepak, Kshirsagar., Sandeep, Kumar. (2021). An efficient feature reduction method for the detection of DoS attack. ICT Express, 7(3):371-375. doi: 10.1016/J.ICTE.2020.12.006

[28] Vimal, Gaur., Vimal, Gaur., Rajneesh, Kumar. (2021). Analysis of Machine Learning Classifiers for Early Detection of DDoS Attacks on IoT Devices. Arabian Journal for Science and Engineering, 1-22. doi: 10.1007/S13369-021-05947-3

[29] Subhashini, Peneti., Hemalatha, E. (2021). DDOS Attack Identification using Machine Learning Techniques. doi: 10.1109/ICCCI50826.2021.9402441

[30] P., Ravi, Kiran, Varma., K., V., Subba, Raju., Suresh, Ruthala. (2021). Application of Whale Optimization Algorithm in DDOS Attack Detection and Feature Reduction. 93-102. doi: 10.1007/978-981-33-4305-4_8

[31] Heru, Nurwarsito., Muhammad, Fahmy, Nadhif. (2021). DDoS Attack Early Detection and Mitigation System on SDN using Random Forest Algorithm and Ryu Framework. 178-183. doi: 10.1109/ICCCE50029.2021.9467167

[32] Mona, A., Alduailij., Qazi, Waqas, Khan., Muhammad, Zeeshan, Tahir., Muhammad, Sardaraz., Mai, A., Alduailij., Fazila, Malik. (2022). Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method. Symmetry, 14(6):1095-1095. doi: 10.3390/sym14061095

[33] Trevor, Pinto., Yakub, Sebastian. (2021). Detecting DDoS attacks using a cascade of machine learning classifiers based on Random Forest and MLP-ANN. doi: 10.1109/MASCON51689.2021.9563266

[34] H. Nurwarsito and M. F. Nadhif, "DDoS Attack Early Detection and Mitigation System on SDN using Random Forest Algorithm and Ryu Framework," 2021 8th International Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, Malaysia, 2021, pp. 178-183, doi: 10.1109/ICCCE50029.2021.9467167.

[35] P. Krishna Kishore, S. Ramamoorthy, V.N. Rajavarman, ARTP: Anomaly based real time prevention of Distributed Denial of Service attacks on the web using machine learning approach, International Journal of Intelligent Networks, Volume 4, 2023, Pages 38-45, ISSN 2666-6030, https://doi.org/10.1016/j.ijin.2022.12.001.