Denial of Service Attack Detection using Feature Selection and Machine Learning Algorithm

Nachiketa Ambarkhane[1], Esha Kutty[2], Priyanka Rathod[3], Manali Karande Patil[4], Shraddha R. Khonde[5], Deepali D. Ahir[6]

***Abstract*—**Denial of Service attack is a rapidly increasing threat in today's Internet World. A denial of service (DoS) attack is a type of Cyber-attack in which the offender aims to deny the services on a network or a server by flooding the traffic on the network or a server with unneeded requests which makes it incapable to serve requests from permissible users. DoS attack tries to shutdown traffic flow to and from the targeted system. The IDs are overflowing with an abnormal amount of traffic, which the system can't handle and shuts down to protect itself. Thus, this prevents the traffic from visiting the network. This document is intended to give a detailed description of how to determine those inherent features which could help to investigate the DoS attack. Such that we can diminish time in detecting the DoS attack and alleviating it in real-time.

***Index Terms*—DoS Detection, Weka Tool, Wireshark, Machine Learning, Features Selection, CICIDS Dataset.**

1. INTRODUCTION

Due to high growth of computer network, all the computer suffers from security susceptibilities which are exhausting and expensive to be solved by the manufactures. As we know that the number of hacking and intrusion detection incidents is increasing year by year as technology rolls out, unfortunately, in today's interconnected digital world there is no place to hide.

A DoS attack is one in which, a multitude of computer systems attack a selected target, thereby causing a denial of service for permissible users of the targeted system. The flood of incoming traffic to the targeted system through the network essentially forces it to shut down, thereby denying service to permissible users.

As the Machine Learning scope is increasing day by day, there are many areas where researchers are working towards modernizing the world for the future. Machine learning has involved computers so they can perform tasks without being explicitly programmed. To maintain security of the server and other online transactional services, it is important to secure it from DoS attack.

The system can be extended to be a true Intrusion Detection System(IDS) for detecting any kind of harmful attack on the users machine. Intrusion Detection explores the various techniques used to detect attacks as they occur. Feature selection is also one of the main problems for machine learning and data mining. The objective of the feature selection is to select the best features among others from the dataset under a certain evaluation criterion.

During a DoS attack, the target is either flooded with a high number of data packets or by sending the target information that triggers a crash. Hence DoS attack is classified in two forms-Flooding Attacks and Crash Attacks.

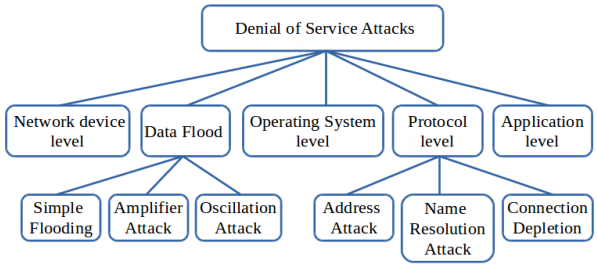

Fig. 1: DoS Attack Classification

The higher version of the DoS attack is the Distributed Denial of Service(DDoS)attack. A DDoS attack occurs when multiple computers system manages to synchronize DoS attack to a single target. The main difference between them is that in a DoS attack the system is being attacked from one location and in DDoS, the target is attacked from many locations at once.

To overcome the effects of the DoS attack various techniques have been explored during this research. Among these, data mining classification techniques are a valuable tool to identify the various threats and various attacks.

2. MOTIVATION

TYPES OF DOS ATTACK:

1. **Hacktivism:** Attacking Site due to political reason.

2. **Commercial Benefit:** They do this because of trival of competition.

3. **Cyber Welfare:** Attacking is done buy another country to compromise with infrastructure and cause embarrasment.

EFFECTS OF DOS ATTACK:

1. Due to overwhelming data packets on the server, the performance of the network becomes slow, and then it becomes difficult to access and open files
2. Due to overwhelming data packets on the particular website, the website goes down.

3. LITERATURE SURVEY

Zhiyuan Tan et al.[6]demonstrated that for receiving accurate network traffic characterisation, the hidden correlations between the features of network traffic were extracted by Multivariate Correlation Analysis technique. The system also recognises individual attack records hidden. Finally, the system proposes the principle of Earth Mover's Distance (EMD) and object shape recognition in the design for DoS attack detection.

S.R.Khonde V.Ulagamuthalvi [15] proposed novel hybrid architecture for intrusion detection systems. The authors used feature selection techniques for reducing the number of fea- tures. The feature selection used is based on the average probability score of each feature. The features having less AP score are removed from the set used for training and testing classifiers. Performance parameters used by authors are true- positive, true-negative, and accurate. Authors make use of various semi-supervised classifiers for intrusion detection. All classifiers used the NSL KDD dataset for intrusion detection. With experimental results, the authors proved that theaccuracy of the hybrid system increased by 10

Suad Mohammed Othman et al. [16] introduced the Spark-Chi-SVM model to deal with Big Data for intrusion detection. The proposed model used Spark Big Data Tool which can process and analyze data. KDD was used as a dataset and SVM was used to classify data whether it was an attack or not. The results demonstrated the model in the reduction of false positives rate and have high performance.

Narmeen Zakaria et al. [8] proposes an adaptive mechanism for the detection and mitigation of DDoS attacks in a large- scale network. They have provided an in-detail survey and study of SDN-based DDoS attack detection as well as mitiga-tion mechanisms and classified them concerning the detection techniques. They proposed and demonstrated an SDN-driven DDoS Defense Framework for the characteristics of SDN in network security. The paper also focuses on open research future research, challenges, and recommendations related to SDN-based DDoS detection.

Shraddha Khonde and V.Ulagamuthalvi [12] used Random Forest supervised algorithm for intrusion detection. Kdd dataset is used for training and testing. For feature selection probability score is used for calculating score of each feature. Depending on probability and Gini index features having highscore are selected for testing. Reduced features are passed to random forest classifier. Random forest works in distributed manner. Random forest algorithm gives 96 percent of accuracy as per authors. Authors used 50 random forest trees for increasing accuracy and reducing false alarm rate.

Manjula Suresh and R. Anitha [13] have used Informa- tion gain feature selection mechanisms and chi-square for extracting the useful attributes for higher accuracy. With the selected attributes, several machine learning models, like K-means, Fuzzy c-means clustering, Navies Bayes, SVM, and KNN are used for efficient detection of DDoS attacks. Their experimental results showcase the Fuzzy C-means clustering to give better accuracy in the identification of DDoS attacks. Meejoung Kim [14] proposed the application of supervised learning algorithms, a basic neural network and a long short-term memory recurrent neural network, to three different network traffic including DDoS attacks. LSTM RNN and BNN are used for feature extraction. The effects of these methods and hyper-parameters for machine learning are surveyed. The values representing attack characteristics are extracted from data-sets and are pre-processed by two methods. Binary clas- sification and two optimizers are used. Some hyper-parameters are obtained for speedy and accurate detection.

Khonde S Ulagamuthalvi [7] In this paper, the authors used a combination of supervised and unsupervised classifiers for the intrusion detection system. Authors used various feature selection techniques to improve the performance and accuracy of intrusion detection. Authors reduced features up to 7 from 42 of the KDD data-set to improve system performance. The architecture proposed by the authors is based on the hybrid pipeline structure of classifiers. In total seven classifiers are used for testing system performance. In this paper, the authors proved that system performance increases and reduce the false alarm rate. Data-set used is NSL KDD.

4. EXPERIMENTATION

A. Proposed Methodology

There are many Intrusion Detection System datasets, the dataset we have selected for detection of DoS attacks is CIC IDS 2018 dataset. We are detecting DoS attacks using the packet capturing technique. For capturing packets we can use Wireshark. Wireshark is the network protocol analyzer used to capture data packets and we can also analyze it, also can create statistics and graphs of our network traffic. Network traffic is recorded on the victim machine using Wireshark.

For detecting DoS attacks we can make use of different machine learning algorithms. To run various algorithms, we can use the weka tool. Weka tool is an open-source machine

learning software and has a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, association rules, visualization, classification, regression, clustering. A big benefit of using the weka platform is a large number of supported algorithms.

The dataset includes DoS attacks and benign network flows. The attacks used for the generation of data are Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet, and DDoS. The dataset involves about eleven criteria such as attack diversity, labeling, complete capture, complete interaction etc. The network topology used for capturing the network traffic includes most of the network devices and the dataset is complete capture.

Network protocols e.g. HTTP, HTTPS, SSH, FTP, SMTP, IMAP, and POP3 have been used for the generated traffic to make it more real. Feature selection as the name suggests, is the selection of the features out of the total available attributes or features in order to reduce the computational latency and complexity.

It is three types:

1) Filter methods

2) Wrappers Methods

3) Embedded methods.

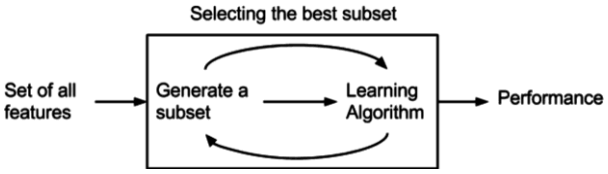

Fig. 2: Feature Selection

In filter methods, correlation of the features with the dependent variable as certains their relevance while wrapper methods use the actual classifier to weigh the worth in subset of the attributes. Filter methods are much faster and less computationally intensive as compared to the wrapper methods as they do not involve training the models. Feature selection methods categorize into Filter, Wrappers, and Embedded Methods.

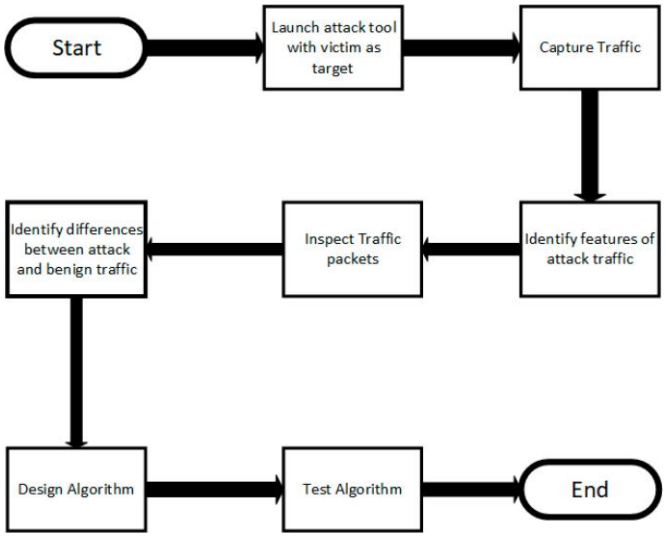

Fig. 3: DoS Attack

B. WEKA

Weka is a collection of machine learning algorithm for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, visualization. A big benefit of using the Weka platform is that it provides implementation of several algorithms, so we can select an algorithm of our choice, set the desired parameters and run it on the dataset. Then, weka would give the statistical output of the model processing. Various models can be applied on the same dataset.

C. ALGORITHM

There are four types of machine learning algorithm namely Supervised Learning, Unsupervised Learning ,Semi-supervised Learning ,Reinforcement Learning.

This is the list of commonly used machine learning algorithms. These algorithms can be applied almost to any data problem:

1. Linear Regression
2. Decision Tree
3. Naive Bayes
4. KNN
5. K-means
6. Random Forest
7. SVM

Random Forest - Random Forests are popular machine learning algorithm and it is use to solve complex problems. They are often precise, do not require feature scaling, categorical feature encoding, but need little parameter adjusting. They are more understandable than other complex models. A random forest consists of multiple random decision trees. Two types of randomnesses are built into the trees:

1] Each tree is built on a random sample from the original data.

2] At each tree node, a few number of features are randomly selected to generate the best split.

The reason that the random forest model works so well is: A large number of comparatively unrelated models (trees) operating together will defeat any of the individual component models. So in random forest, trees are not only trained on different sets of data but also use different features to make decisions. It reduces over fitting in decision trees and makes it more efficient.

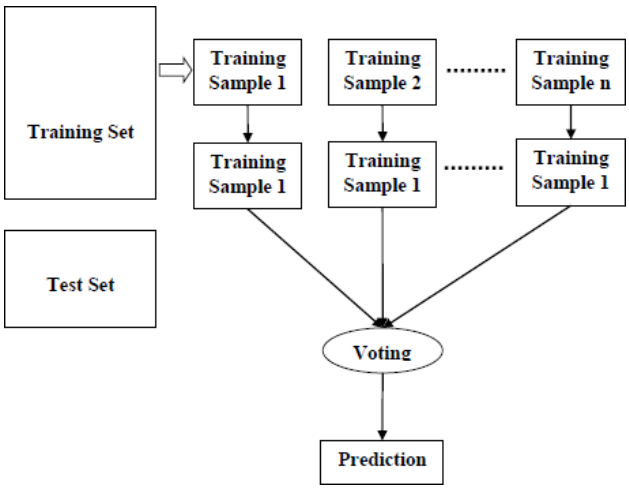

Fig. 4: Random Forest Classifier

K-Nearest Neighbor-K Nearest Neighbour is based on the Supervised Learning technique. K-NN algorithm identifies the similarity between the new set of data and the available set of data and put the new case into the category that is most similar to the available case categories. This means when new data appears then it is compared with the available cases and then it can be easily classified into a good suite category by using machine learning algorithm K-NN. K-NN algorithm can be used for both Regression and Classification but mostly it is used for Classification problems. K-NN is a non-parametric algorithm, which means it is computationally slower but make fewer assumptions about the underlying data.

K-NN algorithm stores the dataset and then splits it into training dataset and testing dataset and while in traning phase when it gets a new data then it classifies the data into the available category which is similar to the new data. Since the K-NN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm. K-NN is very easy to implement. There are only two parameters required to implement K-NN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.) three methods namely-

Filter Method - In this method, the features are filtered based on general characteristics of the dataset. It is performed without any predictive model. It is faster and usually better approach when the number of features are huge.

Example- Correlation, Chi-square, Information Gain.

Wrapper Method - The feature selection algorithm exists as a wrapper around the predictive model algorithm. It is computationally expensive. Example- Forward Selection, Backward Elimination, Stepwise Selection.

Embedded Method - Feature selection process is embedded in the learning or the model building phase. It is less computationally expensive than wrapper method. Example- Regularization methods such as LASSO, Ridge regression, Elastic net.

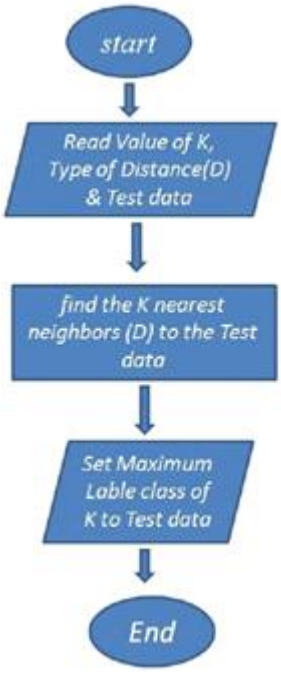

Fig. 5: K-NN Classifier

D. FEATURES SELECTION

Feature Selection means selecting and retaining only the most important features in the model. Feature Selection is important as it simplifies the model, reduces training time, improves accuracy of the model, avoids overfitting, avoids error increases with the increase in the number of features i.e avoids curse of dimensionality.

The purpose of feature selection is to select a feature subset that performs the best under a certain evaluation criterion. It has

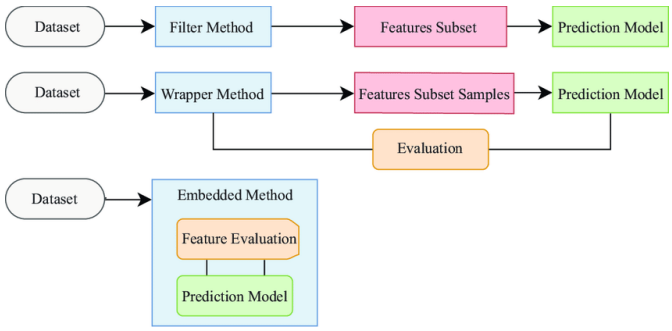

Fig. 6: Feature selection Methods

5. CONCLUSION

In this paper, main purpose is to improve the avail- ability of many modern machine learning detection meth- ods. Generally, using different services DoS attacks target one host or server by generating numerous venomous packets. Before it crashes the system or chokes the network it is important to detect a DoS attack. The aim of the proposed method is to determine those potential features which could help security administrators to investigate the DoS attack. Such that we can reduce time in detecting the DoS attack and attenuating it in a real time.

ACKNOWLEDGMENT

We are really grateful because we managed to complete our project "Denial Of Service Attack Detection Using Feature selection and Machine Learning Algorithm" within the time given by our guide.It gives us great pleasure and satisfaction in presenting this project .

I would like to express my deep sense of gratitude towards my group members who helped me a lot in improving and polishing my knowledge.

I have furthermore to thank my project guides Prof. S.R.Khonde and Prof.D.D.Ahir to encourage me to go ahead and for continuous guidance. I would also like to thank them for their assistance and guidance in preparing the report.

I would like to thank all those, who have directly or indirectly helped us with the completion of the work during this project.

Nachiketa Ambarkhane Esha Kutty Priyanka Rathod Manali Patil

REFERENCES

1) Manjula Suresh and R. Anitha,"Evaluating Machine Learning Algorithms for Detecting DDoS Attacks",: CNSA 2011, CCIS 196.

2) Meng Wang, Yiqin Lu, Jiancheng Qin, "A dynamic MLP-based DDoS attack detection method using feature selection and feedback".

3) Francisco Sales de Lima Filho, Frederico A. F. Silveira

4) ,Agostinho de Medeiros Brito Junior, Genoveva Vargas- Solar,and Luiz F. Silveira ,"Smart Detection: An Online Approach for DoS/DDoS Attack Detection Using Machine Learning".

5) Gupta, Animesh,"Distributed Denial of Service Attack Detection Using a Machine Learning Approach", doi:10.11575/PRISM/32797

6) Ahmad Riza'ain Yusof,Nur Izura Udzir,Ali Selamat,Hazlina Hamdan,Mohd Taufik Abdullah,"Adaptive Feature Selection for Denial of Services(DoS) Attack".

7) Zhiyuan Tan, Member, Aruna Jamdagni, Xiangjian and Priyadarsi Nanda (2014)," Detection of Denial-of-Service Attacks Based on Computer Vision Techniques"

8) S.R.Khonde V.Ulagamuthalvi(2019):Ensemble-based semi-supervised learning approach for a distributed intrusion detection system,Journal of Cyber Security Technology,DOI:10.1080/23742917.2019.1623475,Taylor and Francis.

9) Narmeen Zakaria Bawany, Jawwad A. Shamsi, KhaledSalah (2017), "DDoS Attack Detection and Mitigation Using SDN: Methods, Practices, and Solutions"

10) Mouhammd Al-kasassbeh, Ghazi Al-Naymat, AhmadHassanat, Mohammad Almseidin (2016)," Detecting Distributed Denial of Service Attacks Using Data Mining Techniques" T. Subbulakshmi, K. BalaKrishnan, S. Mercy Shalinie,D. AnandKumar, V.Ganapathi Subramanian, and K. Kan-nathal(2011)," Detection of DDoS attacksusing EnhancedSupport Vector Machines with real time generated dataset"

11) Mouhammd Al-kasassbeh, Ghazi Al-Naymat, Ahmad Hassanat, Mohammad Almseidin (2016)," Detecting Distributed Denial of Service Attacks Using Data Mining Techniques"

12) Shraddha Khonde and V.Ulagamuthalvi (2019)," Fusion of features selection and Random Forest for an Anomalybased intrusion detection system, Journal of Computationaland Therotical Nanoscience", Volume16,PP3603-3607

13) Manjula Suresh and R. Anitha(2011)," Evaluating Ma- chine Learning Algorithms for Detecting DDoS Attacks"

14) Meejoung Kim : (2019),"Supervised learning-based DDoSattacks detection: Tuning hyperparameters"

15) Khonde, S U lagamuthal viVenugopal(2019),Hybrid Architecture for Intrusion detection system, Ingenieriedes Systemesd' Information Volume24,No.1,PP–19- 28February2019.

16) Suad Mohammed Othman, Fadl Mutaher Ba- Alwi, Nabeel T. Alsohybe, Amal Y. Al-Hashida(2018), " Intrusion detection model using machine learning algorithm on Big Data environment"