International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

Prediction of Distributed Denial of Service Attacks using a New Ensemble Model

Sheha Tantuway[1], Prof. Jayesh Jain[2], Prof. Saurabh Sharma[3], Prof. Saurabh Kapoor[4]
[1,2,3,4]*Global Nature Care Sangathan Group of Institutions, Jabalpur(M.P), India*

Abstract:-- **The internet has grown to play a significant role in everyone's life in the current era. Nowadays, the majority of firms are digital, and the internet significantly contributes to their success. Due to the internet, businesses now have access to a wider market for their products. The server is a crucial element of the internet. In general, any internet-connected device is either a server or a client. Clients access the data that is stored on the server. Network Denial-of-Service A network attack on the servers is called service. This attack floods the server with internet traffic in an effort to reduce its availability. As a result, the server is unable to provide services to authorised users. The organisation suffers huge losses as a result of DDoS attacks. In this research, we discuss a method to lessen this issue based on ensemble modelling. In ensemble modelling, several fundamentally diverse models are integrated to create a single classifier model. Using an ensemble model that combines Nave Bayes, Random Forest, Multilayer Perceptrons, and Stochastic Gradient Descent, we attempt to predict the sort of DDoS assault. The accuracy rating of 98.62% has been attained. We can draw the conclusion that this method effectively prevents DDoS attacks.**

Keywords:-- **DDoS attack,http-flood, udp-flood, smurf , machine learning , ensemble model**

I. INTRODUCTION

The role of the internet in business is further strengthened by developments in communication and information technology. Internet is frequently utilised in businesses to market and promote their goods and services. British computer scientist Tim-Berners Lee came up with the idea for the internet (also known as the World Wide Web) in 1989. Since then, there has been a dramatic increase in the number of people using the internet. It is crucial to remember that the internet is a relatively new phenomenon that only began a few decades ago. The development of the World Wide Web is still in its infancy. Although the internet has undoubtedly changed a lot of things, there is still much room for improvement.

As was previously mentioned, the rise in popularity of the internet has brought both new opportunities and challenges. Making data always accessible to users, ensuring data integrity and privacy, and preventing unauthorised changes to the data are some of the major challenges for data stored online (Confidentiality). The CIA triumvirate is another name for this.

Multiple compromised hosts that are openly available provide a fictitious flood of internet traffic, which is what causes a distributed denial of service assault. The term "zombies" is also sometimes used to describe these corrupted hosts. The fundamental internet security premise is broken using distributed denial of service, according to the CIA (Confendiality , Integrity, Availability).

Typically, a single host cannot be used by an attacker to launch a Denial of Service attack. As a result, the attacker uses malware to assault the many vulnerable devices accessible online. Once these devices have been compromised, the attacker utilises them to bombard the server with internet traffic. Most of the time, those gadgets' owners aren't even aware that they've been compromised. Due to the fact that the assault is being carried out through compromised devices that appear to be legitimate clients, the attacker's identity is also kept a secret. The genuine clients are unable to access the server since it is now being used to serve the hacked devices. This results in significant losses for the company.

In this research, we use an ensemble modelling technique based on machine learning to attempt to address the following problem. The majority of publicly accessible datasets are not protected against the most recent DDoS attacks. The most recent DDoS attacks have been covered by the dataset utilised in this study. There are 28 columns and 2 million rows in the dataset. To decrease the amount of features for the prediction without compromising accuracy, feature selection has been used to the attributes.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

For various computational needs, feature engineering has been done on the values of the attributes to convert from string format to numerical representation (a hashing-based solution is detailed in depth in the Implementation section).

In order to increase accuracy while using much less training data, an ensemble-based modelling approach combining Nave Bayes, Random Forest, Multilayer Perceptrons, and Stochastic Gradient Descent has been developed.

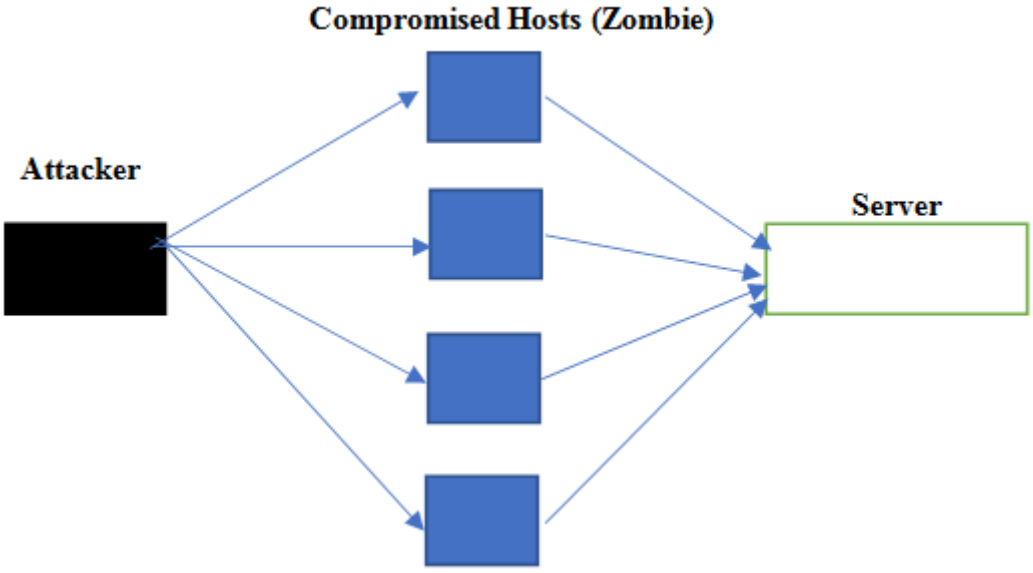

Fig. 1 Typical Structure of a DDoS attack

II. Literature Survey

The author of paper [1] used the Random Forest Classifier approach to categorise the packets as normal or anomalous. The well-known NSL-KDD dataset, with its meagre 23000 or so rows, was employed. The dataset includes 42 characteristics in total. With somewhat less attributes, similar accuracy can be attained. The complexity of the model and the time it takes to develop it both rise as the number of attributes grows. Additionally, it is a bi-classification problem, which divides the packet into the normal and anamolous classes. Our study offers a technique for precisely identifying the DDoS assault type.

The mix of supervised and unsupervised machine learning techniques is employed by the author in paper [2]. In the first step, aggregative clustering is utilised to group together packets that are similar to one another before they are manually labelled. The model is then built using this labelled data as an input. The accuracy of the model, which was built using k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Random Forest (RF), was 95%, 92%, and 96.6% respectively. This paper had an accuracy rate of about 98 percent.

The author of paper [3] employed a firewall-based strategy. The first point of entry for a packet into a server or computing device is a mechanism called a firewall. The packet is examined for authenticity using the access control list (ACL). To distinguish between analogous and regular addresses, one uses the IP addresses. When the attack magnitude increases, firewalls are not a reliable defence.

The author of article [4] used to categorise the ICMPv6 DDoS assault. The acronym ICMP stands for Internet Control Message Protocol, which is used to transmit error and control messages. They have compared the outcomes using well-known classification approaches as Neural Network, Decision Tree, Naive Bayes, Support Vector Machine, and K-Nearest Neighbors.

In paper [5], the author employs the ensemble modelling technique, in which four trained classifier models are voted upon for the final prediction. The well-known NSL-KDD dataset was the source for the study employed in this paper. The four methods utilised for classification are MLP (NN), SMO (SVM), IBK (KNN), and J48 (DT-C4.5). To aggregate the results of various models' predictions, a process known as hard voting is performed.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

Hard Voting, on the other hand, ignores the fact that each model has a varying accuracy and gives each model equal weight when determining the final result.

The authors of paper [6] suggest the RNN-based DeepDefense concept. The model does not employ machine learning-based techniques, in contrast to earlier methods that have been presented. Instead, the authors divided ongoing network traffic into sequences of network packets by extracting 20 attributes from packet headers and applying sliding windows. The last packet in a particular sequence of packets is classified by the model as either legitimate or attack traffic. In comparison to conventional machine learning methods, their suggested deep learning model delivers lower error rates. reliance on packet-by-packet, though.

III. Methods And Materials

The following chapter is further divided into multiple sub-chapters where each sub-chapter describes the multiple processes for achieving the final result.

A. Data

The data is taken from the public dataset [7]. The data consists of the modern DDoS (Denial Distributed of Service) attacks. The attacks include the HTTP Flood, UDP Flood, SIDDOS, Normal, Smurf. These attacks do not cover the entire DDoS attacks but contain the most common ones. The data is present in the text format with the values for each features separated using comma. There are around 2 million rows and 28 features. We believe we have trained the model with all possible scenario. Some of the features which do not contribute in the enhancement of the accuracy have been removed.

Fig. 2 Sample of the Structure of the Data

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

Fig. 3 Features of the dataset

B. Data Pre-Processing

The data is present in the text format and requires pre-processing. In the file, each row is taken as one whole string and then converted to a string array using split function available in python. A separate list data structure is created for every feature. The value is then put to its respective list.

C. Conversion to Numeric Value

Some of the features had values which are of string datatype and had to be converted to a numeric value. The task here is to represent the each string value uniquely using a numeric value. The reason for this is that each of the string has a unique meaning associated with it.

For example, consider the feature node_name_from which says from which type of node the packet is being sent. For this conversion, we use hashing. Hashing is the process of transforming string into a unique key and the value of the hash for a different string is completely different. The value returned by the hash function is also a string which can be then converted to a numeric value using ascii of the characters in the string. We have taken summation of the ascii value of the characters in the hash. This process makes sure that each unique string is assigned unique integer value for computation. However, for the values of the target variable we have assigned unique number asshown the figure 4.

```
for i in range(len(templist)):
    temptext = templist[i]
    hash_object = h.md5(temptext.encode())
    md5_hash = hash_object.hexdigest()
    value = 0
    for j in range(len(md5_hash)):
       value = value + ord(md5_hash[j])
    valuelist[templist[i]] = value
return valuelist
```

Fig. 3 Python Implementation for Conversion to Numeric Value

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

Fig. 4 Numeric value assigned for the values of the target variable

D. Scaling of data

Scaling of data is important for getting the maximum accuracy for any classifier. Scaling can be defined as the process of fitting the data on a specific scale. Consider an example of a data which consists of currencies like rupee and dollar. We know that 1 US dollar = 80 rupees. So therefore, for consistent results, the whole data should be converted to either dollars or rupee. If scaling is not done on the data then the classifier considers a difference in price of 1 rupee as important as a difference of 1 US dollar. This would lead to wrong results and reduce the effectiveness of the classifier.

E. Ensemble Learning

In this paper we implement a ensemble model which is combination of four classifiers which are Stochastic Gradient Classifier, Multilayer Perceptron (also known as Artificial Neural Network), Random Forest Classifier and Naïve Bayes Classifier. Brief working of each of the classifier has been given for completeness. Ensemble model is created to combine the results and improve the accuracy of the model .We combine the results of these models by using a mathematical function which is discussed below.

In the first step, we assign weights for each of the model. The weights signify how much the classifier shall contribute in giving the final prediction. Weights are the accuracy of the classifier when run seperately on the dataset. Consider the Table 1 for weights assigned to the classifiers.

Table 1

Serial No.	Classifier	Weights (Accuracy)
1	Stochastic Gradient Descent	98.50 %
2	Multilayer perceptron	98.62 %
3	Random Forest	97.23 %
4	Naïve Bayes	96.97 %

Figure 5 shows the 2 dimensional list containing the class probabilities of each data point for stochastic gradient descent..Class probabilites are nothing but the probability of the values of the target-variable given the values of the feature variable have already occurred. Now , in order to predict the final result we multiply the weight of the each classifier with the class probability.

Once it is computed for all the classes (or values) of the target variable and the one which has the maximum value is chosen as the final prediction. Consider figure 6 which is the python implementation of the above idea. We design iteration on the values (class) of the target variable (type of distributed denial of service) and calculate according the equation shown figure 6. Once the calculation is completed for all the values of the target variable, then simply the one with maximum number is chosen as the predicted value.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

```
[[0.00000826 0.01097373 0.00027594 0.98526049 0.00348158]
 [0.00000262 0.01066955 0.00022107 0.98515328 0.00395347]
 [0.00000122 0.01039839 0.00015766 0.98524586 0.00419688]
 ...
 [0.00000171 0.01053096 0.00019098 0.98517701 0.00409933]
 [0.00002233 0.01083333 0.00016394 0.98444659 0.0045338 ]
 [0.0000014  0.01058564 0.0002341  0.98541981 0.00375906]]
0.9850214790307719
```

Fig. 5 two dimensional list of class probabilities of each data point for stochastic gradient descent

Fig. 6 Python implementation of the ensemble model

For better understanding, we predict the value of the target variable for the first data point using our ensemble learning apporach. The class probabilities of the first data point for naïve baiyes figure 7, stochastic gradient descent figure 5, multilayer perceptron figure 8, random forest figure 9 is shown.

```
[[0.         0.         0.         0.99999998 0.00000002]
```

Fig. 7 Naïve bayes class probabilities forfirst data point

```
[[0.         0.00866969 0.00013761 0.98697545 0.00421724]
```

Fig. 8 Multillayer perceptron class probabilities forfirst data point

```
[[0. 0. 0. 1. 0.]
```

Fig. 9 Random forest class probabilities for first data point

Now, to predict the value of the target variable we mulitply the weight and the class probability. The f(x) represents the function of the ensemble learning and x takes all the values of the target variable. Therefore,the numeric value for the first class is f(x1) = 0.0 * 0.9697 + 0.0 * 0.9862 + 0.0*0.9723 + 0.00000826 * 0.985 and f(x1) = 0 . f (x2) = 0.0 * 0.9697 + 0.00866969 * 0.9862 + 0.0*0.9723 + 0.010973 * 0985 and f(x2) 0.0193591723. Similarly, f(x3) = 0.0040 , f(x4) =3.88593, f(x5) = 0.00758839.

Clearly, f(x4) has the maximum value and x4 is predicted as the ouput for the target variable. x4 represents Normal packet. So , the first datapoint corresponds to a normal packet.

This approach takes into account that not all classifiers have same accuracy and classifiers which have better accuracy should have more contribution in the prediction. This method has increased the accuracy of each of the classifer and given a more robust , effective and more accurate model.

F. Naïve Bayes

Naive Bayes Classifier is a probability based classifier. The main mathematical concept behind the usage of this classifier is bayes theorem.Bayes theorem is used to calculate the probability of an event A happening given that the event B has already occurred. Therefore, it can be said that the event B is the evidence and event A is the hypothesis. It is important to understand here that Naive Bayes assumes the features are independent. The presence of one particular feature does not affect the other. Since , broadly the classification problem can be further divided into either multiclass prediction or biclassification (yes or no). The mathematical formula differs for both the types. Consider the Figure 10 for bi- classification problems and Figure 11 for multiclass problem [8].

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

$$P(y|x_1,...,x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

Fig. 10. Bayes Theorem for bi-classification

$$P(y|x_1,...,x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

Fig. 11 Bayes Theorem for multiclass classification

G. Stochastic Gradient Descent

Stochastic Gradient Descent Classifier has two main mathematical concepts. The first concept is the Gradient Descent Algorithm and the other part being Stochastic. Stochastic in plain term means random. Gradient Descent is an optimization algorithm. It is used to find local manima of a differential function. Consider an example of a parabola as shown in the figure 12. In order to find the local minima dx/dy should be equal to zero. Gradient Descent takes larger step size when the value is far the from the local minima and small step sizes when the value is near the local minima. As mentioned earlier stochastic means random. However the dataset consists of multiple rows (or samples) and each row having multiple features and each contributing to the prediction of the target variable (type of Distributed Denial of Service attack) .Stochastic Gradient Descent randomly takes one feature randomly instead of iterating over all features in the samples of the dataset. This reduces the computation enormously and makes it suitable as a classifier.

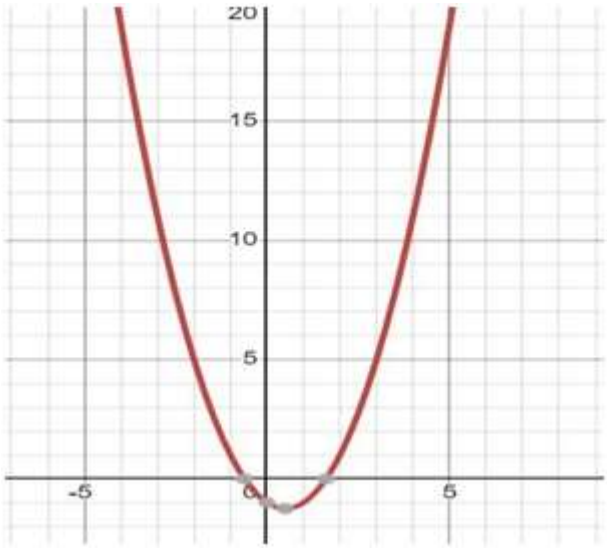

Fig. 12 Diagram of a parabola

H. Multilayer Perceptron

Multilayer Perceptron or sometimes also referred as the Neural Network or the Artificial Neural Network is based on the simple model of the brain. The builiding block of the Artificial Neural Network is the neurons. Each neuron takes weighted input signals and computes the output signal based on the activation function. There are mainly three layers in the multilayer perceptron – input layer, hidden layer, output layer. Usually , there is only one input layer,one output layer and many hidden layers. The activation function is the sigmoid function. The sigmoid activation function takes real numbers as input and converts it into value between 0 and 1 .

I. Random Forest

Random forest Classifier applies decision tree algorithm on various subsets of the dataset and combines the results for prediction of the target variable. Decision tree constructs a flow-chart like tree structure where the internal node denotes the features , the branch denotes the values that the feature holds and the leaf node consists of the values held by the target variable.

IV. Results And Discussion

The ensemble model is based on the mathematical equation as discussed in the implementation section. The ensemble model has an accuracy of 98.621 percent which is higher than all the individual classifier models. Though the difference in accuracy is in the range of 0.5 percent to 1.5 percent as compared to the famous classifiers used in this paper, however 1 % increase in accuracy on a dataset containing 2 million dataset is predicting 20,000 more rows correctly which is a very good increase. Any machine learning model is judged based on the f1-score, precision and recall score.

Precision score quantifies the number of positive class predictions that actually belong to the positive class [9] .The figure 13 is the classification report .Recall quantifies the number of positive class predictions made out of all positive examples in the dataset. F-Measure provides a single score that balances both the concerns of precision and recall in one number. Confusion matrix as shown in the figure 14.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

	precision	recall	f1-score	support
2	0.98	0.94	0.96	1352
3	1.00	0.90	0.95	66274
4	0.88	0.95	0.91	2130
5	0.99	1.00	0.99	639048
6	0.96	0.32	0.48	4217
accuracy			0.99	713021
macro avg	0.96	0.82	0.86	713021
weighted avg	0.99	0.99	0.99	713021

Fig. 13 Classification Report

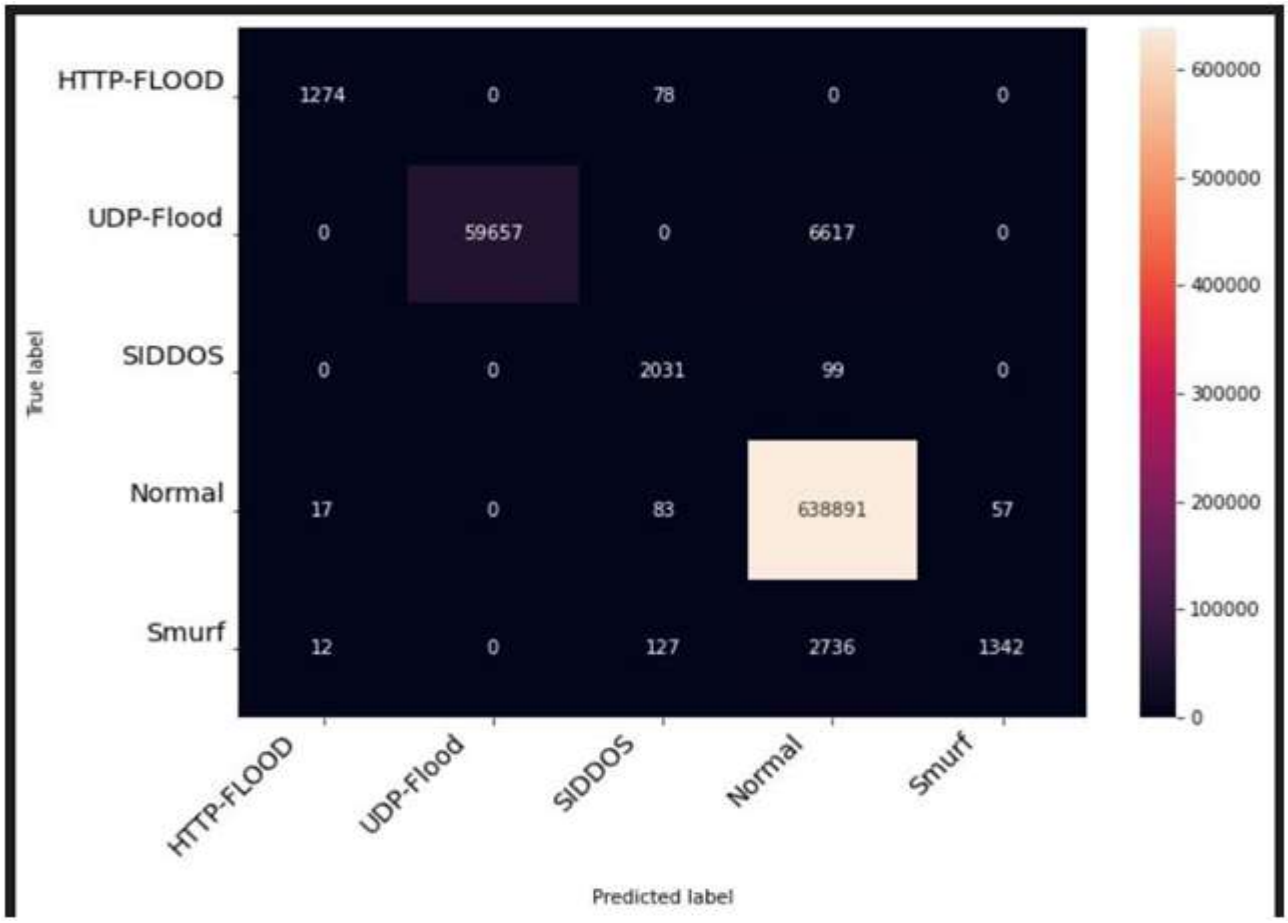

Fig. 14 Confusion matrix

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 11, November 2022)

V. CONCLUSION

The network security of numerous servers is seriously threatened by distributed denial of service (DDoS) assaults. DDoS attacks can cost businesses billions of dollars in sales because to the ever-expanding nature of online business. By overloading the server with packets, these attacks have the potential to refuse services to the legitimate user. In this essay, we offer a potential fix to lessen the impact of this issue. The accuracy of the innovative ensemble model developed in this research is 98.62%. The four well-known classifiers—Naive Bayes, Stochastic Gradient Descent, Multilayer Perceptron, and Random Forest—were used to create the ensemble model. 28 features and 2 million datasets in all are included in the data.

REFERENCES

[1] Pande, S., Khamparia, A., Gupta, D., Thanh, D.N.H. (2021),“ DDOS Detection Using Machine Learning Technique.In: Khanna, A., Singh, A.K., Swaroop, A. (eds) Recent Studies on Computational Intelligence“, Studies in Computational Intelligence, vol 921. Springer, Singapore.

[2] NG, B. A., & Selvakumar, S. (2019), “Deep radial intelligence with cumulative incarnation approach for detecting denial of service attacks”, Neurocomputing.

[3] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in Proc. ECOC’00, 2000, paper 11.3.4, p. 109.

[4] Elejla, Omar E., Bahari Belaton, Mohammed Anbar, Basim Alabsi, and Ahmed K. Al-Ani, "Comparison of classification algorithms on ICMPv6-based DDoS attacks detection." In Computational Science and Technology, pp. 347-357,Springer, Singapore, 2019.

[5] 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 (QRS-C), 2019, pp. 471-477.

[6] X. Yuan, C. Li and X. Li, "DeepDefense: Identifying DDoS Attack via Deep Learning," 2017 IEEE International Conference on Smart Computing (SMARTCOMP), 2017, pp. 1-8.

[7] Dataset used in this paper which contains 2 million rows and 28 features.[Online]. Available: https://www.kaggle.com/datasets/jacobvs/ddos-attack-network-logs?resource=download.

[8] Geeksforgeeks article describing naïve bayes classifier. [Online]. Available: https://www.geeksforgeeks.org/naive- bayes-classifiers.

[9] Definition of precision score, f1 socre and recall score. [Online]. Available: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification.