Hybrid Technique for DDOS Attack Detection Using Machine Learning

Deepak Singh Rajput, Dr. Arvind Kumar Upadhyay

[1]Research Scholar, [2]Professor, Dept. of Computer Science and Engineering

Amity School of Engineering &Technology (ASET)

Amity University, Gwalior Madhya Pradesh

Article Info

Page Number: 7847 - 7870

Publication Issue:

Vol 71 No. 4 (2022)

Abstract

The demand for the Internet is increasing every day, raising concerns about network security. Distributed Denial of Services (DDoS) attacks have harmed network availability for the decades and there is still no technique available to protect completely against these attacks. DDoS attacks are one of the biggest and fastest-growing cyber threats to network security. A DDoS attacks are an attempt to make a service resource unavailable, making it unusable for some time. Therefore, recognizing different forms of DDoS attacks with improved algorithms and higher accuracy while keeping the computing time less under control has become the most challenging factor. Firewalls and antivirus software are no longer sufficient in today's world to keep a company safe from the variety of assaults it faces. Traditional intrusion detection systems and firewalls can detect attacks based on signature patterns. Existing developments are insufficient to detect unknown threats. In order to identify and classify different types of anonymous attacks, it is necessary to apply intelligent technologies. Machine Learning (ML) has advanced significantly in terms of technology, bringing up lots of new research opportunities for addressing current and future network security challenges. In this research, machine learning methods and the significance of security in the context of various types of DDoS attacks and various ML classification algorithms such as Random Forest (RF), Naive Bayes (NB), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost), etc. have been compared in terms of attack detection. In addition, a faster and more accurate machine learning-based attack detection system has been proposed for accurate and fast DDoS attack detection.

Article History

***Article Received**: 25 March 2022*

***Revised**: 30 April 2022*

***Accepted**: 15 June 2022*

***Publication**: 19 August 2022*

Keywords: Distributed Denial of Services, Machine Learning, K-Nearest Neighbour, Support Vector Machine, eXtreme Gradient Boosting, Random Forest, Naive Bayes, Adaptive Boosting

1. Introduction

DDoS attacks are amongst the most popular and significant cyber attacks in recent history [1]. The goal of a DDoS attack is to consume the victim's resources. The attacker sends a large amount of traffic to the victim. As a result, these services will not be used for some time and will not be able to serve legitimate customers. This is the most common and most annoying problem for both the service providers and their users [2]. DDoS attacks primarily threaten the availability of computer resources and can lead to financial loss or loss of trust. Availability can be impacted by a variety of factors, including software or hardware failure, power outages and human mistake. Perhaps the most well-known access attacks are intentional and malicious blocking of the availability of a system, server, web application, web service, or the entire system. A denial of service attack makes them disappear. DDoS attacks are the biggest threat to the IT industry, and their number is increasing significantly every year [3]. However, due to computing power limits, central servers cannot process large amounts of data, such as large amounts of Internet traffic, in a short amount of time. When launching a DDoS attack, a large amount of internet traffic needs to be tracked. This is a difficult task for the server. Some monitoring systems use packet sampling to reduce the amount of input, but the output is not accurate. Furthermore, a single server makes the system prone to failure. If the server crashes, you can't fix it right away without interrupting your work. Online flow monitoring is similar to flow analysis with an unlimited input range.

Therefore, an intrusion detection system (IDS) is always needed to solve DDoS problems and maintain confidentiality and integrity several types of DDoS attacks are common. The most common are UDP, ICMP HTTP and SYN attacks. As the number of cyber attacks on critical network resources increases and some network monitoring technologies does not detect them, advanced techniques should be investigated and used to detect and report such attacks. Artificial Intelligence (AI) and Machine Learning (ML) are two of the biggest technological advancements that can transform modern security architectures. All technologies that allow computers to mimic human behaviour are artificial intelligence [4]. Machine learning is the ability to learn without explicit programming. Both are widely used in many industries such as healthcare, finance, and warehousing. For this reason, researchers track past DDoS attacks.

Several detection algorithms have been proposed to detect DDoS attacks. However, current attack detection methods still have problems with true negative values, low accuracy, and accuracy. Therefore, it is difficult to guarantee reliability, stability, and versatility. To solve the above problems, this article discusses current machine learning methods and proposes an improved method for detecting DDoS attacks by training a set of hybrid multi-classifiers.

2. Literature Review

Wu Zhijun et al. [18] proposed an investigation into multi-character DDoS attacks. SVM and Self-Organizing Map (SOM) were used to find out about DDoS. This only works with specific data, not with general forecasting tasks.

Shi Dong et al. [19] Uses best KNN and four attributes (flow length, stream size and throughput) detects DDoS attacks. It uses a grouping of the DDADA algorithm and the DDAML algorithm, but there are much further research is needed.

Abdulhamad et al. [20] proposed a ML-based IDS, which includes classification algorithms and feature selection methods. Various classifiers are used such as AdaBoost, RF, RT, J48, Logit Boost, MLP, ZeroR. They chose 4 useful sets of functions 5, 10, 7, and 32 to train the model. Using a random forest classifier with 32 picked feattures, their assessment findings indicate the best results. The Precision 0.995 and Feedback 0.966 classification algorithms have a performance level of 99.64%.

Saba al-Zahrani et al. [21] proposed a signature based ANN. It consist a signature based approach in which, if the attack has known features, use the ANN based approach; otherwise anomaly based neural networks are used to detect unknown DDoS attacks.

Sayakat Das et al. [22] use a combination of different techniques, such as ANN, SVM, MLP, NB, Multiple Adaptive Regression spline (MARS), K nearest neighbour (KNN) to detect DDoS attacks.

Suman Nandy et al. [23] used combination of Decision table, Naive Bayes, random forest, J48 and other five methods of selecting features to obtain information, rate of return, chi- square, relief f, and symmetric uncertainty.

The paper [24] by Belouch et al. estimated the performance of 4 classification algorithms: SVM, NB, DT and RF. It has the 42-attributes method to build the model. These results show that the random forest classifier, which has a precision of 97.55%, is the best amongst the other classifiers. The random forest algorithm has a precision of 93.53%.

Gray et al. [25] came up with ways to detect intrusions that used Naive Bayes machines and supporting vector machines (SVMs) as classification algorithms. Selecting only 24 of the 42 features from the NSL-KDD dataset is done by using the subset type, which is used when selecting features. Experiments show that the SVM classifier is better than the Naive Bayes classifier, with an overall accuracy of 93.95%, which is better than the Naive Bayes classifier.

Iman et al. [26] used the random forest as a classification algorithm with a choice of 34 features. The result of the proposed model was 0.99, which was evaluated in terms of accuracy, sensitivity and specificity.

Swati Sahu et al. [28] proposed a model that detects traffic in the domain is normal or not. Further it forms a filter to classify traffic as suspicious or normal. If suspicious then it is passed to a honeypot. It is installed at the server level, not at the client level. The Honeypot presumes that the attack must be observable using a signature-based detection mechanism.

Kachavimath et al. proposed a DDoS attacks detection model by using machine learning techniques to improve network security [29]. The K-nearest neighbour algorithm and the Naïve Bayes algorithm are used to classify them on the basis of eight different features. Various parameters was used to evaluate the accuracy rate of the classification algorithm 98.51%, recall rate 97.8%, sensitivity 97.8%, measured value f 1.005%, efficiency 98.48%, coefficient error 1.50%, specificity 99.12% and ROC 0.99%.

Yuze Su et al. [30], used the technique known as phase space reconstruction techniques to characterize the original flow. To identify DDoS attacks, RBF neural network is deployed to train network traffic patterns.

Shanmuga Priya et al. [31] used 3 classification methods, KNN, NB and RF divide DDoS data packets into two characteristics in general data packets, namely incremental time and data packets.

Sah et al. [32] proposed the intelligent IDS using various NB, KNN, RF and SVMs. Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) methods were used to reduce feature set. This task was tested using 41 features and comparing the feature reduction of 11, 12, 13, and 15 groups of different selected features with RF Classification.

Gaganjot Kaur et al. [33] used Bayesian Networks, Waves, SVM and kNN. There were parameters such as data packet bandwidth, duration and accuracy used for traffic monitoring. The precision level is applied to the KNN dataset.

Box, K. et al. [34] uses a hybrid SVM method to which combines SVM and Self Organized Map (SOM). Again both the methods were used separately for detecting the DDoS attacks in network traffic.

Shuang Wei et al. [35] proposed 2-tier architecture. It enters and collects aggregated information about the console process. DDoS attack detected using KL distances of real time flow distribution with normal time and the packet rate.

Fitni et al. [36] proposed a method for heterogeneous IDS based on the integration of DT, gradient boosting and LR as a classifier. After the implementation of the hybrid scheme, 23 items out of 80 were selected as feature-set. The results of comparison of seven different classifiers show that the 3 classifiers outperformed in terms of accuracy 98.8%, memorization 97.1%, accuracy 98.8% and F1. 97.9%.

Alireza Seifousadati et al. [37] proposed a ML approach for DDoS detection by implementing classifiers like NB, SVM, KNN, AdaBoost, Random Forest and XGBoost with dataset CICDDoS2019. They found that AdaBoost and XGBoost were producing extremely accurate results. Also it was observed that XGBoost provides slightly better training and detection time than AdaBoost.

Mugunthan, S. R. [38]. employed a Hidden Markov model to analyze network traffic flow characteristics, which are then used to train a random classifier to detect irregular network traffic flow. The entropy predicts attack probability, whereas the Hidden–MM predicts attack severity. The RF is trained using bootstrap aggregation methodology to identify normal traffic from attacked flow. The model outperforms the previous models ABC-ANN and ATBA in classification accuracy using the KDD CUP 99 data set.

Vivekanandam, B. [39] suggested a machine learning method for solving functional selection problems using GA (genetic algorithms). The method has the highest chance of overcoming function selection issues during training for different population sizes and identifies distinct malware groups. The method improves the mean and standard deviation in the optimization process for various datasets.

2.1 Performance comparisons of different classification algorithms based upon the literature review.

Ref. and Year	Dataset	Number of Features	Classification Algorithm	Evaluation Metrics	Limitations / Negative aspect s
[20] 2018	AWID	32 set,10set 7 set,5 set	Random Forest, AdaBoost, MLP J48, logit Boost,	Best results with Random Forest including 32 features. Precision 0.995, Accuracy 99.64%, and recall 0.966.	Correlation feature selection algorithm is used which is heavily dependent on the model, so they can fail to fit the data well.
[24] 2018	UNSW-NB15	42 features out of 49	SVM, Random Forest, Naïve Bayes,	Best results with Random Forest classifier. Sensitivity 93.53, Accuracy 97.49, Specificity 97.75 .	Any specific advanced feature selection method was not used and random selection of features can cause outliers and over-

			Decision Tree		fitting issues in machine learning model.
[25] 2019	NSL–KDD	24 features	Naïve Bayes, SVM	SVM best accuracy of 93.95	Any optimization method was not used in proposed system. Results can be improved by using optimization methods.
[26] 2020	NSL-KDD	34 Accepted features	Random Forest	Accuracy 0.9989, Specificity 0.9993 and Sensitivity 0.9985,	The boosting algorithms were not included in implementation. The outlier detection was not available in proposed method. It can cause incorrect classification results.
[29] 2020	NSL-KDD KDD Cup 99	8 Accepted features	KNN , Naïve Bayes	Best results with KNN classifier. Precision 98.9% Accuracy 98.51, Recall 97.8%, Sensitivity 97.8%, F-measure1.00%, Specificity99.12%, efficiency 98.48%, BCR 98.5%, ROC 0.99% and Rate of Error 1.50%.	With no specific advanced method of feature selection, very less correlated features were included for analysis. It can cause serious underfitting issues.
[32] 2020	NSL KDD	Selected different set 11,12,13, 15	SVM, KNN, Random Forest, NB	Optimized result by Random Forest classifier with DoS class. F-score 99.58%, precision 99.53, accuracy 99.63%, and recall 99.6%,	Proposed reduction methods are not able to deal with the high dimensionality of datasets. Also any of the boosting methods were not included.

[36] 2020	SE-CIC-IDS2018	23 Accepted features	Logistic regression, DT, Gradient boosting ensemble	Recall 97.1%, Accuracy 98.8%, Precision 98.8%, and F1 97.9%.	Single classification method was used for classification of feature. Ensemble method may give better classification results.

Table 1.1 Performance comparisons of different classification algorithms

2.3 Problem formulation

Previous researches and literature study evaluations have revealed that machine learning algorithms have a significant promise for accurate, precise, and quick DDoS attack detection. Several studies and research projects have previously been performed in this area, although there are still certain gaps in the detection of DDoS attacks:

i. The term "availability" refers to the data or service being available when it is needed. DDoS attacks are the most serious threat to this facility readiness.

ii. Machine learning is a relatively new technology that provides a set of classification methods for data classifiers that may be used to identify DDoS assaults accurately, quickly, efficiently.

3. DDoS Attack

The DDoS attacks are malicious attempt to block access to online services, typically by disrupting or temporarily shutting down hosting servers. DDoS attacks are non-intrusive network attacks that delays or disables a targeted service resource by flooding application network or server with forged traffic. A small amount of traffic is enough to attack the costly endpoints of vulnerable resources. DDoS attacks are an integral part of your security environment and are a risk that website owners should be aware of. Navigating the different types of DDoS attacks can be difficult and time-consuming. The sole intention of a DDoS attack is to overload the service provider’s resources. DDoS attacks can be used as a form of extortion. For example, the website owner can pay a ransom to the attacker to launch a DDoS attack [5].

3.1 Classification of DDoS Attacks

i. Volume-based

Large-scale attacks involve redirecting a enormous amount of requests to the target system. These requests are considered valid (fake package) or invalid (poorly formatted package) by the system. Hackers use joint attacks to disrupt network capabilities. These requests can be sent through multiple ports on your device. One of the techniques which are used by hackers to send data requests to third-party servers is UDP attacks. Because of this, they use your server's IP address as the return address. An enormous amount of data is then sent from the external server to the internal one. A third-party server that has magnified the data is all that is needed for a hacker to attack your server. Tens, hundreds, or even thousands of systems may be involved in such attacks. Large-scale attacks include flood attacks like UDP, ICMP floods and other false floods. The goal of this attack is to increase the bandwidth of the site and reduce its size to bits per second (Bps) [6].

ii. Protocol-Based

In this type of attack, hackers take advantage of flaws in web-server or application to suspend or shut down the web-server. Persistent application-based attacks involve partial server requests that attempt to hijack the entire server-to-database connection pool to intercept legitimate requests. These include slow and slow attacks, GET / POST floods, Apache, Windows or OpenBSD attacks, etc. These attacks aim to crash web servers with legitimate and harmless requests, the size of which is measured by the number of requests per second [6].

4. Machine Learning (ML)

ML is a subset of AI. It needs to provide algorithms to identify and predict future data patterns, in this case a model made up of data sets. Each model has its own formula to support the analysis of the provided data. As compare with other solutions, there are different types of advanced and fast ML algorithms available that can be used for accurate data classification and analysis for detection of DDoS attacks.

4.1 Classifier Methods in ML

An introduction to the many types of machine learning employed by intrusion detection systems is provided here.

A. LR (Logistic Regression)

The likelihood of an event failing or succeeding can be predicted with the help of a statistical technique known as logistic regression. In the context of binary variables, LR is utilized. Check the relationship of a specific tagged data set to classify the data. It studies the linear correlation of a given data set and then presents the nonlinearity as a sigmoid activation function.

Advantages

- LR algorithm is effortless to learn, use and understand.
- LR can be easily extended by displaying common probabilities for multiple class and range predictions.
- LR allows you to classify unknown records very quickly.
- LR provides excellent accuracy for simple data sets and is suitable for linearly segmented data sets.
- LR is a lesser amount of causing to over-fitting.

Disadvantages

- The most important limitation of LR is the concept of linearity between the dependent and independent variables.
- It predicts only inefficient tasks.
- Do not choose logistic regression if few observations are available. This can lead to over-fitting.
- Logistic regression usually produces linear boundaries.

The equation for LR is referred to in Eq. 1 here, where Pb represents the probability that the event will occur between 0 and 1.

$Z = \ln(P_i / (1- P_i)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta nxn$ (1)

By multiplying the exponents on either side, we arrive at eq. 2.

$P = e(y =1|x_{\ i}) = ez/(1 + ez)=e\alpha+\beta_{\ i}\ x_{\ i} /(1+ e\alpha + \beta_{\ i}\ x_{\ i})$ (2)

B. NB (Naive Bayes)

Naive Bayesian is usually a basic Bayesian probabilistic model. It is based on the concept of strong independence. For n functions, Naive Bayes gives 2n! Independent perception [8]. The naive Bayesian method gives accurate results.

Considering the feature vector: $x_i = (x_1, x_2, ..., x_n)$ and the class variable Bk, Bayes's principle states:

Where:

For $P(Bk|x) = p(x|Bk)\ p(Bk) / (p(x))$, $k = 1, 2, ...n$ (3)

P(Bk|X) - denotes posterior probability

P(x|Bk)-represents the probability,

p(Bk) - represents the prior type probability of the square

P(X) - Represents the prior type probability of the predictor variable.

Calculating future probabilities and previous probabilities from probabilities is what we're after.

It takes a long time to recursively compute all of the probability for all possible values. Because of conditional independence, it is impossible for the probability of one attribute to affect the likelihood of another.

Conditional independence is provided as follows:

$$P (b \mid a_1, ..., a n) = (P (a_1 \mid b) P (a_2 \mid b) ... P (a_n \mid b) P (b)) / (P (a_1)) P (a_2) ... P (a_n))) \qquad (4)$$

And, the posterior probability is equated as:

$$P_{post} (b \mid a1, ..., an) \alpha P (b) \pi n (ai \mid b) \qquad (5)$$

All values are then divided by the same numerator. Gauss, Polynomial, and Bernoulli are three types of naive Bayes classifiers. It takes functions that follow a normal distribution and is optimized for Gaussian classification. Polynomials are used in various calculations. Bernoulli classification is used for binary feature vectors.

Advantages

• If the independent prediction assumptions are true, the naive Bayes classifier works with other classifiers. Other models, such logistic regression, take longer to get to the same conclusion.

• Smart and compact training data that can estimate test data. • Simple yet easy to use. It is fast and can make probabilistic predictions because of its simplicity. For linear measurements,

• The Naive Bayes scale makes use of many data points and predictive features.

• Naive Bayes is capable of dealing with binary classification problems, numerous classes, categorical and continuous data, and many other types of data structures.

Disadvantages

• Naive Bayes expressly admits that all qualities together are independent, yet this is not the case.

• Assuming there are no categories in the training set, but there are in the test data set, this model has a 0 probability and cannot be predicted.

C. DT (Decision Trees)

Decision trees are used as an auxiliary tool with possible outcomes and as a tree structure for modelling results. Provide a way to express an algorithm using conditional statements. The steps of decision-making that can lead to positive outcomes are represented by branches. Regression and classification problems are no problem with DT. The attribute's outcome is referenced in each branch [9]. The classification rule is represented by the path from the leaf to the root.

Advantages

- Rule-based decision trees don't need to be standardized or generalized.
- When dealing with non-linear factors, decision trees outperform alternative techniques based on curves.
- It is possible to use decision trees to handle missing values more effectively.
- Decision trees have a quick learning curve because there is just one tree to memorize..

Disadvantages

- The decision tree gave incorrect predictions due to over-processing. The tree is very complex as we are constantly creating new nodes to process the data. This leads to a loss of the ability to generalize.
- Decision trees do not work properly with discarded data.
- Over-fitting can lead to high contrast and inaccuracy.
- Every time a new data point is added, the tree is reconstructed and all nodes need to be recalculated and reconstructed.
- A small amount of noise destabilizes the decision tree and makes false predictions.

D. AdaBoost

Adaptive Boosting is the abbreviation for AdaBoost. Improve the performance of machine learning algorithms with the AdaBoost algorithm. To boost a weak classifier, you first need to make it stronger. Weak classifiers rely on simple thresholds for a feature to classify an object. If a characteristic's value exceeds the cut-off point, it is assumed to be positive; otherwise, it is assumed to be negative. To begin, you must define the weights of each individual sample. As indicated in Equation 6, the answer can be found.

Sample weight = (1 / Number of samples) (6)

After that use the formula of Equation 7 to compute the Gini errors for each variable.

Gini impurity=1-(true probability)2-(false probability)2 (7)

When the best partition is selected from the root node and the partition is successfully partitioned, the Gini Impurity is a technique that is built into the decision tree algorithm. Total Gini impurity for each variable can be determined by computing each node's Gini impurity. Each node's impurity weighted average is used to calculate this value. For the final step, we'll utilize the formula in Figure 8 to figure out how many operations can be performed.

Total = log ((1-general error) / common error) (8)

The sum of the weights of the incorrectly identified samples is used to calculate the overall error in this case.

Advantages

- This algorithm is straightforward, simple and fast to implement.
- AdaBoost can be used with any machine learning technique.
- Binary, text, and numeric data types are all supported.

Disadvantages

- AdaBoost is extremely noise-sensitive.
- Poor rankings lead to overtraining and low profitability.
- It is critical to assure the availability of high-quality data, as advanced technology is only being learned at a sluggish pace.

E. RF (Random Forest)

Random forest is utilized both as classification and regression. As the name suggests, RF is made up of many trees. RF includes a package of DTs in terms of classification and is considered a saving method in case of over-fitting DTs. The DT has high variance and low bias, resulting in an undesirable output. RF can focus on variables of training data along with interpretation to develop discrete decision trees and obtain overall averages for classification or focused query or regression problems [10]. When RF is used to solve regression problems, it uses Mean Squared Error (MSE). The formula for calculating MSE is shown in Equation 9.

MSE = 1/Z ∑Zi = 1 (Xi – Yi)2 (9)

The number of data points is denoted by the letter Z. The model's output is Xi, and the actual value for data point I is Yi.

When dealing with categorization data, the Gini index is used. Equation 10 shows the formula for determining the Gini index.

$$\text{Gini} = 1 - \sum c \quad (pi)2 \qquad (10)$$

Where pi denotes the class's relative frequency in the dataset and c is the number of classes.

Advantages

- Reducing over-fitting and handling classification and regression issues are two ways to increase accuracy in RF.
- RF can handle problems related to classification and regression.
- Both continuous and sequential values are well-suited to RF.
- Because RF employs a rule-based approach, data normalization is not necessary.

Disadvantages

- To generate many trees for joint production, RF needs a lot of computational power and resources, and it takes a long time to train because of the several DTs involved.
- Interpreting RF is possible, but each variable's significance cannot be assessed in full detail.

F. KNN (K-Nearest Neighbours)

The K-nearest neighbour method belongs to a class of supervised ML methods for predictive problem regression and classification. However, most of its applications can be found in classification prediction problems. KNN has no specific training phase and uses all available training data in the classification process. KNN learns the concept of feature similarity and assigns new data points based on the similarity of data points in the training set. KNN works as follows:

- Load the training and test data first.
- Select the value of the nearest data point named K, which can be an integer.
- Use an appropriate distance calculation method (such as Euclid, Manhattan and Minkowski) to calculate the distance between the test data and each row of training data as shown in Eq.11, 12, 13.

$$\text{Euclidean} = \text{Sqrt}\ (\sum_{i=1} N \ (a - b)*2 \qquad (11)$$

$$\text{Manhattan} = \sum_{i=1} N \ |a - b| \qquad (12)$$

$$\text{Minkowski} = \sum_{i=1} N \ [\ (|a - b|)q] * 1/q \qquad (13)$$

Where x, y are two variables involved.

- Sort the distance values in ascending order.

- Select the top N rows of the ordered array.
- Adjust the box to have breakpoints sorted by the most frequently occurring cells in these rows.

Advantages

- KNN is easy to understand.
- KNN is the value for both classification and regression.
- Declaration works perfectly with multilevel problems.

Disadvantages

- ANN is very memory-intensive and parallelism requires large computing resources.
- The KNN display is sensitive to the amount of data.
- In rare cases, the KNN target type may not work.
- ANN interferes with many explanatory variables.

G. Support Vector Machine (SVM)

Support vectors are a set of supervised training programs that focus on regression, classification, and externalization. SVM can be used depending on the problems encountered. SVM is unique in the way it chooses decision boundaries, using distances from adjacent data points in the class under study. The extent of the solution obtained is called the maximum super area or maximum area classifier. A straight line is generated between 2 categories by the SVM classifier. One set of data points falls in one region, whereas the second set of data points falls in another. It is used for facial identification, biometric recognition, text categorization, handwriting recognition and other applications.

This is the case in SVM, where the values of each feature are shown as points in n-dimensional space. Splitting into different categories requires finding the best hyper-level to better separate the two categories. A generalization of a plane called a super-plane can be a 2-D line, a 3-D plane, and a multidimensional super-plane [11]. Line functions in a two-dimensional space can be represented as y=mx+cy where the features of the line, which are designated as x1, x2,..., xn, can be rewritten as x1, x2,…xn. On defining x = (x1, x2) and w = (m, -1), one gets the equation w.x + c = 0 Hyper - plane position and orientation can be influenced by data points that are close to the hyper-plane, known as support vectors. As shown in the following equations 14a and 14b, we can define a hypothesis function h.

$$h\,(x_i) = +1 \qquad \text{if, } [(w_* x) + b] >= 0 \qquad (14a)$$

$$h\,(x_i) = -1 \qquad \text{if, } [(w_* x) + b] < 0 \qquad (14b)$$

SVM classifier reduces the turn of phrase to the eq. 15.

$[1/n\sum n\ \max(0,1 - y(w.x - c))] + \lambda||w||2$ (15)

Advantages

• Reliable optimization - convex optimization results in a global minimum rather than a local minimum.

• Effective use - People who have used SVMs have found them to be very good at managing both non-linear and linear data. SVMs can be used with both labelled and unlabeled data in semi-supervised learning models, but there is a small problem called transformation SVM that needs to be solved before they can be used with both types of data.

• Feature mapping - Using a simple dot product, the SVM model maps features to each other. This makes it easier to figure out how well you did at training.

Disadvantages

• Unable to process text structure - SVM cannot process text structure. Loss of sequence information will result in poor performance.

• Difficulty in selecting cores- Multiple cores makes it complicated to choose the right core. This is the main limitation of SVM.

5. Feature Selection by Boosting

Feature selection refers to the process of selecting features that are very important for prediction. It sounds trivial, but it is one of the most multidimensional problems encountered when building new ML models. The biggest challenge in machine learning is to choose the right functions as inputs to the model. The features chosen to train the model have a significant impact on efficient performance [12]. Inappropriate actions can affect the performance of the model being built and make changes a problem for data scientists. Feature selection algorithms help to overcome these shortcomings by identifying suitable features from available features without leaving out a lot of information [13]. Boosting is amongst the most important ways to learn in the last twenty years. This used to be a factor in classifying errors, but now it is being used to take priority regression. The main goal of the development was to figure out how to combine the results of several weak classifiers so that the committee would be impressed. This is part of machine learning called ensemble learning. Gradient boosting is part of this group. Ensemble methods is a subfield of machine learning which allows multiple features to be trained and predicted at the same time to get the same great results [14].

- **Bootstrap Aggregation (Bagging)**

Bagging involves training and predicting two independent tasks. When performing an exercise, sagging affects the bootstrap technique. The bootstrap technique divides the training data into different random sub-samples [15] for different iterations of the training model. To make predictions, the packaging classifier trims its output based on the prediction that got the most votes in the participant model. The bagging regression is responsible for averaging all the models to produce the output.

- **Stacking**

The stacking treats the outputs of multiple models in such a way that they have vastly different input efficiencies. For example, overlay the training data with KNN, LR and DT, then aggregate the obtained output and merge it with LR [16]. Its purpose is to reduce over-fitting and improve accuracy.

- **Boosting**

Boosting is one such way of converting weak learners into responsibilities of strong learners. Growth refers specifically to trees. Gradually increase the number of iterations of the model and adjust it according to the weight of weaker object. This reduces model bias and greatly improves accuracy. Popular boosting algorithms are AdaBoost and XGBoost. XGBoost is a powerful gradient boosting algorithm designed to solve today's data science problems and tools. XGBoost provides tree pruning, parallelization and regularization to avoid over-fitting, and handles strings with missing values. In other words, it calculates the significance of the function for all functions and returns the last predicted value of the function [17].

Advantages of XGBoost technology:

- Easier to use, faster to run, and generally better than other algorithms.
- Could be used to resolve regression and classification issues in supervised ML.
- Efficient sorting, internal planning, and tree parallelization.
- Provides optimized memory management for large datasets.
- Provides various settings to help reduce over-fitting.

- Handle missing values efficiently.

Disadvantages:

- Suitable for small datasets.
- Not directly related to unconditional characteristics.

6. Proposed Enhanced DDoS Attack Detection Mechanism

The aim of this research is to propose and build the perfect model for accurately detecting DDoS attacks. There are many systems available for DDoS prevention. A defence can be made by any of the existing methods after accurately detecting attacks once. The most important and initial tasks are fast and accurate identification of DDoS attacks in network security domain. So, the proposed work is limited to DDoS detection only. A machine learning classifier-based model is proposed in this paper. KNN, random forest, Adaboost and Gboost. Classifiers in this approach operate autonomously of one another to create a new data model. The primary goal of this research is to evaluate the ability of machine learning algorithms to identify network assaults and threats. The main methods of machine learning include the following steps:

- Feature Engineering: Feature selection is central to a growing understanding of systems. The included modes' features have shown to be more accurate than those of the excluded modes. Denial-of-service (DDoS) assaults can't reverse the profile feature's effects, therefore it's essential for all attacks.
- Select the appropriate machine learning algorithm. (For example, more complex or faster classification or regression algorithms)
- Train and evaluate your model. (Evaluate and select the most efficient model for different algorithms.)
- Use trained models to classify or predict unknown data.

The figure 1.1 shows the proposed approach for detecting malicious behaviour and DDoS threats to the system. The system process looks like this:

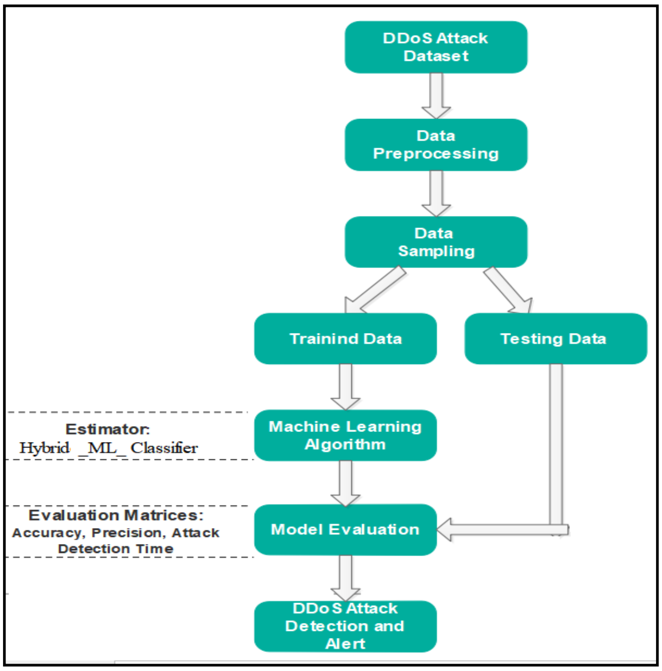
DDoS Attack Dataset
Data Preprocessing
Data Sampling
Trainind Data
Testing Data
Estimator: Hybrid _ML_ Classifier
Machine Learning Algorithm
Evaluation Matrices: Accuracy, Precision, Attack Detection Time
Model Evaluation
DDoS Attack Detection and Alert

Figure 1.1 Proposed Enhanced DDoS Attack Detection Mechanism

i. Data collection

Datasets (e.g. Kaggle Dataset, Google Dataset).

ii. Pre-processing of data

Process the saved data in a uniform format. The data is converted or encoded so that machine can easily analyze it.

iii. Data sampling

Create a subset of all your data to uncover meaningful data points and to uncover patterns and trends in large datasets discovered by ML algorithms.

iv. Machine learning algorithms

A machine learning algorithm is designed to analyze data based on feature vectors. Use the proposed Hybrid_ML_Classifier algorithm to classify the data and detect attacks.

6.1 Proposed Hybrid_ML_Classifier Algorithm

```
If : Request is HTTP
    Create : Class_HTTP_Attack
    for each (H$i$) HTTP request
    Create Set (dataframe)
    where i = 1...N
  Prune the dataframe according to the features
  For each dataframe:
    Packet features Extraction: (Feature Set1, Feature Set2)
Feature Set 1{If-Match, Max-Forwards, Proxy-      Authorization, Referrer, User-Agent, From}
  Feature Set 2{Connection, Authorization, Date, Via,      Warning, Expect}
for each feature selection
     for features in range(i...vi) :
Import the pre-processed dataset file
for each subset size $Si$ where i = 1...S do
   Hybrid_ML_Classifier[(classifier1,Feature_Set1: filled=True), (classifier2, Feature_Set2: filled=True)]
Use the trained the model to predict the class variable values for the training data
Train the Hybrid_ML_Classifier using (classifier1, Feature_Set1),
Train the Hybrid_ML_Classifier using (classifier2, Feature_Set2)
Use the trained model to predict the class variable values for the testing data
Test the Hybrid_ML_Classifier using (classifier1, Feature_Set1),
Test the Hybrid_ML_Classifier using (classifier2, Feature_Set2)
Return (Accuracy, Precision, Detection_Time)
Analyze Retune (Parameters)
      If (Traffic_Abnormality = true)
          HTTP Flood DDoS Attack Detected
Store the entire results
Increment Counter
Iterate (0 to n)
```

v. Evaluation

The estimation parameters accuracy, accuracy, and attack detection time are used to assess the proposed model. Each classifier's performance is assessed in terms of DDoS attacks classifier by 4 metrics namely, F-measure, accuracy, recall and precision [31].

The **Accuracy** is specified by the equations given below, which provides a value between 0 and 1:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Precision is the ratio of data classified as an attack accurately and the total amount of data classified as attack.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall is the proportion of data that is appropriately classified as an attack to the total number of attacks in the data.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The **F-measure** is a proportion of recall and weighted average of precision.

$$\text{F} - \text{measure} = 2 \times \frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}$$

True Negative (TN):- The data is classified as standard data.

True Positive (TP):- The irregular data classified as an attack.

False Negative (FN):- The irregular data is classified as normal data.

False Positive (FP):- The data classified as attack.

The Confusion Matrix:

Actual Class

		Attack	Normal
Predicted	Attack	TP	FP
Class	Normal	FN	TN

vi. Outcomes

Objective of the study is to propose an accurate DDoS attack detection mechanism using Hybrid_ML_Classifier ensemble technique. The final results will be presented in different formats based on evaluation parameters discussed above.

7. Conclusion

This paper addresses potential problems and issues with machine learning techniques. Confidentiality, integrity and availability are the three major threats to all networks. Availability makes data or services available as needed. The biggest threat to service availability is DDoS attacks. A hybrid approach has been proposed using different classifiers with the aim of using machine learning classification algorithms to address security concerns and detect DDoS attacks more accurately and quickly. This proposed method is under development phase and improved results are expected to generate after complete implementation. This study provides various machine learning based method for detecting DDoS threats in network systems. It is found in previous research and literature reviews; there is a tremendous range and scopes are available for detection of DDoS attack using machine learning techniques and algorithms, which can be used to detect DDoS attacks accurately and quickly.

REFERENCES

[1] A. Alsirhani, S. Sampalli, and P. Bodorik, DDoS attack detection system: Utilizing classification algorithms with apache spark," in 2018 9th IFIP International Conference on New Technologies, Mobility and Security.IEEE, 2018, pp. 1.

[2] O. Osanaiye, H. Cai, K.-K. R. Choo, A. Dehghantanha, Ensemble-based feature selection method for ddos detection incloud computing,"EURASIP Journal on Wireless Communications and Net-working, vol. 2016,1, pp.1{10}

[3] O. A. Wahab, J. Bentahar, H. Otrok, and A. Mourad, Optimal load distribution for the detection of vm-based ddos attacks in the cloud," IEEE transactions on services computing, vol. 13, no. 1, pp. 114.

[4] R. V. Deshmukh and K. K. Devadkar, \Understanding ddos attack & its e_ect in cloud environment," Procedia Computer Science, vol. 49, pp. 202.

[5]. Nagesh K., Sumathy R., Devakumar P., Sathiyamurthy K.(2016). A survey on denial of service attacks and preclusions. In: International conference on informatics and analytics, p. 118.

[6]. Lucian Constantin (2017) Hackers Use Thousands of Infected Android Devices in DDoS Attacks, https://www.forbes.com/sites/lconstantin/2017/08/28/hackers-use-thousands-of-infected-android-devices-in-ddos-attacks.

[7]. Arpitha K.S, S. Koushik Reddy, Punith Babu G.U (2020). Ddos Attacks Using ML, Journal of Xi'an University of Architecture & Technology, Vol XII, Issue IV, pp: 3380.

[8]. Hussain, J. and Lalmuanawma, S. (2016) ‘Feature Analysis, Evaluation and Comparisons of Classification Algorithms Based on Noisy Intrusion Dataset’, Procedia ComputerScience,92,pp.188–198.doi: 10.1016/j.procs.

[9]. Khan, F. A. and Gumaei, A. (2019) A Comparative Study of Machine Learning Classifiers for Network Intrusion Detection, Springer International Publishing. doi: 10.1007/978-3-030-24265-7_7.

[10]. Farnaaz, N. and Jabbar, M. A. (2016) ‘Random Forest Modeling for Network Intrusion Detection System’, Procedia Computer Science, 89, pp.213–217. doi: 10.1016/j.procs.2016.06.047.

[11]. Khraisat, A. et al. (2019) ‘Survey of intrusion detection systems: techniques, datasets and challenges’, Cybersecurity, 2(1). doi: 10.1186/s42400-019-0038-7.

[12]. Costa, K. A. P. et al. (2019) ‘Internet of Things: A survey on machine learning-based intrusion detection approaches’, Computer Networks, 151.

[13]. Liu, J. and Xu, L. (2019) ‘Improvement of SOM classification algorithm and application effect analysis in intrusion detection’, in Advances in Intelligent Systems and Computing. Springer Verlag, pp. 559–565.

[14]. Karatas, G., Demir, O. and Sahingoz, O. K. (2020) ‘Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset’, IEEE Access, 8, pp. 32150–32162.

[15]. Yao, H. et al. (2019) ‘MSML: A novel multilevel semi-supervised machine learning framework for intrusion detection system’, IEEE Internet of Things Journal, 6(2), pp. 1949–1959.

[16]. Gao, X. et al. (2019) ‘An Adaptive Ensemble Machine Learning Model for Intrusion Detection’, IEEE Access, 7, pp. 82512–82521. doi:10.1109/ACCESS.2019.2923640.

[17] Wei, P. et al. (2019) ‘An optimization method for intrusion detection classification model based on deep belief network’, IEEE Access, 7, pp.87593–87605.

[18]. W. Zhijun, X. Qing, W. Jingjie, Y. Meng and L. Liang, “Low-Rate DDoS Attack Detection Based on Factorization Machine in Software Defined Network” in IEEE Access, vol. 8, pp. 17404-17418, 2020.

[19]. S. Dong and M. Sarem, “DDoS Attack Detection Method Based on Improved KNN With the Degree of DDoS Attack in Software-Defined Networks” in IEEE Access, vol. 8, pp. 5039-5048, 2020,

[20]. Abdulhammed, R., Faezipour, M., Abuzneid, A., & Alessa, A., 2018, Effective features selection and machine learning classifiers for improved wireless intrusion detection. In 2018 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1-6) IEEE.

[21]. S. Alzahrani and L. Hong, ”Detection of Distributed Denial of Service (DDoS) Attacks Using Artificial Intelligence on Cloud” 2018 IEEE World Congress on Services (SERVICES), San Francisco, CA, 2018, pp. 35-36.

[22]. S. Das, A. M. Mahfouz, D. Venugopal and S. Shiva, ”DDoS Intrusion Detection Through Machine Learning Ensemble” 2019 IEEE 19th International Conference on Software Quality Reliability and Security Companion(QRSC), Sofia, Bulgaria, 2019, pp. 471-477,

[23]. S. Nandi, S. Phadikar and K. Majumder, ”Detection of DDoS Attack and Classification Using a Hybrid Approach” 2020, 3rd ISEA Conference on Security and Privacy (ISEA-ISAP), Guwahati, India, 2020, pp. 41-47,

[24]. Belouch, M., El Hadaj, S., & Idhammad, M. (2018). Performance evaluation of intrusion detection based on machine learning using Apache Spark. Procedia Computer Science, 127, 1-6.

[25]. Gulla, K. K., Viswanath, P., Veluru, S. B., & Kumar, R. R. (2020). Machine learning based intrusion detection techniques. In Handbook of computer networks and cyber security (pp. 873-888). Springer, Cham.

[26]. Iman, A. N., & Ahmad, T. (2020, February). Improving Intrusion Detection System by Estimating Parameters of Random Forest in Boruta. In 2020 International Conference on Smart Technology and Applications (pp. 1-6). IEEE.

[27]. D. Erhan and E. Anarim, ”Hybrid DDoS Detection Framework Using Matching Pursuit Algorithm,”inIEEEAccess,vol.8,pp.118912-118923,2020, [28]. S. Sahu and A. Verma, ”DDoS attack detection in ISP domain using machine learning,” 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-4, [29].Kachavimath, A. V., Nazare, S. V., & Akki, S. S. (2020) Distributed Denial of Service Attack Detection using Naïve Bayes and K-Nearest Neighbor for Network Forensics. In 2020,2nd International Conference on Innovative Mechanisms for Industry Applications(ICIMIA)(pp.711-717). IEEE.

[30]. Y. Su, X. Meng, Q. Meng and X. Han, “DDoS Attack Detection Algorithm Based on Hybrid TrafficPredictionModel”2018IEEEInternationalConferenceonSignalProcessing, Communications and Computing,Qingdao,2018pp.1-5.

[31]. S. S. Priya, M. Sivaram, D. Yuvaraj and A. Jayanthiladevi, ”Machine Learning based DDOS Detection,” 2020 International Conference on Emerging Smart Computing and Informatics, Pune, India, 2020, pp. 234-237.

[32]. Sah, G., & Banerjee, S. (2020, July). Feature Reduction and Classifications Techniques for Intrusion Detection System. In 2020 International Conference on Communication and Signal Processing (ICCSP) (pp. 1543-1547). IEEE.

[33]. G. Kaur and P. Gupta, “Hybrid Approach for detecting DDOS Attacks in Software Defined Networks” 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, pp. 1-6, doi: 10.1109/IC3.2019.8844944.

[34]. V. Deepa, K. M. Sudar and P. Deepalakshmi, “Detection of DDoS Attack on SDN Control plane using Hybrid Machine Learning Techniques,” 2018 International Conference on Smart 2 Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2018, pp. 299-303, doi: 10.1109/ICSSIT.2018.8748836.

[35]. S. Wei, S. Dai, X. Wu and X. Han, “STDC: A SDN-Oriented Two-Stage DDoS Detection and Defence System Based on Clustering” 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference on Big Data Science and Engineering(TrustCom/BigDataSE), New York, NY, 2018, pp.339-347,doi: 10.1109/TrustCom/BigDataSE.2018.00059.

[36]. Fitni, Q. R. S., & Ramli, K. (2020, July). Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems. In 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT) (pp. 118-124). IEEE.

[37]. Alireza Seifousadati, Saeid Ghasemshirazi, Mohammad Fathian, “A Machine Learning Approach for DDoS Detection on IoT Devices” ©2021 IEEE.

[38]. Mugunthan, S. R. "Soft computing based autonomous low rate DDOS attack detection and security for cloud computing." J. Soft Comput. Paradig. (JSCP) 1, no. 02 (2019): 80-90.

[39]. Vivekanandam, B. "Design an Adaptive Hybrid Approach for Genetic Algorithm to Detect Effective Malware Detection in Android Division." Journal of Ubiquitous Computing and Communication Technologies 3, no. 2 (2021): 135-149.