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], *S. R. Khonde*[5], *D. D. Ahir*[6]

***Abstract*—A Denial of Service attack that is DoS attack is meant to shut down a machine or network, making it recondite to its deliberate users. The concept is to send more traffic to a network address than the programmers have built the system to handle. The primary goal of a Denial of service attack is not to slither information but to slow or take down a web site. Because of their distributed nature it is extremely difficult to defend from this attack. To differentiate permissible Web traffic from requests that are part of the DDoS attack is difficult. The best way to detect and identify a DoS attack would be via network traffic monitoring and analysis. Network traffic can be monitored via a firewall or intrusion detection system. To get rid of DoS attack we have the intrusion detection systems but we need to maintain the performance of the intrusion detection systems. To maintain the performance of intrusion detection system we have used random forest classifier and information Gain (IG) model. Random Forest (RF) is an ensemble classifier and performs well as compared to other traditional classifiers for effective classification of attacks. By use of optimal feature subset selection using Information gain ,intrusion detection system is made fast and efficient . To evaluate the performance of our model, we conducted experiments on CICIDS2018 dataset.**

***Index Terms*—DoS Detection, Random Forest, Information Gain, Machine Learning, Features Selection, CICIDS Dataset.**

I. Introduction

Due to the use of network based technologies and the sensitive information in the network, network security is getting more importance. Many security technologies are developed like Intrusion prevention, information encryption and access control to protect the network based system but still they are not enough to detect many intrusions.

There is a serious problem for computer scientists and practitioners for detection and prevention attacks and it have become a major focus of as computer attacks have become an increasing threat to commercial business as well as our daily lives. Intrusion detection system is intending to monitor the events in a system or network by determining whether is an intrusion or not. It also monitors the network traffic for suspicious activity and alert the network or system administrator about those attacks when occurred. The objective of this system intends to cover the availability, confidentiality and integrity of critical networked information system.

II. System Description

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 .We can collect the packets from the Dataset input .Then we will perform feature extraction and feature selection which is also called Preprocessing . Feature Extraction means getting useful features from existing data and Feature Selection means Choosing a subset of the original pool of features .After getting the feature selected subset of training and testing dataset we can use classification algorithm to train our model using which we can detect the malicious activities like DoS attack and its types .

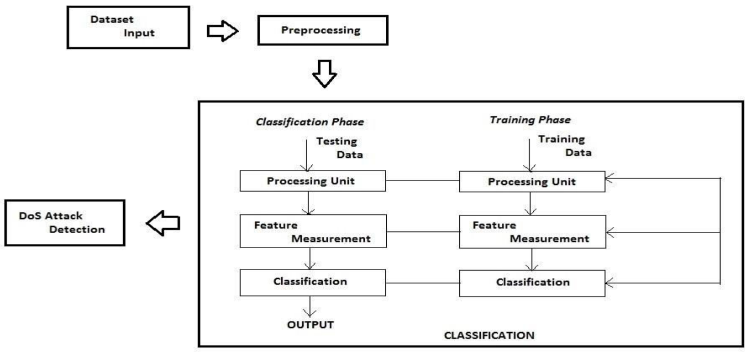

Fig. 1: Feature Selection

III. Methodology

A. *Dataset Description*

Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs) are the most important defense tools against the ever-growing network attacks. Due to the lack of testing and validation of datasets, intrusion detection approaches are suffering from compatible and precise performance evolutions.

In the proposed system, we have used Canadian Institute for Cybersecurity Intrusion Detection Evaluation Dataset(CICIDS2018). The dataset contains most up-to date common attacks, which resembles the true real-world data. There are various features such as source and destination

IPs, destination port,flow duration etc.

B. Preprocessing

To transform the raw data in a useful and efficient format data preprocessing is used. As our dataset contains minus infinity values, infinity values, and missing values so by using imputer library we have converted or replace this values with the column mean. Imputer library is used to replace the missing values in the numeric feature with some meaningful value like mean, median or mode.

C. Feature selection

Feature Selection means selecting and retaining only the most important features in the model. Feature Selection plays a pivotal role as it simplifies the model. It facilitates the machine learning algorithm to train faster. It reduces the complexity of a model and makes it easier to interpret. It improves the accuracy of a model if the right subset is chosen. It reduces overfitting problem as less redundant data means less scope to make decisions based on noise. It improves accuracy as it avoids error increases with the increase in the number of features i.e avoids curse of dimensionality .

The feature selection technique we have implemented in our project is Information gain. Information Gain is well suitable feature selection technique used in Intrusion Detection System (IDS) research. The experiment results show that the number of relevant and significant features yielded by Information Gain affects the improvement of detection accuracy and execution time significantly. Information Gain, or IG for short, measures the reduction in entropy or surprise by splitting a dataset according to a given value of a random variable of samples. Information gain provides a way to use entropy to calculate how a change to the dataset impacts the purity of the dataset, e.g. the distribution of classes. A smaller entropy suggests more purity or less surprise. Mutual Information is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. Mutual information is always larger than or equal to zero, where the larger the value, the greater the relationship between the two variables. If the calculated result is zero, then the variables are independent.

Mutual information is often used as a general form of a correlation coefficient, e.g. a measure of the dependence between random variables. Mutual Information and Information Gain are the same thing, although the context or usage of the measure often gives rise to the different names. Notice the similarity in the way that the mutual information is calculated and the way that information gain is calculated; they are equivalent:

- $I(X ; Y) = H(X) - H(X — Y)$

and

- $IG(S, a) = H(S) - H(S — a)$

As such, mutual information is sometimes used as a synonym for information gain. Technically, they calculate the same quantity if applied to the same data.

D. Random Forest

Random Forests are popular machine learning algorithm and it is use to solve complex problems. It builds multiple decision trees and combines them together to get a more stable and accurate prediction. 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:Each tree is built on a random sample from the original data.

At each tree node, a few number of features are randomly selected to generate the best split. It is usually trained with bagging method is that a combination of learning models increases the overall result. The factor due to which random forest algorithm works significantly 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 various sets of data but also use different features to make decisions. It reduces over fitting problem in decision trees and makes it more efficient. We can understand the working of Random Forest algorithm with the help of following steps:

Step 1 First, random samples are selected from a given dataset.

Step 2 Next, this algorithm will construct a decision tree for every sample created. Then it will get the prediction result from every decision tree.

Step 3 In this step, on every predicted result, voting will be performed.

Step 4 At last, the most voted prediction result is selected as the final prediction result.

IV. Experiment Analysis

For the proposed system, the performance is measured by finding out the accuracy, f-measure, recall and precision. The accuracy is calculated by finding out the ratio between the correctly predicted attack to the total attack given as an input. In this model, the dataset is split into training and testing parts as 80 percent and 20 percent respectively. The attack classes are classified as Benign, DoS Slowloris, DoS Slowhttptest, DoS Hulk and DoS Golden Eye.

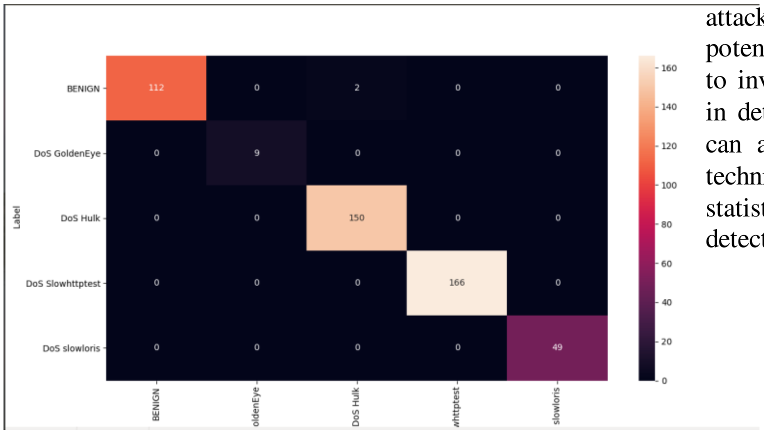

Fig. 2: Confusion Matrix

V. Result

In order to detect the attacks 3000 samples are given as an input and Random Forest model is built which is used to detect the attacks and categories them into different types . Below table shows 10 sample from the training dataset along with its actual and predicted values.

SR. No.	Actual Values	Predicted Values
2102	DoS GoldenEye	DoS GoldenEye
1399	DoS Hulk	DoS Hulk
956	DoS Slowhttptest	DoS Slowhttptest
2277	DoS GoldenEye	DoS GoldenEye
2066	DoS GoldenEye	DoS GoldenEye
2224	BENIGN	BENIGN
2013	DoS Hulk	DoS Hulk
179	BENIGN	BENIGN
741	DoS Slowhttptest	DoS Slowhttptest
1718	DoS Hulk	DoS Hulk

Fig. 3: Actual and Predicted values

For the proposed model overall accuracy is calculated using the accuracy score function of metrics library of machine learning .The overall accuracy of our model is 99.72 percent.

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 future we can also add prevention measures detection using various techniques and algorithm. And as of now model only detects statistically further we can also improve it for real time detection.

Acknowledgment

VI. Conclusion And Future Scope

Main purpose is to improve the availability of many modern machine learning detection methods. 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

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 ,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".

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

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

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