DETECTION OF Distributed Denial of SERVICE (DDoS) Attacks

Thota Srushti
Artificial Intelligence
MallaReddy University
Hyderabad, India
2011cs020369@mallareddyuniversity.ac.in

Tatipalli Vaishno Shivena
Artificial Intelligence
MallaReddy University
Hyderabad, India
2011cs020370@mallareddyuniversity.ac.in

Talagaturi Venkata Sai Meenan
Artificial Intelligence
MallaReddy University
Hyderabad, India
2011cs020371@mallareddyuniversity.ac.in

Yanamala Charan Tej Reddy
Artificial Intelligence
MallaReddy University
Hyderabad, India
2011cs020372@mallareddyuniversity.ac.in

***Abstract*— Detection of Distributed Denial of Service(ddos) attacks has become one of the main Internet security challenges today. The difficulty that is faced by the DDoS attack detection system comes from two issues. First, is how to distinguish between normal packets and attack packets. Second, is how to detect different attack categories accurately and at a low cost of resources. The right detection and prevention methods can help stop a DDoS event before it gains enough momentum to topple company networks. This type of attack consumes actual server resources and is measured in packets per second (Pps). Therefore, the detection solution must be an efficient and less complex detection solution that has low false alarms.**

Keywords—DDoS;machinelearning;SupportVector Machine; SVM; Logistic Regression; Decision Tree;

I. INTRODUCTION

A distributed denial-of-service (DDoS) attack is a malicious attempt to disrupt the normal traffic of a targeted server, service or network by overwhelming the target or its surrounding infrastructure with a flood of Internet traffic.There could be many reasons for launching DDoS attacks.These reasons may include networks disruption and financial earning.DDoS attacks can cripple networks and services by overloading servers, networks links, and network devices such as routers and switches with illegitimate packets.

These attacks can cause huge losses by either degrading the service or complete denial of service. The growing dependence of the ICT infrastructure has given a rise in the need for efficient solutions for protection against DDoS attacks.Although existing DDoS attack detection and mitigation solutions are numerous and varied, DDoS attacks continue to grow in sophistication and intensity. Fast detection and mitigation of DDoS attacks have become severely challenging as attackers continue to use new methods to launch DDoS attacks.The rising number of

DDoS attacks has made the DDoS attack detection and mitigation the topmost priority. The goal of Software Defined Networking (SDN) is to make the network more agile.

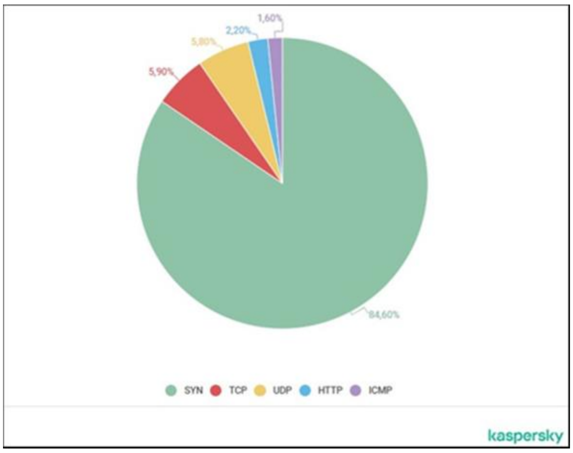


Fig.1Distribution of DDoS attacks by type for the fourth quarter of 2020

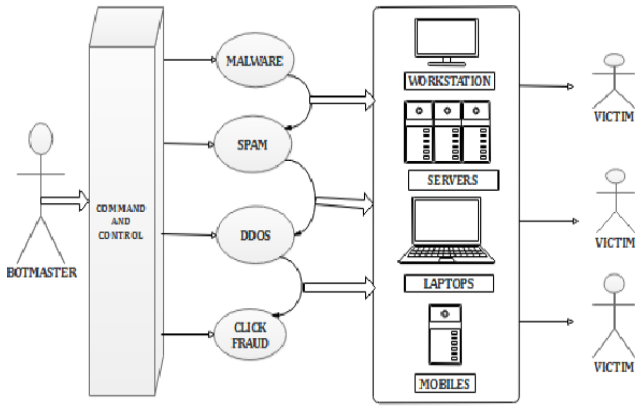
II. TASK OF TRAFFIC CLASSIFICATION

Identification and classification of network traffic is used to solve problems such as prioritizing the formation of bandwidth for individual traffic, establishing rules for network management, network security and diagnostic monitoring. In general form, the task of classifying network traffic can be formulated in the following way: receiving some characteristics of the network traffic at the input with outputting the class, which the particular type of traffic belongs to. The packet data and the various frequency characteristics both can perform as the input characteristics. As the output – the identifier of the specific application that is responsible for generating this traffic and the identifier of the traffic type, for example, VoIP (Voice over IP) traffic.

III. Research Method

In this study the dataset used was obtained from Kaggle, Dataset used is called "DDoS SDN dataset". There are 104345 rows and 23 columns. There is a one target variable called label: contains only 1(malicious) and 0(benign). Dataset contains 3 categorical and 20 numeric features(including the Label). Our goal is to classify whether the traffic is normal or not using classical Machine Learning algorithms.

UML diagram:

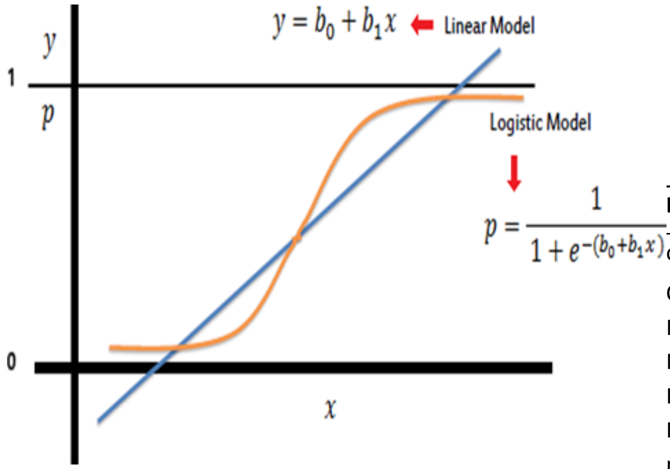


GENERALISED DDOS ATTACKS MODEL IN UML DIAGRAMS

IV. Algorithms and Techniques

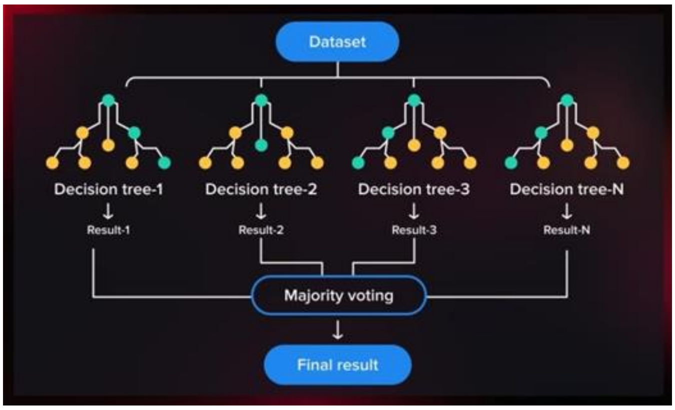
A. Logistic Regression

Logistic regression is a machine learning technique that can be used for classification problems. Logisticregression works well on the binary class label. In LR, weights are multiplied with input and pass them to the sigmoid activation function. In the proposed work, we apply LR on selected features for DDoS attack detection. The weights are optimized, using the lbfgs optimizer with C = 0.03.



B. Random Forest

RF is an ensemble-learning algorithm that grows many decision trees, independently, and combines the output. Decision trees consist of internal and leaf nodes. The selected features are used to make a decision in the internal node, and it divides the dataset into two separate sets, with similar responses. The features in an internal node are selected by the Gini impurity criterion. The feature that has the highest decrease in impurity is selected for the internal node.



The RF model is comprised of decision trees and can be used for classification or regression. In the classification case, prediction is based on a majority vote of prediction using decision trees, but in the case of regression, the result is the averaging of the tree's output. During the training phase, a training set T_i is created for each tree, based on the samples in the original training set, T, and to build each tree split, m features are randomly selected and, then, analyzed by a measure to determine which one should cause the separation.

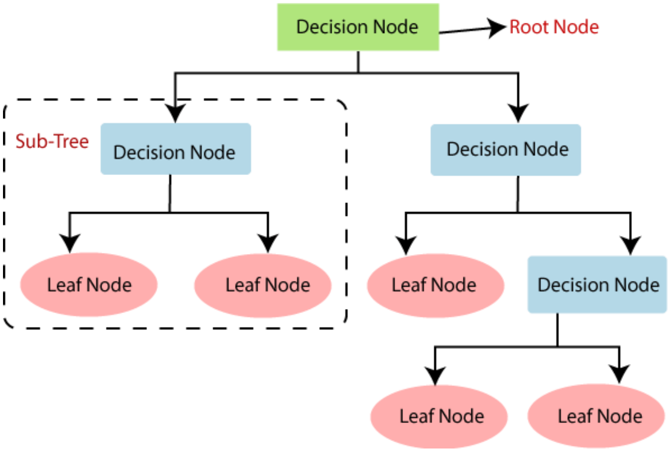
Due to this randomization, multiple trees are produced, which usually result in better prediction performance, if combined. RF models has several advantages over generally used machine learning methods, including lowest model training time, intensity to handle inconsistent datasets, classification mechanism for embedded features, and inner metrics for determining the impact of features. RF is trained for DDoS attack detection, by using different feature sets. Table 2 shows parameters used for RF.

Hyper Parameter used for RF

Hyper Parameter Name	Values
obb score	True
Criterion	Gini
Min samples	split 10
N estimators	500
Random state	1
Max features	Auto
n_jobs	-1

C. *Decision Tree*

A supervised learning algorithm is a decision tree. This method is commonly used in classification problems. It works well with both continuous and categorical attributes. Based on the most significant predictors, this algorithm divides the population into two or more similar sets. The Decision Tree algorithm first computes the entropy of each attribute. The dataset is then split using the variables or predictors with the highest information gain or lowest entropy. These two steps are repeated with the remaining attributes.



The decision tree accuracy obtained is 98.22

V. Evaluation Measures

Evaluation metrics are used to evaluate the performance of the prediction model. This study used accuracy, precision, recall, and F score to evaluate the performance of machine learning, for DDoS attack detection.

a) Accuracy:

The basic performance metric is accuracy, which is the proportion of correctly predicted observations to all observations. Accuracy is a useful evaluation measure, only when the datasets are uniform, and the false positive and false negative values are almost comparable. Accuracy tells how correctly the classifier is predicting the data points, as shown in Equation (2).

Accuracy = (TP /TP + TN + FP + FN)*100%

b) Precision:

Precision is defined as the proportion of accurately predicted positive observations to all expected positive observations. High precision is associated with a low false-positive rate. Precision gives a probability of how correctly the classifier is predicting the positive class. Precision is calculated with Equation (3).

Precision = (TP/ TP + FP)*100%

c) Recall:

Recall is defined as the ratio of accurately predicted positive observations to all observations in the actual class. Precision gives a probability of how correctly the classifier is predicting the actual positive class, as shown in Equation (4).

Recall = (TP /TP + FN)*100

d) F1 Score:

F1 Score is a normalized average of precision and recall. As a result, this score includes both false positives and false negatives. Although F1 score is simpler than accuracy, it is more useful, especially if class distribution is irregular. F1 score is a harmonic mean of precision and recall, as shown inEquation.

FMeasure = 2 × PR/ (P + R)

VI. Dataset

There are 104345 rows and 23 columns. There is a one target variable called label: contains only 1(malicious) and 0(benign).

These 23 attributes include:

1.Packet_count *– refers to the count of packets*

2.byte_count *– refers to the count of bytes in the packet*

3.Switch-id *– ID of the switch*

4.duration_sec *– packet transmission (in seconds)*

5.duration_nsec *–packet transmission (in nanoseconds)*

6.Source IP *– IP address of the source machine*

7.Destination IP *– IP address of the destination*

8. ***Port Number*** *– Port number of the application*

9. ***tx_bytes****– number of bytes transferred from theswitch*

10. ***rx_bytes****– numberof bytes received on switch port*

11. ***dt field*** *– shows the date and time which has been converted into a number and the flow is monitored at a monitoring interval of 30 seconds*

12. ***Byte Per Flow*** *– byte count during a single flow*

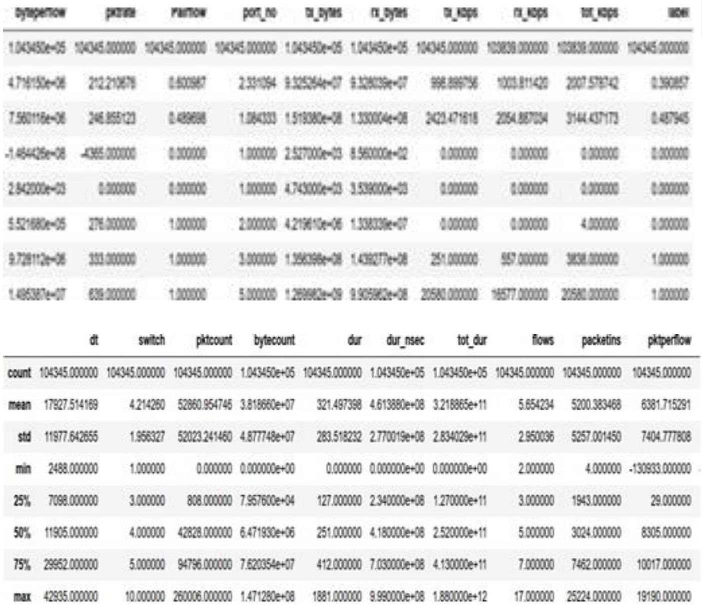
13. ***Packet Per Flow*** *– packet count during a single flow*

14. ***Packet Rate*** *– number of packets transmitted per second and calculated by dividing the packet per flow by monitering number of Packet_ins messages– messages that are generated by the switch and is sent to the controller.*

15. ***Flow entries of switch*** *– entries in the flow table ofa switch which is used to match and process packets.*

*16. **tx_kbps** – Speed of packet transmission (in kbps)* ***rx_kbps -*** *Speed of packet reception (in kbps)*

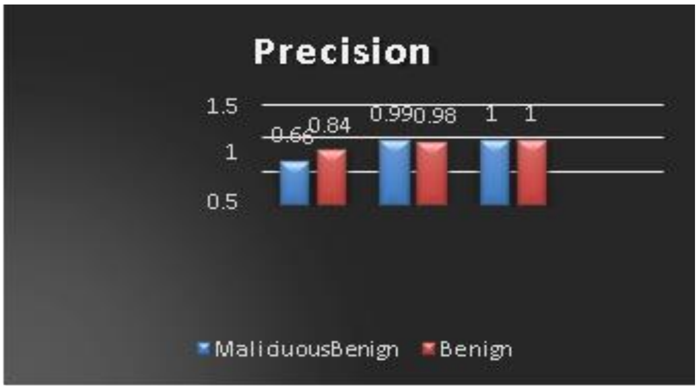
17. Port Bandwidth *– Sum of tx_kbps and rx_kbp*

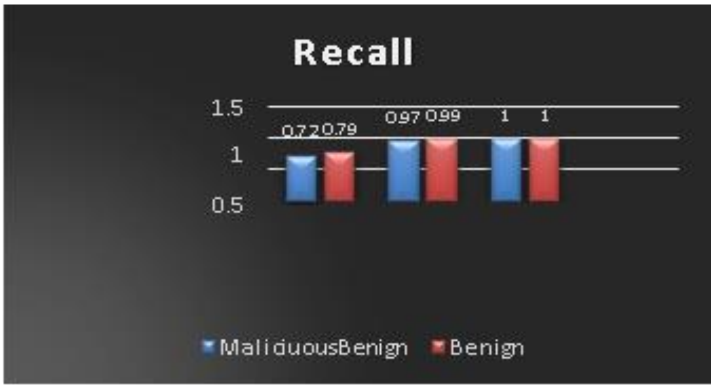


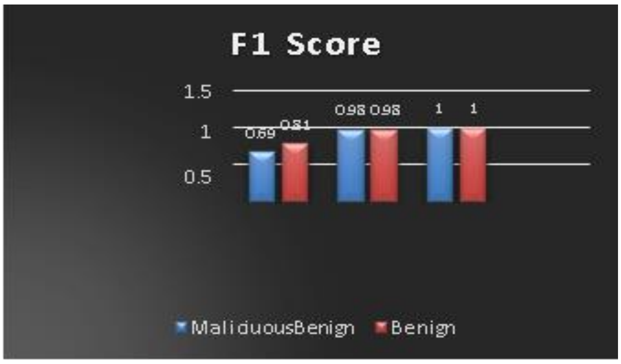
byteperflow	pktrate	Pairflow	port_no	tx_bytes	rx_bytes	tx_kbps	rx_kbps	tot_kbps	label
1.043450e+05	104345.000000	104345.000000	104345.000000	1.043450e+05	1.043450e+05	104345.000000	103839.000000	103839.000000	104345.000000
4.716150e+06	212.210676	0.600987	2.331094	9.325264e+07	9.328039e+07	998.899756	1003.811420	2007.578742	0.390857
7.560116e+06	246.855123	0.489698	1.084333	1.519380e+08	1.330004e+08	2423.471618	2054.867034	3144.437173	0.487945
-1.464426e+08	-4365.000000	0.000000	1.000000	2.527000e+03	8.560000e+02	0.000000	0.000000	0.000000	0.000000
2.842000e+03	0.000000	0.000000	1.000000	4.743000e+03	3.539000e+03	0.000000	0.000000	0.000000	0.000000
5.521680e+05	276.000000	1.000000	2.000000	4.219610e+06	1.338039e+07	0.000000	0.000000	4.000000	0.000000
9.728112e+06	333.000000	1.000000	3.000000	1.356398e+08	1.439277e+08	251.000000	557.000000	3838.000000	1.000000
1.495387e+07	639.000000	1.000000	5.000000	1.269982e+09	9.905962e+08	20580.000000	16577.000000	20580.000000	1.000000

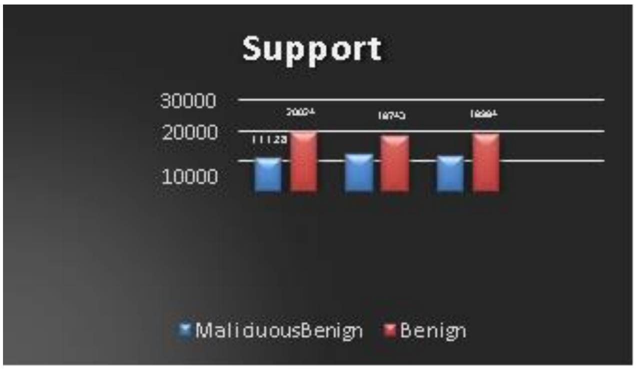
	dt	switch	pktcount	bytecount	dur	dur_nsec	tot_dur	flows	packetins	pktperflow
count	104345.000000	104345.000000	104345.000000	1.043450e+05	104345.000000	1.043450e+05	1.043450e+05	104345.000000	104345.000000	104345.000000
mean	17927.514169	4.214260	52860.954746	3.818660e+07	321.497398	4.613880e+08	3.218865e+11	5.654234	5200.383468	6381.715291
std	11977.642655	1.956327	52023.241460	4.877748e+07	283.518232	2.770019e+08	2.834029e+11	2.950036	5257.001450	7404.777808
min	2488.000000	1.000000	0.000000	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	2.000000	4.000000	-130933.000000
25%	7098.000000	3.000000	808.000000	7.957600e+04	127.000000	2.340000e+08	1.270000e+11	3.000000	1943.000000	29.000000
50%	11905.000000	4.000000	42828.000000	6.471930e+06	251.000000	4.180000e+08	2.520000e+11	5.000000	3024.000000	8305.000000
75%	29952.000000	5.000000	94796.000000	7.620354e+07	412.000000	7.030000e+08	4.130000e+11	7.000000	7462.000000	10017.000000
max	42935.000000	10.000000	260006.000000	1.471280e+08	1881.000000	9.990000e+08	1.880000e+12	17.000000	25224.000000	19190.000000

VII. Results

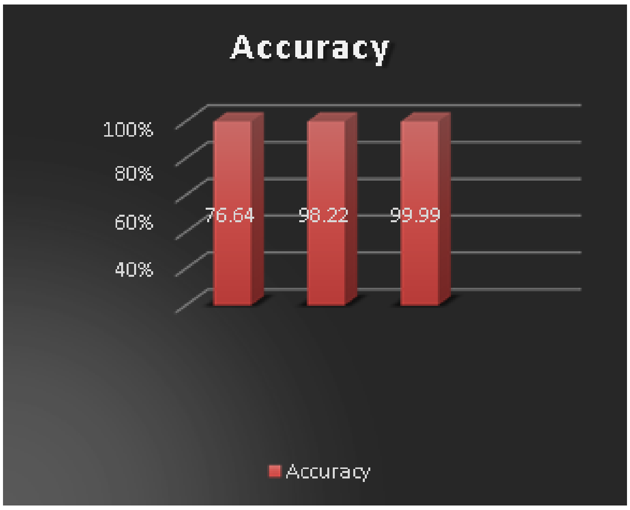






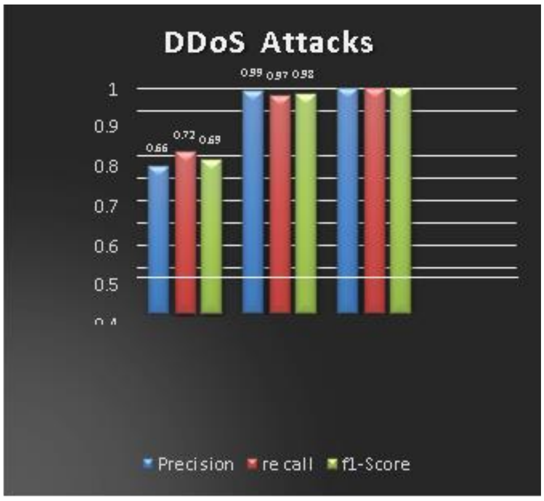


VIII. Conclusion



In this paper, we put together a Software Defined Networking dataset(SDN_Dataset) that covers the types of modern attacks, which were not used in previous studies. The dataset contains 23 features . Attacks are carried out directly to the target server and capture packet data.

In a cloud environment, DDoS attack detection and prevention are critical issues.DDoS attack detection is a binary class problem with labels for benign and Malicious attacks.Benign is a regular class. Because the interest is in detecting an attack, we consider the existence of an attack to be a positive class, and benign to be a negative class.Reviewing and understanding the DDoS attack architecture is thought to be a vital step in deploying the right method to stop this attack before the attacker overwhelms the legitimate internet applications.Performance comparison is done with the ultimate purpose of promoting real- time DDoS attack avoidance.



Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

[1] Kuzmin V.V. Classification and Identification of Motion in a Multiservice Network of a Communication Operator. *Sovremennyye problemy nauki i obrazovaniya*[Modern problems of science and education], 2014, № 5.

[2] Cheskidov P., Nikolskaia K., Minbaleev A. Choosing the Reinforcement Learning Method for Modeling DDos Attacks. *Proceedings of the IEEE International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*, 1-4 Oct. 2019

[3] DDoS attacks in the fourth quarter of 2019. Available at: https://securelist.ru/ddos-report-q4-2019/95568/ (Accessed: 18.03.2020)