JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Detection of DDoS Attack using Machine Learning Algorithms

[1]C M NalayiniI, [2]Dr. Jeevaa Katiravan
[1] Assistant Professor, [2]Professor
[1]Department of Information Technology
[1]Velammal Engineering College, Chennai, India

Abstract: ***Distributed Denial of Service (DDoS) attack is one of the common network attacks. DDoS attack occurs when a website or server is targeted by a malicious user to deny the services by flooding with unwanted information. This causes delay of services to legitimate user. Denial of Service (DoS) attack happens when the attack is from single source, whereas Distributed Denial of Service attack (DDoS) happens when the attack is from many number of sources say Botnet which controls the devices remotely for malicious purpose. A set of eight supervised machine learning algorithms are selected to detect DDoS attack and found the best model in terms of accuracy, precision, recal and false alarm ratel. For experimental results, a standard benchmark dataset CIC-IDS2017 is used for training and testing purpose. K-Fold cross validation is performed during the preprocessing stage. Then the eight models are trained and tested via K-Fold cross validation to find the best one to detect the DDoS attack at the earliest stage. In the testing phase we tested the trained models with the parameters Accuracy, Precision, Recall and FAR. Among eight models we found that Random Forest is the best model by considering all parameters into account. It has produced 99.885% accuracy, 99.88% Precision, 100% Recall and 0.05% False alarm rate to detect DDoS attack at the earliest.***

IndexTerms **- DDoS-Distributed Denial of Service, K-Fold Cross Validation, Machine Learning Algorithms.**

I. INTRODUCTION

DDoS attack is one of the most dangerous and growing attack with monstrous growth of Internet [1]. Botnet is a network of bots works together to target the victim abundantly [2] as shown in the Figure 1. It is a complex problem because it involves bots in distributed environment in the Internet and affecting numerous networks [3]. It is a big threat to all cloud related platforms say like Amazon Cloud servers [4, 5]. In February a very largest DDoS attacks was launched with 1.3 of Terabits per second traffic transfer [6]. The most severe attacks are Traffic attack, Application attack and Volume attack or bandwidth attack [11]. Traffic based attacks[21,22] targets the victim by sending large volume of TCP and UDP packets via botnet to degrade the performance of the server to make the targeted network down. In volume based attack, first the attacker identifies the total bandwidth of the targeted network and sends bulk volume of unwanted data to occupy maximum bandwidth to deny the services of the legitimate user. Another type of attack is mainly targeting specific applications to make application access difficult and unavailable to legitimate users. In traditional Systems we generally use filtering [12] and Threshold based techniques for detection and trace route mechanism for mitigation of DDoS. But identifying unknown attacks is a big task in the traditional methods. Most of the attackers targets application layer to damage the access severely [13]. Its main focus is to affect the normal behaviour of the system [14]. Since security is the major challenge in Networking Field [15], it's very essential to efficient tools and methodology to defend against DDoS attack. Now a day machine leaning is one efficient methodology help us to detect known and unknown attacks accurately with the help many available machine learning algorithms. DDoS attack comes under the classification problem [8,9]. Since the attack traffic size is relatively larger its crucial to find the best algorithm to detect it accurately, there is need to find a good algorithm [10]. In machine learning, dataset is fed as the input to various models to train and test it efficiently for efficient detection and prediction [16]. DDoS attack detection and prediction can be done by comparing various parameters such as accuracy, precision, recall and false alarm rate[17,18]. Therefore it is necessary to find best machine learning model to detect DDoS at an earlier stage to avoid major security issues in the network.

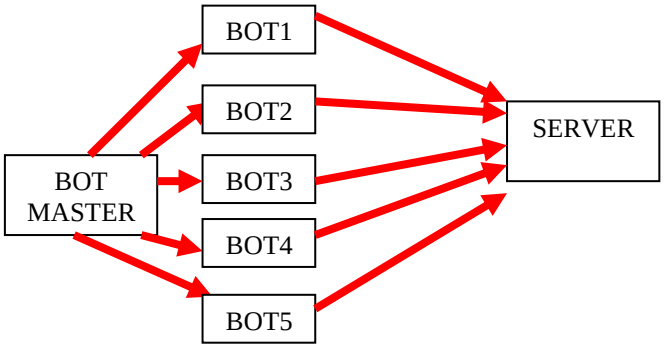

Figure1 A DDoS Attack Scenario

Machine Learning Algorithms

i. Logistic Regression

It is one of the supervised learning algorithm help us to classify the existence of DDoS or Non-DDoS based on the probability.

ii. Random Forest Classifier

Random Forest is a supervised learning classification and regression algorithm. It involves many decision trees to do the best prediction of DDoS attack. It ensembles all the constructed decision trees to have better prediction [20].

iii. Support Vector Machine (SVM)

It is a supervised learning algorithm and applicable for both classification and regression problems. Hyper plane is the decision boundary which segregates classes based on the vector points[23]. There are 9 kernel functions in SVM. For our model we have taken 3 kernel functions. SVM with a polynomial kernel gives a non-linear decision boundary to separate the classes. Gaussian Radial Basis Function is mostly applicable for non-linear data. Gamma values lies from 0 to 1. Sigmoid kernel function acts as an activation function for artificial neurons. Detection of unknown malicious activity can be found be efficient with the help of this artificial neurons.

iv. Decision tree

It is a supervised machine learning algorithm applicable for both classification and regression problems. It involves Root node, Internal nodes and leaf nodes. Root node is the cause of division of the features set. Internal nodes are the split nodes or decision nodes derived from root node. Decision node can be divided further based on the decision rules. leaf nodes represents the final output.

v. Gaussian Naive Bayes

Gaussian Naïve Bayes is a supervised classification algorithm. Bayes theorem works with conditional attributes for the selected features to yield the class value.

vi. K-Nearest Neighbours

It is the supervised learning technique. It can be used in both regression and classification. Cases or data stored in K-Nearest Neighbour are classified based on the similarity. It classifies the data point based on its neighbour classification.

II. LITERATURE REVIEW

S.No	Name of the paper	Authors	Merits	Demerits	Reference No.
1	DDOS Attack Detection Using Machine Learning	Ashutosh Nath Rimal and Dr.Raja Praveen	Proposed a DDoS attack detection using ML algorithms and packet analysis in a smart way.	This paper proposed SVM to be the better algorithm than Naive bayes but the false alarm rate of Naive bayes is lower than SVM.	[1]
2	A survey of defence mechanisms against distributed denial of service(DDoS)flooding attacks	Zargar S T, Joshi J, Tipper D.	Classification of various DDoS detection mechanism explained in detail for beginners.	Compared various defence mechanisms performed at network/transport-level, application-level, but effective solution is missing.	[2]
3	DDoS attacks in cloud computing	Somani, G, Gaur, M.S, Sanghi, D, Conti, M, Buyya, R.	Detailed taxonomy on DDoS detection in cloud computing.	Focused on only a particular environment that is cloud computing.	[5]
4	Detecting DDoS attacks against data center with correlation analysis.	Xiao, P, Qu, W, Qi, H.	Different types of DDoS attack anomalies are elaborated neatly.	KNN is proposed as the best algorithm to detect DDoS attack. For all applications KNN is not the best choice	[7]

5	A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear Regression	Swathi Sambangi and Lakshmeeswari Gondi	Proposed classification ML algorithms based on the attack type.	If application wise the best models are suggested then it would have been a better solution.	**[9]**
6	Research and Implementation of DDoS Attack Detection Based on Machine Learning in Distributed Environment	Tan Miao	The severity of DDoS is well portrayed in a distributed environment	The context of the data stream is not fully used.	**[10]**
7	Taxonomy of Distributed Denial of Service mitigation approaches for cloud computing.	Alireza Shameli-Sendi, et al.	Taxonomy of DDoS and Counter measures are clearly explained in a cloud environment	Context-aware mechanism is needed.	**[12]**
8	Prevention Mechanism on DDOS Attacks by using Multilevel Filtering of Distributed Firewalls	Sagar pande and Gadicha	Proposed a framework which supports multi-level filtering at firewalls.	Number of features used is not in satisfactory limit.	**[15]**
9	A Comparison of Various Machine Learning Algorithms in a Distributed Denial of Service Intrusion	SH Kok, Azween Abdullah, et al.	The accuracy measures of algorithm are well derived.	Further enhancements to achieve betterment of the model would have been explained	**[16]**
10	Detecting Distributed Denial of Service Attacks Using Data Mining Techniques	Mouhammd Alkasassbeh,Ahmad B.A Hassanat,Ghazi Al-Naymat,Mohammad Almseidin	A network simulator (NS2) was used in this work which is capable of producing valid results and reflects a real environment.	Only concentrated on network layer and application layer.	**[19]**

III. METHODOLOGY

We used CIC IDS 2017 dataset as the input to our proposed system. It involves data pre-processing, K fold Cross Validation which intakes all model into account for train and test as shown in Figure 2. Finally based on the parameters accuracy, precision, recall and FAR, we conclude that Random forest is the best model to detect DDoS at the earliest.

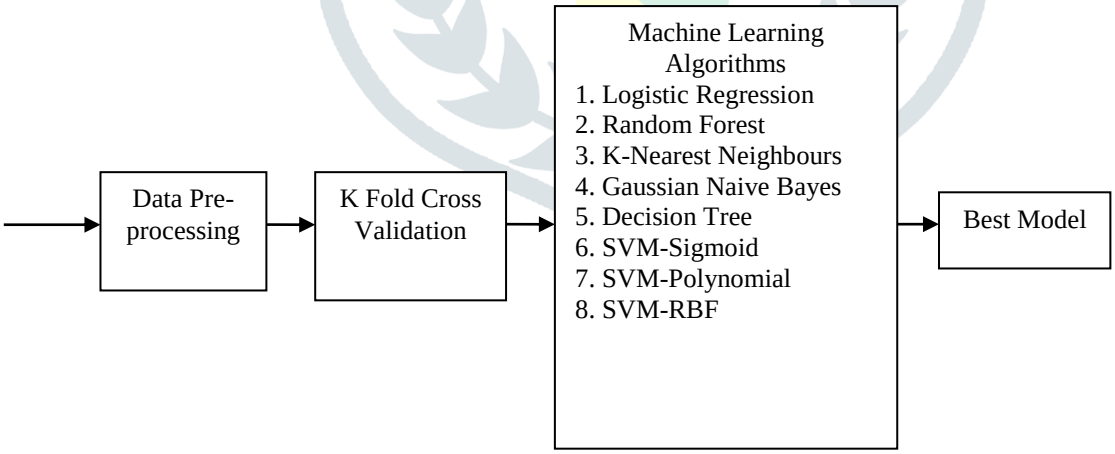

Figure.2 Proposed Model

Dataset:

CIC IDS 2017 dataset involves most recent attacks. It has the network traffic analysis result collected based on the parameters source IP, Destination IP, source and destination ports, timestamp, protocols etc.[2]. It is one most widely used dataset to classify the existence of DDoS or Non-DDoS[19].

Data pre-processing:

It is the first step in creating a machine learning model. It is a process of making the raw data to fit the machine learning model [11]. The dataset was imported by the function *read_csv()* from pandas library which is used to read csv files. We have used *iloc[ ]* method from pandas library to extract both independent and dependant variables from the dataset.

K-Fold Cross Validation:

In K fold cross validation we considered k value as 5 as shown in Figure 3. Our dataset is divided into 5 equal splits. k=5 times the process gets executed as shown in the figure 3. Among 5 splits, one split is for testing and the remaining splits for training. Fit the machine learning models on the training splits and evaluate it on the testing split. For each iteration, accuracy, precision, recall and FAR values are recorded. Based on the estimated values we found that Random Forest produced the best results.

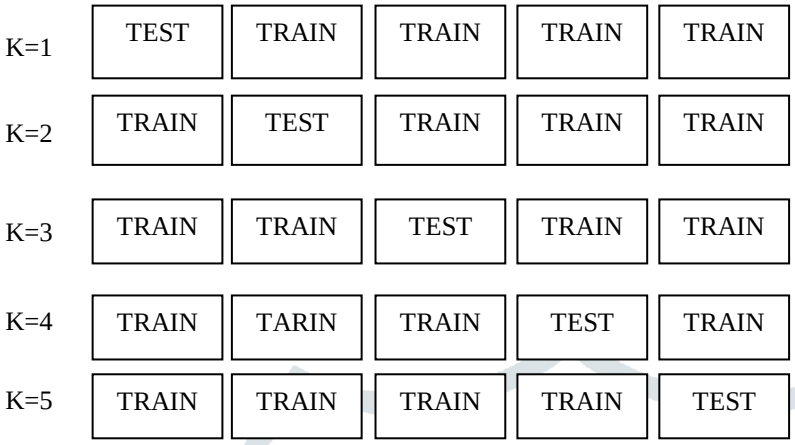

Figure.3 K Fold Cross Validation

IV. RESULTS AND DISCUSSIONS

Performance metrics

Performance is evaluated based on the confusion matrix which is shown in Table 1. The terms used in the confusion matrix are,

- **TP (True Positive)**

No of malicious requests predicted as anomaly.

- **TN (True Negative)**

No of legitimate requests predicted as normal.

- **FP (False Positive)**

No of legitimate requests predicted as anomaly.

- **FN (False Negative)**

No of malicious requests predicted as normal.

Table 1: CONFUSION MATRIX

Class	Positive prediction	Negative prediction
Positive class	True positive (TP) [00]	False Negative (FN) [01]
Negative class	False Positive (FP) [10]	True Negative (TN) [11]

- **Accuracy:**

No of correct prediction decides the model's accuracy:

Accuracy = (cm[0][0]+cm[1][1])*100/(cm[0][0]+cm[1][1]+cm[0][1]+cm[1][0])
(Here, cm refers to confusion matrix) (ie).,
Accuracy= (true positive+true negative)*100/(true positive+true negative+
false negative+false positive)

- **Precision:**

No of correctly identified malicious requests by the model out of all the malicious requests.

Precision = (cm[0][0])*100/(cm[0][0]+cm[1][0])

Precision= (true positive)*100 / (true positive+false positive)

- **Recall / TPR (True Positive Rate):**
  The recall is the measure of our model correctly identifying True positive. Thus, for all malicious requests, recall tells us how many we correctly identified as malicious requests.

  Recall = (cm [0][0])*100/(cm[0][0]+cm[0][1])
  Recall = (true positive)*100 / (true positive+false negative)

- **FAR (False Alarm Rate):**
  The value of FAR must be less in a good machine learning model. Because, it calculates how often the model predicts positive output wrongly.

  FAR =cm[1][0]*100/(cm[1][0]+cm[1][1])
  FAR = (false positive)*100 / (false positive+true negative)

Using the confusion matrix, the scatter plot of eight ML algorithms has been plotted. This scatter plot depicts the attack and normal flood traffic of various ML algorithms. SSIP (Speed of source IP), SSP (Speed of Source Port) and RPF (Rate Pair Flow) indicated in axes in order to plot the data points in the pane using the dataset.
As shown in Table 2 and Figure 4-15 Logistic regression produced 99.86% accuracy, 99.86% precision, 100% recall and 10.53% false alarm rate. Random Forest produced 99.88% accuracy, 99.88 precision, 100% recall and 0.05 false alarm rate. K-Nearest Neighbour produced 99.80% accuracy, 99.90 precision, 99.90 recall and 0.27 false alarm rate. Guassian Naive Bayes produced 61.22% accuracy, 100% precision, 61.16% recall and 0.23% false alarm rate. Decision Tree produced 99.94% accuracy, 99.96% precision, 99.98 recall and 0.24 false alarm rate. SVM-sigmoid produced 99.78 accuracy, 99.86 precision, 99.92 recall and 6.75% false alarm rate. SVM-Polynomial produced 99.86 accuracy, 99.86 precision, 100% recall and 2.34% false alarm rate. SVM-RBF produced 99.84% accuracy, 99.84% precision, 100% recall and 0.62% false alarm rate. By comparing the parameters accuracy, precision, recall and false alarm rate of eight algorithms, Random Forest produced best results to detect the existence of DDoS attack with low false alarm rate 0.05%.

Table 2. Estimated average values of eight ML algorithms

Machine Learning Algorithms	ACCURACY	PRECISION	RECALL/TPR	FAR
LOGISTC REGRESSION	99.86	99.86	100.00	10.53
RANDOM FOREST	99.88	99.88	100.00	0.05
K-NEAREST NEIGHBOURS	99.80	99.90	99.90	0.27
GAUSSIAN NAIVE BAYES	61.22	100.00	61.16	0.23
DECISION TREE	99.94	99.96	99.98	0.24
SVM-SIGMOID	99.78	99.86	99.92	6.75
SVM-POLYNOMIAL	99.86	99.86	100.00	2.34
SVM-RBF	99.84	99.84	100.00	0.62

```
Accuracy - Logical Regression :  99.86108354832308
Precision - Logical Regression :  99.86105597459309
Recall - Logical Regression :  100.0
FAR - Logical Regression : 10.53
```

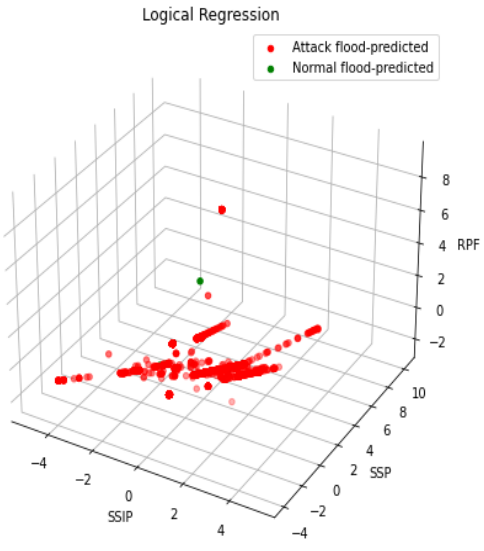

Figure 4 Logistic Regression

```
Accuracy - Random Forest :  99.8809287557055
Precision - Random Forest :  99.88088147706968
Recall - Random Forest :  100.0
FAR - Random Forest : 0.05
```

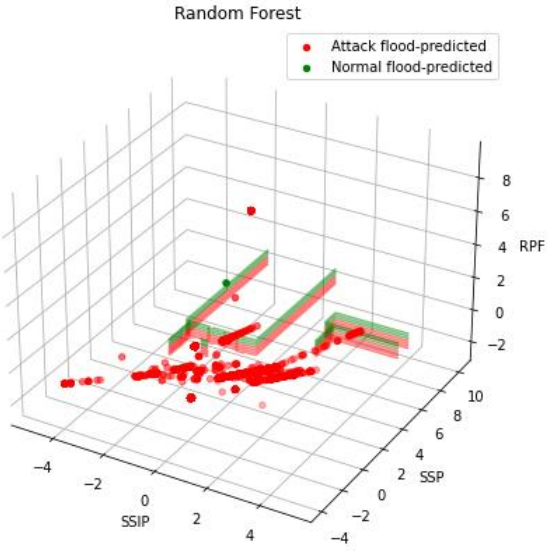

Figure 5 Random Forest

```
Accuracy - K-nearest Neighbours :  99.80154792617583
Precision - K-nearest Neighbours :  99.90061617968594
Recall - K-nearest Neighbours :  99.90061617968594
FAR - K-nearest Neighbours : 0.27
```

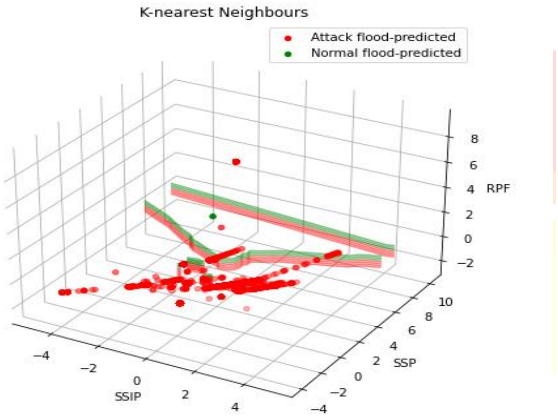

Figure 6 K-nearest neighbours

```
Accuracy - Gaussian Naive Bayes :  61.222464774756894
Precision - Gaussian Naive Bayes :  100.0
Recall - Gaussian Naive Bayes :  61.16080302126814
FAR - Gaussian Naive Bayes : 0.23
```

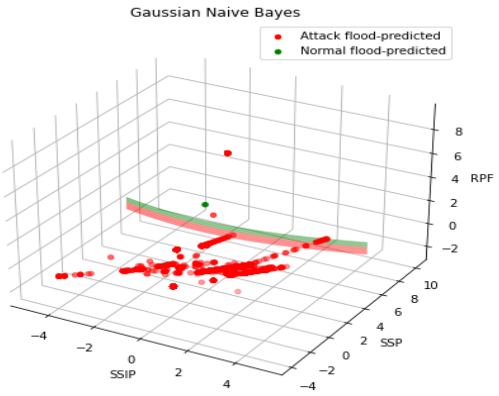

Figure 7 Gaussian naive bayes

```
Accuracy - Decision Tree :  99.94046437785275
Precision - Decision Tree :  99.96025437201908
Recall - Decision Tree :  99.98012323593719
FAR - Decision Tree : 0.24
```

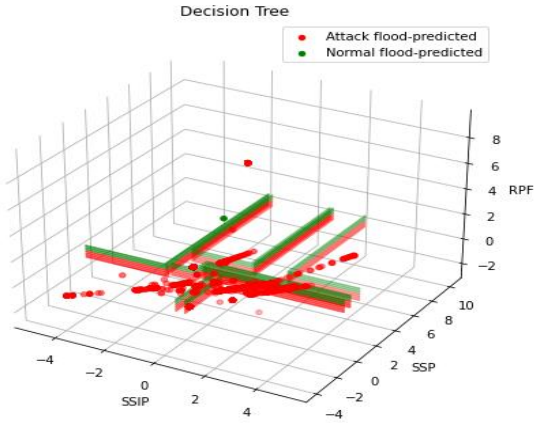

Figure 8 Decision Tree

```
Accuracy - Support Vector Machine - sigmoid :  99.78170271879341
Precision - Support Vector Machine - sigmoid :  99.86094557012316
Recall - Support Vector Machine - sigmoid :  99.92049294374876
FAR - Support Vector Machine - sigmoid : 6.75
```

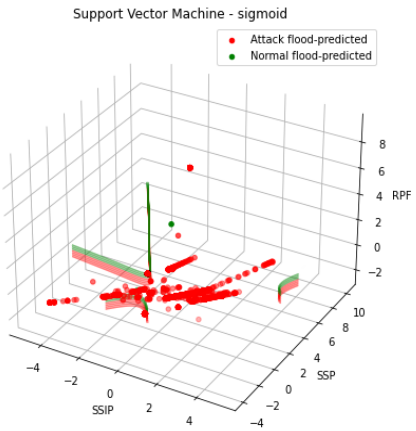

Figure 9 Support Vector Machine-Sigmoid

```
Accuracy - Support Vector Machine - poly :  99.86108354832308
Precision - Support Vector Machine - poly :  99.86105597459309
Recall - Support Vector Machine - poly :  100.0
FAR - Support Vector Machine - poly : 2.34
```

```
Accuracy - Support Vector Machine - rbf :  99.84123834094066
Precision - Support Vector Machine - rbf :  99.84123834094066
Recall - Support Vector Machine - rbf :  100.0
FAR - Support Vector Machine - rbf : 0.62
```

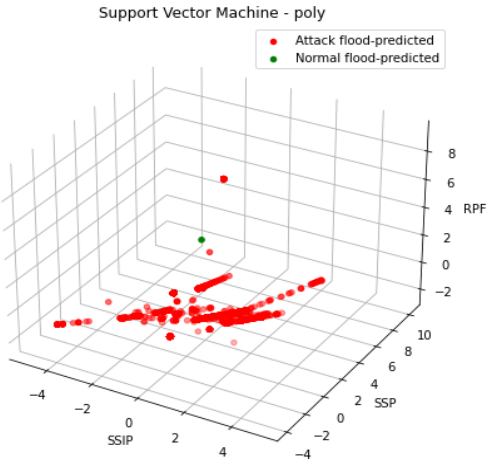

Figure 10 Support Vector Machine - poly

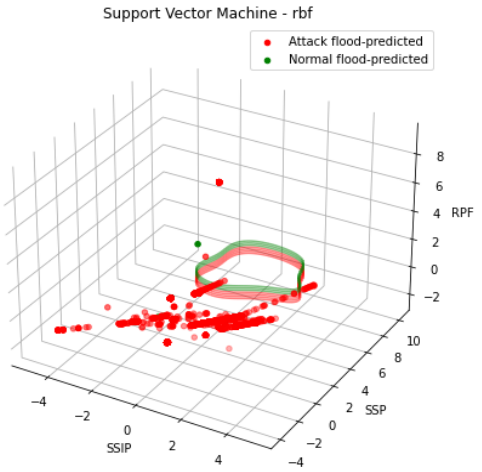

Figure 11 Support Vector Machine - rbf

Accuracy:

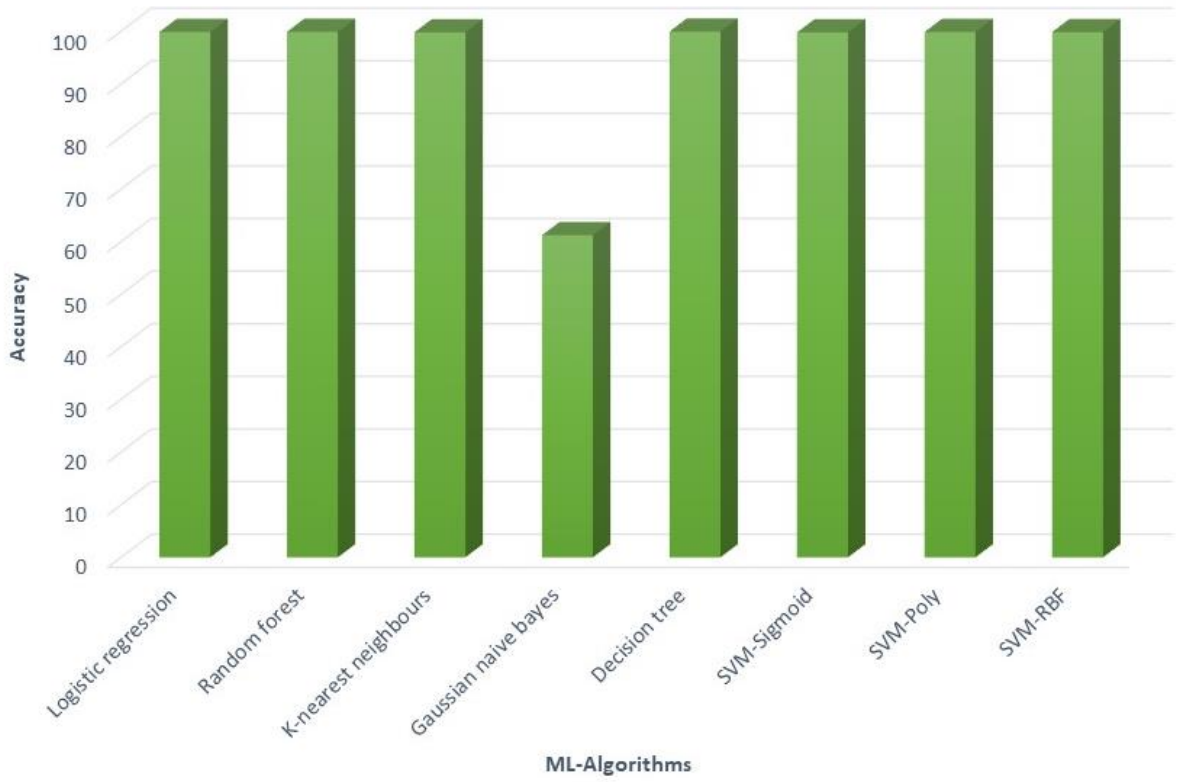

Figure 12 Accuracy

Precision:

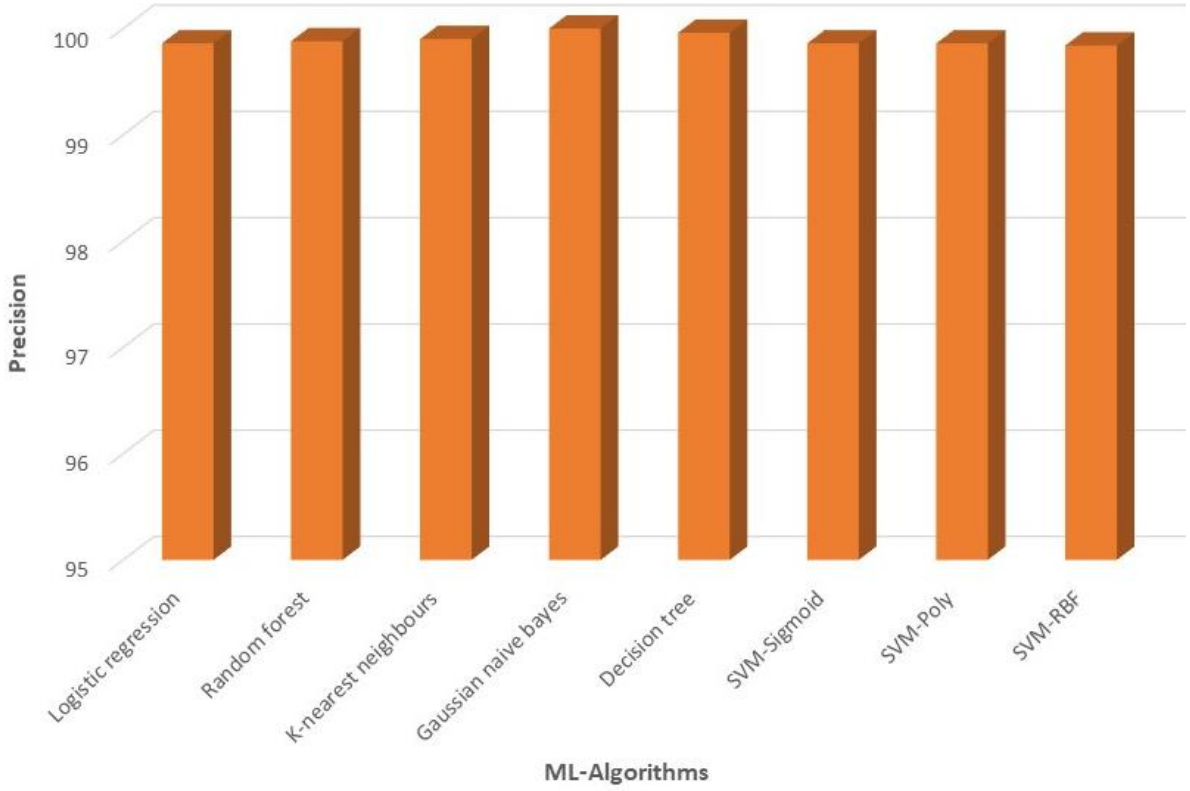

Figure 13 Precision

Recall / TPR:

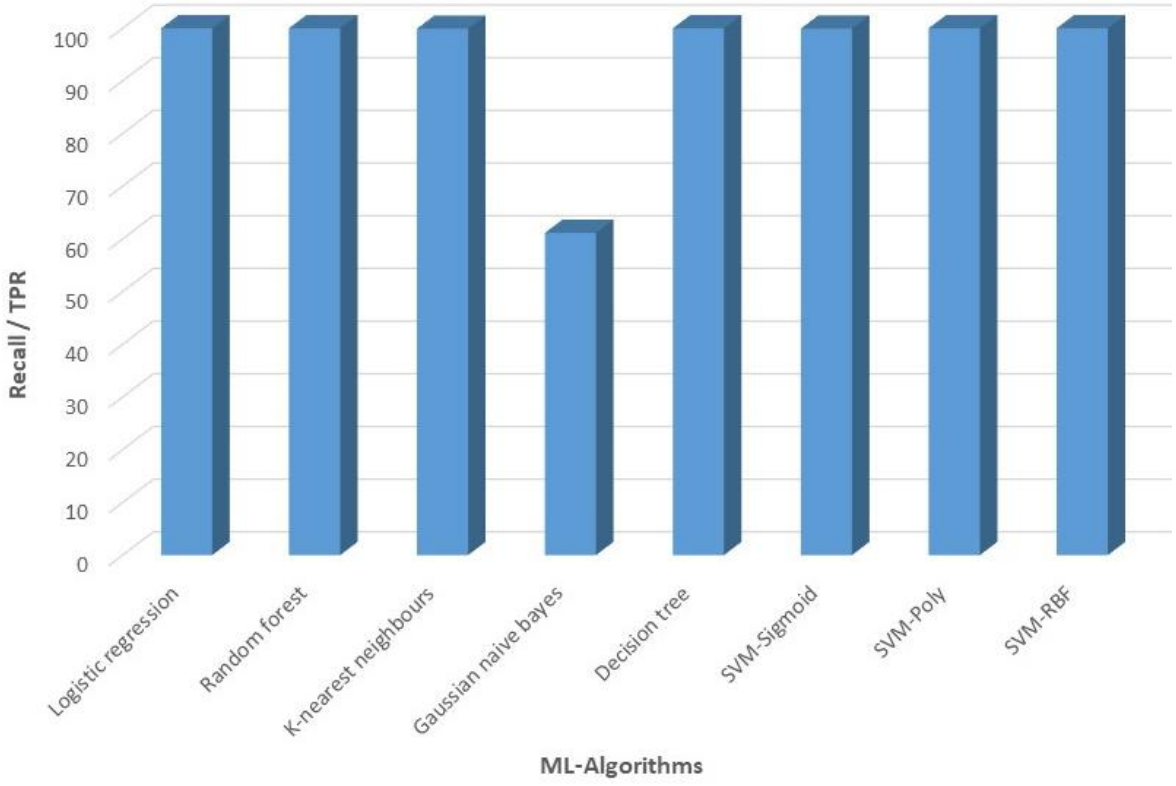

Figure 14 Recall / TPR:

FAR:

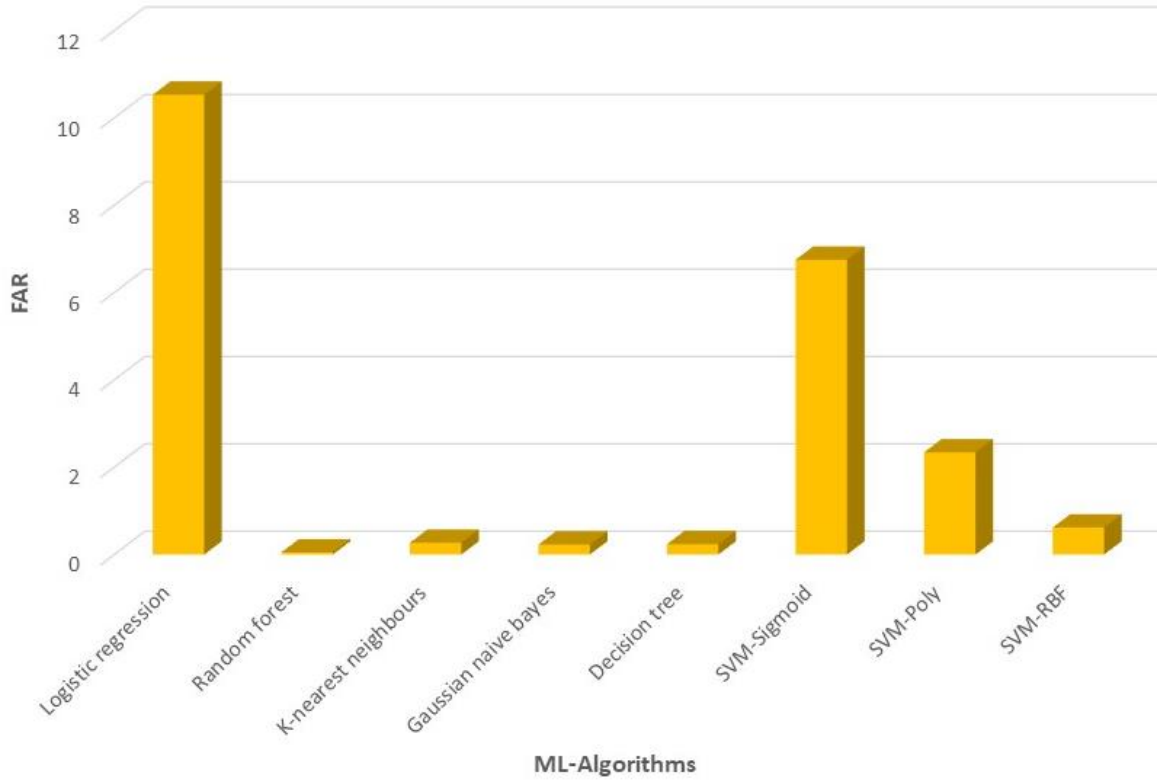

Figure 15 False Alarm Rate

Conclusion

Random Forest algorithm is found to be the best among the eight algorithms to detect DDoS attack in the CIC IDS 2017 dataset by applying k-fold cross validation. It gives high accuracy, precision and recall/TPR at the same time gives low FAR (False Alarm Rate). The FAR of the Random Forest which is very low say like 0.05 is comparatively low than other algorithms. It produces more accuracy by reducing the over fitting in all its decision trees.

Future work:

In our paper we have used Machine learning algorithms to detect DDoS attack and found Random forest is the best algorithm out 8 other ML algorithms. We used k fold cross validation to have better performance of the models. In future we will concentrate on optimized feature selection and to generate own data set to validate the best found models.

References:

[1].Ashutosh Nath Rimal and Dr. Raja Praveen; DDOS Attack Detection Using Machine Learning – 2020.
[2]. Zargar S T, Joshi J, Tipper D. A survey of defense mechanisms against distributed denial of service(DDoS)flooding attacks[J]. IEEE Corn— munications Surveys&Tutorials2013, 15(4) : 2046—2069.
[3]. Jelena Mirkovic, Max Robinson, George Oikonomou, Peter Reiher; Distributed Defense against DDoS Attacks.
[4]. Parvinder Singh Saini, Sunny Behal, Sajal Bhatia; Detection of DDoS Attacks using Machine Learning Algorithms 12-14 March 2020.
[5]. SomaniG, Gaur M.S, Sanghi D, Conti M, Buyya R; DDoS attacks in cloud computing: Issues, taxonomy, and future directions. Comput. Commun. 2017, 107, 30–48.
[6]. https://www.foxnews.com/tech/biggest-ddos-attack-on-record-hits-github
[7]. Xiao P, Qu W, Qi H, Li Z; Detecting DDoS attacks against data center with correlation analysis. Comput. Commun. 2015, 67, 66–74.
[8]. Manjula Suresh and R.Anitha; Evaluating Machine Learning Algorithms for Detecting DDoS Attacks; CNSA 2011: Advances in Network Security and Applications pp 441-452.
[9]. Swathi Sambangi and Lakshmeeswari Gondi; A Machine Learning Approach for DDoS
(Distributed Denial of Service) Attack Detection Using Multiple Linear Regression 2020.
[10]. Jiangtao Pei, Yunli Chen1 and Wei Ji1;A DDoS Attack Detection Method Based on Machine Learning; IOP Conf. Series: Journal of Physics: Conf. Series 1237 (2019) 032040.
[11]. https://cybersecurity.att.com/blogs/security-essentials/types-of-ddos-attacks-explained
[12].https://www.researchgate.net/publication/283295345Taxonomy of Distributed Denial of Service mitigation approaches for cloud computing.
[13].https://cryptome.org/2014/01/ddos-defense.pdf- author Muhammad Aamir and Mustafa Ali Zaidi SZABIST, Karachi, Pakistan.
[14]. S Behal and K Kumar, "Detection of DDoS attacks and flash events using novel information theory metrics", ELSEVIER Computer Networks, vol. 116, pp. 96-110, 2017.

[15]. Sagar pande, S., & Gadicha, A. B. (2015). Prevention mechanism on DDOS attacks by using multi-level filtering of distributed firewalls. ISSN: 2321–8169.
[16]. SH Kok, Azween Abdullah, Mahadevan Supramaniam, Thulasyammal Ramiah Pillai, Ibrahim Abaker Targio Hashem ; A Comparison of Various Machine Learning Algorithms in a Distributed Denial of Service Intrusion 2019.
[17]. Arshi M, Nasreen MD and Karanam Madhavi; A Survey of DDOS Attacks Using Machine Learning Techniques 2020.
[18]. Igor Kotenko and Alexander Uianov;Agent-based simulation of ddos attacks and defense mechanisms; computing, 2005, Vol. 4, Issue 2, 113-123.
[19]. Mouhammd Alkasassbeh,Ahmad B.A Hassanat,Ghazi Al-Naymat,Mohammad Almseidin; Detecting Distributed Denial of Service Attacks Using Data Mining Techniques – 2016.
[20]. Khamparia A, PandeS, Gupta D, Khanna A,Sangaiah A. K. (2020). Multi-level framework for anomaly detection in social networking. Library Hi Tech.2020.
[21] C.M.Nalayini, Dr. Jeevaa Katiravan, Araving Prasad V, "Flooding Attack on MANET – A Survey",International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, Feb 2017
[22] Nalayini, C.M., Katiravan, J. (2019). "Block Link Flooding Algorithm for TCP SYN Flooding Attack", International Conference on Computer Networks and Communication Technologies. Lecture Notes on Data Engineering and Communications Technologies, vol 15. Springer, Singapore. https://doi.org/10.1007/978-981-10-8681-6_83, 18 September 2018
[23] NalayiniI, C.M., Gayathri, T. (2022). A Comparative Analysis of Standard Classifiers with CHDTC to Detect Credit Card Fraudulent Transactions. In: Sivasubramanian, A., Shastry, P.N., Hong, P.C. (eds) Futuristic Communication and Network Technologies. Lecture Notes in Electrical Engineering, vol 792. Springer, Singapore. https://doi.org/10.1007/978-981-16-4625-6_99