International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Security System for DOS based attack using Machine Learning

Vidhi Agrawal[1], Nikhil Suryawanshi[2], Yachna Nankani[3], Yash Barange[4], Aman Pawar[5], Vaishnavi Shelke[6], Ashvin Mathankar[7]

Sanjay Kalamdhad[8]

Dept. of Computer Science and Engineering, Shri Balaji Institute of Technology and Management, Betul, M.P

ABSTRACT - *ML security system for DOS based attack is built to identify and stop the DOS attacking through the computer systems. In this approach we are using MLOPs for creating a basic wall against the attackers. MLOPs includes the Machine Learning, DevOps which is automate the things and always responsible for continues flow of the project. Using clustering approach in machine learning we identify the suspected IP of the attacker and then we command the firewall to block that IP. In this way this project save the server from going down and being not available for visitors.*

Key Words: Machine Learning, Clustering, DOS attack, SecOps, MLOps.

1. INTRODUCTION

Machine learning security system for Do's or DDoS based attack is the security system for the website's server protecting against the Denial of service attack. In this attack attackers targets the webserver and sends the bulk of request through the website and running down the web server. And keeping server busy for the actual visitors. While the servers is hanging on for the actual visitors , the attacker trying to access the database content of the website and trying to misuse them, as well as the attacker trying to break the communication link between the server and the visitors . This type of cyber-attack can affect the company reputation and the stocks of the company. This cyber thread can affect the company shares in the market and then company faces the huge loss in the market. When client hits the site, the request from the client goes to the server through the firewall. The communication of the server and client is done through the IP addresses. An IP (Internet Protocol) is the unique identity of the every device. Using the IP address two devices can communicate each other. Whenever the client send the request to the webserver the request comes through the IP address and the web server creates a log file for the requests and these log files contains the IP address and the other information. From the log directories of web server we fetch the log files and filter the IP addresses from the log file using the software called Prometheus and grafana and then our other software called Jenkins send the filtered log file content to the machine learning program usually a machine learning written in python programming language.

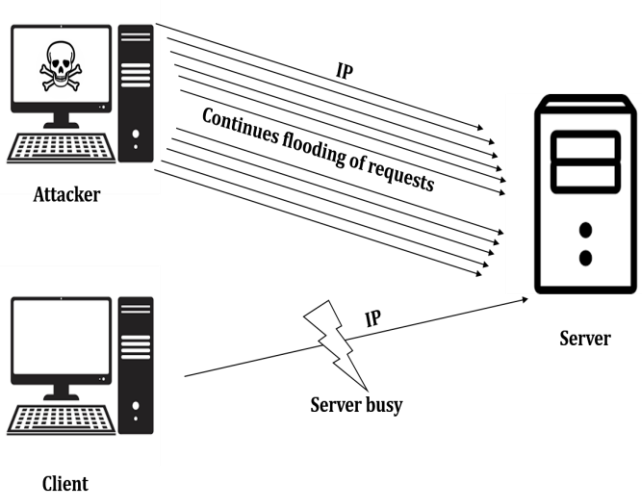

Fig. 1.1

In the machine learning we write the algorithms to creating a cluster of the IP address within specific time. In the machine learning we are using K-MEANS (K nearest neighbor) approach for creating a cluster. Here cluster means a group of item which have same property. This python program written the biggest cluster's IP to the software called Jenkins. This software can be programmed or triggered the operating system command for any software. So it can give the blocking command to the firewall to block that subspecies IP address. So that the attacker's IP block by the firewall and the attacker won't be able to access the site again.

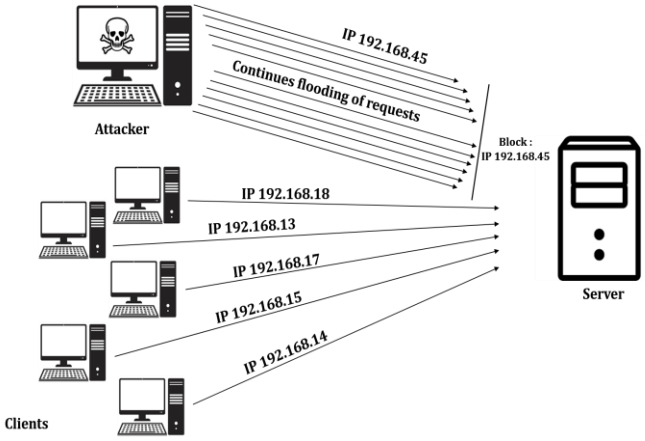

Fig 1.2

1.1 Problem Definition

This modern age of technology providing the best possible services to the client to access the company web site and application over the internet, but these services can facing the major security issues over the internet from the attackers. In present time lots of paid software are available for securing the web server against the server attacks, these attacks lead the trouble the web server and also the database. Our main aim is to secure the web server against the Dos type attacks and maintain the full functionality of the web server. These attacks can affect the database of the particular organization and can access the private information of the visitors and its services related information. Such kind of attacks can lead the undiversed traffic on our webserver. Attackers or hackers can flood the webserver through the Dos attack and can create a security glitch for the client so that the client won't be able to access the web site and can't use services provided by the company. It affect the company market value and this type of security issue can down the company revenue in the market. Attackers or hackers can directly harm any company or individual application or perform a malicious operation on your site or application without our noticing. So we need some fast, accurate and reliable mechanisms or program to detect attack and protect against attacks. Unfortunately, we don't have a unique software to pretend the server 100% secure. So we used the continues pipeline of the process using the software for creating a flow of processes for securing the web server. We know that Computer servers store all the valuable data of the company and the visitors for run a system, whether for its business, academic, communication or many other fields. That's why it should be the most protected part of any web or application server.

1.2 Machine Learning

As we know Machine learning is the trending technology and It is vastly used due to its capabilities. Basically machine learning is nothing but intelligent program or algorithm which can learn and operate the Machine have so many different algorithms for different tasks. We have used K-Nearest Neighbor in this project. By this algorithm we have trained the model and created the cluster of the IP so that it can determine the cluster of the IP which comes under few minutes and show the biggest cluster of attackers IP as a prediction. The machine learning is a technique of the programming in which we used the programming language to create several algorithms which is smart as humans.

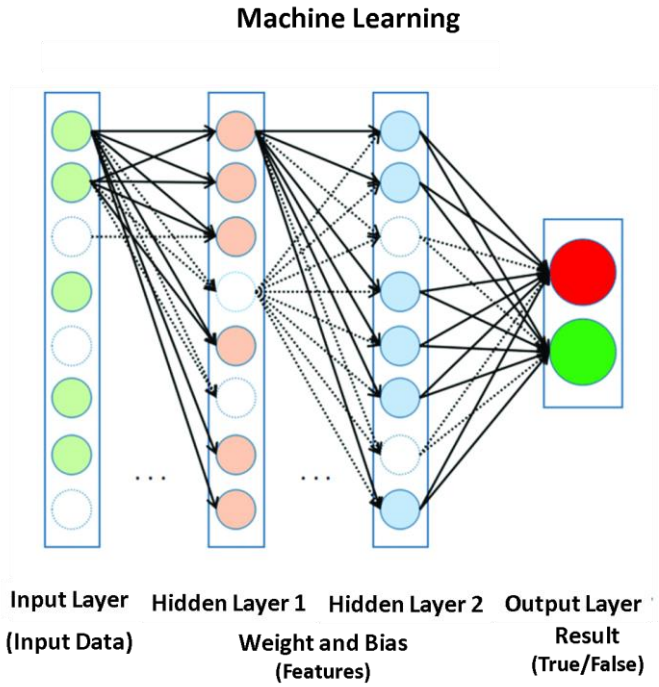

Fig 1.2.1 Machine Learning

In this algorithms the historic data is cleaned and feeded to this algorithms so that it can analyses the data and predict some values which is very useful in some areas. This algorithms generally works in the reference of linear regression. As we know Machine learning is the trending technology and It is vastly used due to its capabilities. Machine learning have so many different algorithms for different tasks. We have used K-Nearest Neighbor in this project. By this algorithm we have trained the model and created the cluster of the IP so that it can determine the cluster of the IP which comes under few minutes and show the biggest cluster of attackers IP as a prediction

1.3 Dos Attack

Definition

DOS attack is basically done by using all the resources of the user make all the resources busy so no one can use it. Dos attack is the easiest and most used attack .It is generally done by flooding the requests in a network ,so that the server gets busy by handling all the requests and uses all the resources of machines in handling requests. This may cause the machine failure or crash the server some of the Dos attack tools are Dossim, pyloris, HULK which are easily available for Linux based OS. DOS attack is caused due to the excess requests comes to the server from the user, that requests are not easy to handle due to excess in numbers and limited resources it can cause the system failure. In this attack the hacker used the tools that is specially created to target IP and send the data packets rapidly until the server gets crashed. This may cause slowdown the whole network due to the flood of data packets in network i.e.

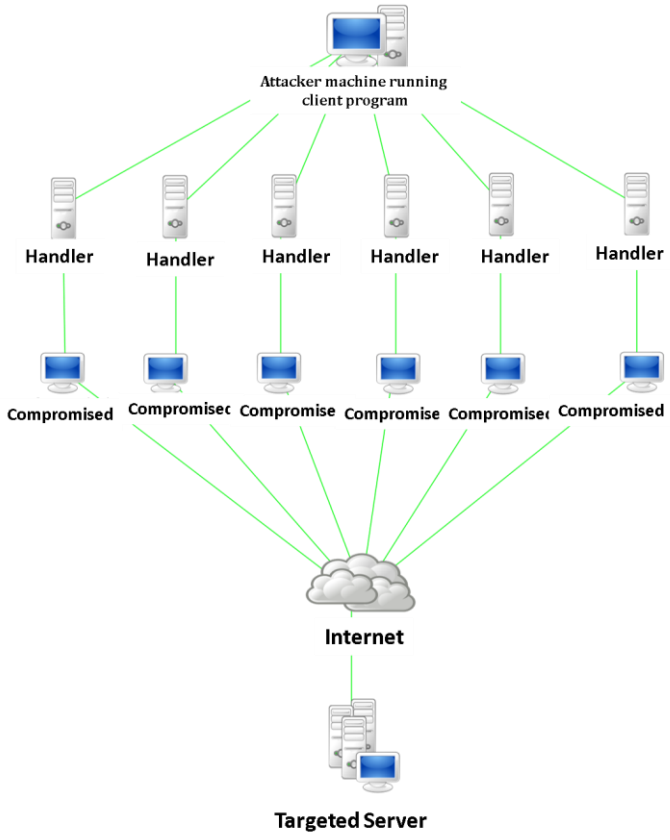

Fig 1.3.1 Dos Attack

Network Flooding. DOS attack is basically done by using all the resources of the user, hacker makes all the resources busy so no one can use it. Dos attack is the easiest and most used attack .It is generally done by flooding the requests in a network ,so that the server gets busy by handling all the requests and uses all the resources of machines in handling requests. This may cause the machine failure or crash the server some of the Dos attack tools are Dossim, pyloris, HULK which are easily available for Linux based OS. The Dos attack is harmful if this can be done to a reputed organization then it may lose the trust of the customers, today's generation no one wants to wait. By the dos attack sometime the site may fails for so long time

1.4 Security:

Definition

The security is the major concerned in today's digital world. This project is so helpful in overcoming from the dos based attacks. As soon as possible it detects the attack and block the IP from which the requests are coming. The major shield for any OS which provides the security is firewall. The firewall provides the basic security to the OS. It monitors the ports of the computer, which allows the network traffic to comes to the computer.

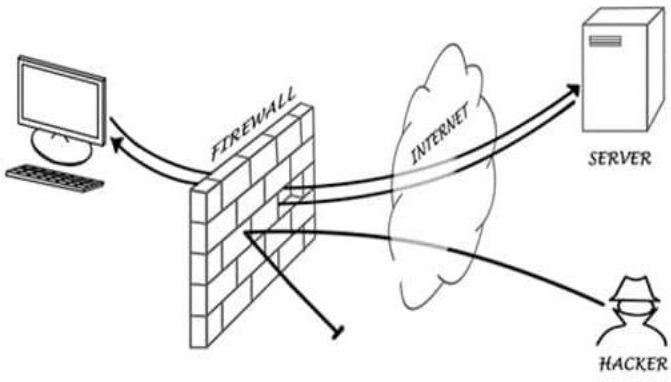

Fig 1.4.1 Firewall security

In the firewall IP tables is a CLI (Command-line interface) application that allows the administrator to configure specific rules that will enforce the Linux kernel (Netfilter framework) to perform an action such as inspect, modify or drop network packets. Enabling this IP tables in any Linux machine or device will be acting as a Network Firewall and/or a router. Different kernel modules and programs are used for different protocols; IP tables applies to IPv4, ip6tables to IPv6, ARP tables to ARP, and IP tables to Ethernet frames. This application is responsible for allowing or disallowing an IP address.

3. OBJECTIVE:

In the initial step we have to identify the suspicious IP of the attacker for this task we have to access the logs of the web server. As we know that every web server is the application software and every software needs the operating system for providing services and every operating system has it's own

firewall to protecting the application running insides the operating system, firewall has the authority to blocking and bypassing the IP of any user so that the application such as web server run smoothly on the operating system. When the visitors hits the web server, every web server creates the log file of the information related to the visitors such as IP address, hitting or request time and etc. Using some specific software we can access the log file and then this log file provides to the Machine Learning for filtering and creating the cluster of the specific data such as IP address and the time of request. As we know requests are hits the web server in every second so we have to access the log file in specific time duration so that we can create a machine learning program to filtering the important features of the log file. For this we can use a software called Jenkins which is used to trigger and manipulate the log files created by the web server. In machine learning using K-MEANS algorithm approach we can create the cluster of the IP address in particular time and get the biggest cluster of the IP then using the Jenkins software we command the operating system firewall to block that IP.

4. PROPOSED SYSTEM

In this section, we discuss various concepts or tools like machine learning approaches, K-MEANS , clustering, Jenkins for controlling the whole flow of the project. These application are used to handle the whole flow of the project and perform given tasks:

4.1 Log file accessing:

In this project we have used the data of the logs generated by the webserver while the rapid requests are coming to the server. This logs helps in training the model of the K-Nearest Neighbor Model. This logs are in the format of time stamp data. In this data the ingress IP, time, hit, response, etc. data are stored. This file needs to be filtered to get the information from this file.

4.2 IP address filtration

IP filtration is the process in which the information is gets extracted from the log file that is the training data for our ML K-Means Model. This process is also done by some DevOps tools like log stash or by the python program with the help of the regular expression (regex). This program filters the information from the log file and arrange it to the csv form so that we can use it further as the pandas data frame.

4.3 Machine learning programming

K-Means is a concept of the machine learning . It is a clustering technique widely used to partition a collection of data in groups k automatically .It create the cluster of the data and find the specific specialty in the data. In this project it is used to find the IP cluster which hits the maximum request to the server in particular duration. And helps to know either the packets is normal or is Dos attack.

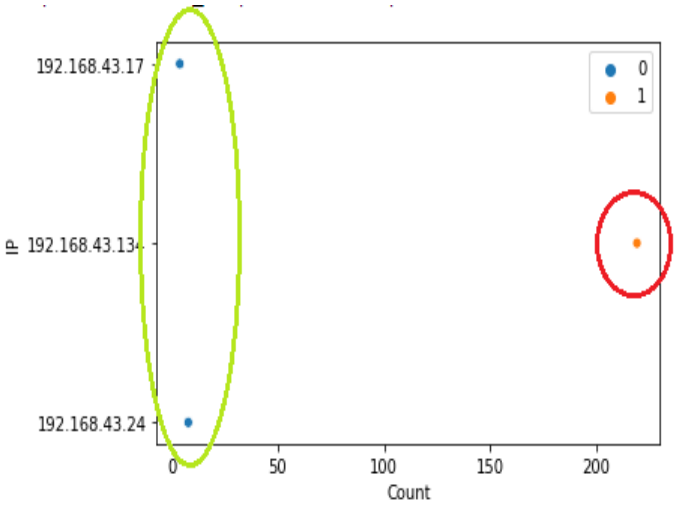

Fig 4.3.1 K-Means Clustering

In above image there are two cluster of IP addresses, one is of IP's with less than 200 hits on server and another with more than 200 hit on the same day. Now this data could be feed to K-Means model and it could predict that the IP should be Blocked or Not on the basis of Cluster.

4.4 Programming in Python

This project is built in the python. The whole machine learning code is written in the python programming language. The python code first filters the log file with the help of the regular expressions. This can arrange the data in some specific pattern so that the python modules can easily categorize them and save it to the csv file to use the data in further future processes. After this the csv file data is used to train the ml model. Python has lots of inbuilt machine learning supporting libraries which are very helpful of create the machine learning code. Python is the most suitable programming language for the machine learning programming. For using python code we have to install the python interpreter in the operating system and for machine learning programming we have to install external libraries.

4.5 Flowchart

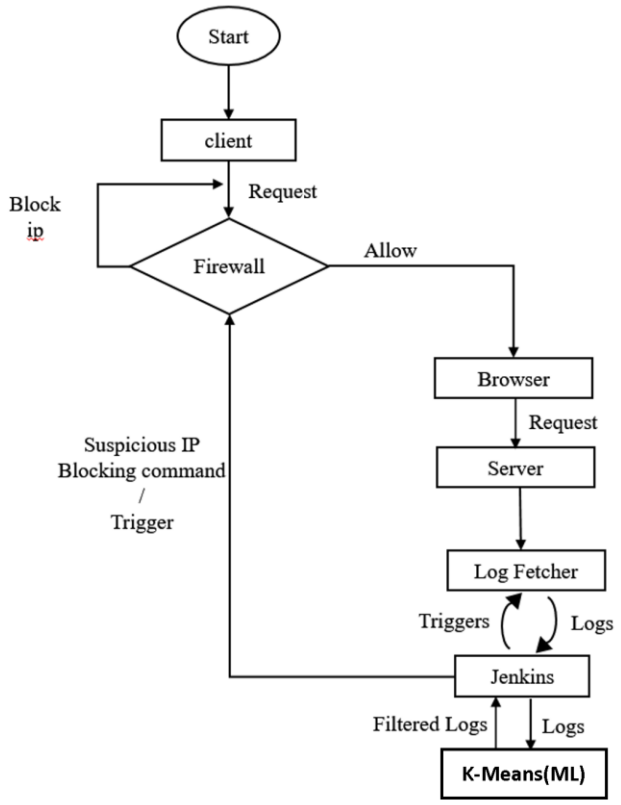

Fig 4.5.1 Flowchart of proposed system

5. LITERATURE SURVEY

In this research paper [1] the author R. K. C. Chang introduced that the internet is not stable due to the flooding-based distributed denial-of-service (DDoS) attack. In the last two years, it was found that the detection of DDoS-based attacks is a hitch because of the modern attacking tools and techniques. In this research paper we get know two things:-

1. Different types of DDoS attack method and all the current defense mechanisms.

2. How to kill the Internet-firewall approach, that attempts to intercept attack packets in the Internet core, well before reaching the victim.

As we studies another research paper[2] named as "D-face: an anomaly based distributed approach for early detection of DDoS attacks and flash events" written by the author S. Behal, K. Kumar, and M. Sachdeva. This paper proposes an ISP level distributed, flexible, automated, and collaborative (D-FACE) defense system. It distributes the storage and computational complexity to the nearest point of presence (PoPs) routers but also leads to an early detection of DDoS attacks and flash events (FEs). The results show that D-FACE defense system outperformed the existing Entropy-based systems on various defense system evaluation metrics.

For understanding the technical view of the Dos based attack we studies the research paper[3] which is about the In this the validation techniques used for DDoS research are investigated and it is proposed to extend them with the inclusion of new validation technique of analyzing real datasets. A brief review of existing real datasets is presented.

For better understanding of the machine learning for Clustering based semi-supervised machine learning for DDoS attack classification we studies the research paper[4] provides a clustering based approach to distinguish the data which represents the flows of network traffic which include both normal and Distributed Denial of Service (DDoS) traffic. The features are taken for victim-end identification of attacks and the work is demonstrated with three features which can be monitored at the target machine. The clustering methods include agglomerative and K-means with feature extraction under Principal Component Analysis (PCA). A voting method is also proposed to label the data and obtain classes to distinguish attacks from normal traffic. After labeling, supervised machine learning algorithms of k-Nearest Neighbors (kNN), Support Vector Machine (SVM) and Random Forest (RF) are applied to obtain the trained models for future classification. It is also validated using a subset of benchmark dataset with new vectors of attack.

6. Tools

The tools which WE have used our projects are Jenkins , python , Docker. Here Jenkins is used to access the log file from the web server which is running on the specific operating system.

6.1 Jenkins

Jenkins is an open-source automation tool written in Java with plugins built for Continuous Integration purposes, in this project Jenkins is used to give the command to the operating system firewall to block the suspicious IP or block those IP (which is given by the machine learning program) has the most of the request to the server. Jenkins generally gives the feature to write the operating system commands and apply the triggers to run the command on specific time of event.

6.2 Docker

Docker is an open source containerization technology which used to deploy the operating system within some seconds. Docker has its own commands. Docker is mainly use to deploy the web server, Here we are using this containerization technology to deploy and test our environment. Overall the docker is as open source project that is used to automate the deployment of the softwere application which is running inside the containers, here containers provides the isolation for the application on the other hand we can say that it provide the isolated part for the application.

7. ALGORITHM

In this section, we will discuss our algorithm which we have developed using different concepts as explained earlier. Accessing the log from the web server which include the time of request, IP address of the visitors and other information. Then we have to extract the essential information from the logs such as IP address and time of request. For this we can use machine learning code written in python. Python is the programming language which is used to create machine learning program. Python has lots of inbuilt libraries which support the machine learning code due to this python is most suitable language in compare to other programming language. In the machine learning we are using K-means algorithm to create the cluster of the IP with respect to time and the biggest cluster indicate the suspicious IP which is targets the web server.

7.1 Creating the cluster of the IP

K-means approach (Machine Learning)

K-Means is a concept of the machine learning . It is a clustering technique widely used to partition a collection of data in groups k automatically .It create the cluster of the data and find the specific specialty in the data. In this project it is used to find the IP cluster which hits the maximum request to the server in particular duration. And helps to know either the packets is normal or is Dos attack. We have used machine learning to get the accurate and fast result for the machine and provide the best possible result to client.

8. RESULT AND LIMITATION

8.1 Result

Proposed algorithm of the machine learning first of all filter the logs and extract the useful information just like time and IP address which is used to creates a cluster or we can said dataset of the IP's and return the IP to the application called Jenkins which is command the operating system's firewall to disallow or block the clustered IP and protect the web server from the attacker.

8.2 Limitation

As this setup is for only DOS attack and it's need some customization .Due to operating system configuration, this project is only work on configured operating system so for using on another operating system of web server it required the reconfiguration of all the software on the operating system and to attain full detection accuracy all routers on the internet will have to follow this detection scheme, because unavailability of this scheme in one router may cause failure to the detection and traceback process.

9. FUTURE WORK

In future a comprehensive study will be carried out on real time data collected from the university network security using ML technique . This will help to find the security of the attacks over the university network, so that appropriate firewall rules will be applied to the network. Future works include analysis of DDoS attacks based on the vulnerabilities of services such as Heartbleed and web brute force attack, enhancement in the multiple-class classification, self-configuration of the system, developing methods for correlating triggered alarms, and formulating protective measures.

10. CONCLUSION

After thorough review , it is concluded that network attacks are very harmful and IDS/IPS does not cater to the latest attack which are affecting the network. With the help of ML technique are playing vital role in accessing the severity of the attack and thus helping the organizations to take appropriate decision to restrict such attack .In this thesis , we proposed an approach to detect a Dos attack using Machine Learning approach .We present the design and implementation of the project . We build test bed, install the software, configure the network environment, and perform extensive tests. The purpose of this project is to help student gain better understanding and more hands-on experience on Internet security.

11. REFERENCES

1. R. K. C. Chang, "Defending against flooding-based distributed denial-of-service attacks: a tutorial," *IEEE Communications Magazine*, vol. 40, no. 10, pp. 42–51, 2002.

2. S. Behal, K. Kumar, and M. Sachdeva, "D-face: an anomaly based distributed approach for early detection of DDoS attacks and flash events," *Journal of Network and Computer Applications*, vol. 111, pp. 49–63, 2018.

3. S. Behal and K. Kumar, "Trends in validation of DDoS research," *Procedia Computer Science*, vol. 85, pp. 7–15, 2016.

4. M. Aamir and S. M. A. Zaidi, "Clustering based semi-supervised machine learning for DDoS attack classification," *Journal of King Saud University—Computer and Information Sciences*, 2019.

5. Sci-kit Learn, Machine Learning in Python, 2017. https://scikit-learn.org/stable/. Accessed November 5, 2017.

6. Types of Machine Learning Algoritms, 2017. https://towardsdatascience.com/types-of-machine-learning-algoritms-you-should-know-953a08248861. Accussed December 12, 2017.

7. Robinson, R. and Thomas, C., Ranking of machine learning algorithms based on the performance in classifying DDoS attacks, Proceedings of the IEEE Recent Advances in Intelligent Computational System (RAICS), Trivandrum, 2015, pp. 185-190

8. Cyber security and its Trends on internet, 2018. https://www.incapsula.com/ddos/attack-glossary/high-orbit-ion-cannon.html.Accessed February 5,2018.

9. DDoS Trend Report and its overview , 2018. https://www.cdnetworks.com/CDNetworks Q3 2017 DDoS%20Attack%20Trends%20Report EN 201712.pdf. Accessed February 26, 2018.

10. DDoS and Dos based Attacks and its trends , 2017. https://en.wikipedia.org/wiki/Denial-of-service attack. Accessed November 14, 2017.