



***Dissertation on***

**“DDoS cyber threat detection and prevention using  
Artificial Intelligence and Machine learning algorithms”**

*Submitted in partial fulfilment of the requirements for the award of degree of*

**Bachelor of Technology  
in  
Computer Science & Engineering**

**UE18CS390B – Capstone Project Phase - 2**

***Submitted by:***

<b>Uddhar Pujari</b>	<b>PES2201800413</b>
<b>Y.R Pavan Sai</b>	<b>PES2201800484</b>
<b>Gokul K M</b>	<b>PES2201800517</b>
<b>Swadin Madi</b>	<b>PES2201800638</b>

*Under the guidance of*

**Prof. Shanthala**

Assistant Professor

PES University

**June - November 2021**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
FACULTY OF ENGINEERING  
PES UNIVERSITY**

(Established under Karnataka Act No. 16 of 2013)

Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India



## PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)  
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

### FACULTY OF ENGINEERING

## CERTIFICATE

*This is to certify that the dissertation entitled*

**‘DDoS cyber threat detection and prevention using  
Artificial Intelligence and Machine learning algorithms’**

*is a bonafide work carried out by*

**Uddhar Pujari  
Y.R Pavan Sai  
Gokul K M  
Swadin Madi**

**PES2201800413  
PES2201800484  
PES2201800517  
PES2201800638**

In partial fulfillment for the completion of seventh semester Capstone Project Phase - 2 (UE18CS390B) in the Program of Study -Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period June. 2021 – Nov. 2021. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 7<sup>th</sup> semester academic requirements in respect of project work.

Signature  
**Prof. Shanthala**  
Assistant Professor

Signature  
Dr. Sandesh B J  
Chairperson

Signature  
Dr. B K Keshavan  
Dean of Faculty

### External Viva

**Name of the Examiners**

**Signature with Date**

1. \_\_\_\_\_  
2. \_\_\_\_\_

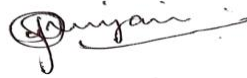
\_\_\_\_\_  
\_\_\_\_\_

## DECLARATION

We hereby declare that the Capstone Project Phase - 2 entitled “**DDoS cyber threat detection and prevention using artificial intelligence and machine learning algorithms** ” has been carried out by us under the guidance of Prof. Shanthala, Assistant Professor, PES University and submitted in partial fulfilment of the course requirements for the award of degree of **Bachelor of Technology in Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester June – November 2021. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PES2201800413

Uddhar Pujari



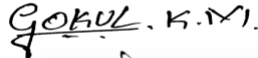
PES2201800484

Y.R Pavan Sai



PES2201800517

Gokul K M



PES2201800638

Swadin Madi



## **ACKNOWLEDGEMENT**

I would like to express my gratitude to Prof. Shanthala, Department of Computer Science and Engineering, PES University, for her continuous guidance, assistance, and encouragement throughout the development of this UE18CS390B -Capstone Project Phase – 2.

I am grateful to the Capstone Project Coordinator, Dr. Sarasvathi V, Associate Professor, for organizing, managing, and helping with the entire process.

I take this opportunity to thank Dr. Sandesh B J, Chairperson, Department of Computer Science and Engineering, PES University, for all the knowledge and support I have received from the department.

I would like to thank Dr. B.K. Keshavan, Dean of Faculty, PES University for his help.

I am deeply grateful to Dr. M. R. Doreswamy, Chancellor, PES University, Prof. Jawahar Doreswamy, Pro Chancellor – PES University, Dr. Suryaprasad J, Vice-Chancellor, PES University for providing to me various opportunities and enlightenment every step of the way. Finally, this project could not have been completed without the continual support and encouragement I have received from my family and friends.

## **ABSTRACT**

DDoS (Distributed Denial of Service) is a form of cyber-threat that is one of several versions of DoS (Denial of Service) that employs IP addresses to attack a specific server or victim. DDoS assaults are well-coordinated attacks that leverage compromised secondary victims to attack a single or numerous victim systems, whether they are huge firm servers or small scale systems. DDoS attacks are expensive in terms of bandwidth and electricity, and can also result in the loss of sensitive data. As a result, developing better algorithms to detect various types of DDoS Cyber Threats with greater accuracy while taking into account the computational cost of detecting these threats has become critical.

The majority of research in the literature treats DDoS threat detection as a binary classification problem, with the findings indicating whether an attack was attempted or not. However, knowing which form of DDoS assault is aimed at the network or system is critical in order to properly prevent the network from causing substantial damage. By converting the problem to a multilabel classification problem, this study presents an Ensemble Classifier that combines the performance of the top performing algorithm and compares it to different Artificial Intelligence and Machine Learning (AI and ML) algorithms to effectively detect different types of DDoS threats.

# TABLE OF CONTENTS

<b>Chapter No.</b>	<b>Title</b>	<b>Page no.</b>
<b>1.</b>	<b>INTRODUCTION</b>	01
<b>2.</b>	<b>PROBLEM DEFINITION</b>	03
<b>3.</b>	<b>LITERATURE SURVEY</b>	04
	3.1 Paper1	04
	3.1.1 Introduction	04
	3.1.2 Objective of paper, technique/methods	05
	3.1.3 Advantages	05
	3.1.4 Limitations	06
	3.2 Paper2	06
	3.2.1 Introduction	06
	3.2.2 Objective of paper, technique/methods	07
	3.2.3 Advantages	08
	3.2.4 Limitations	08
	3.3 Paper 3	08
	3.3.1 introduction	08
	3.3.2 Objective of paper, technique/methods	09
	3.3.3 Advantages	10
	3.3.4 Limitations	10
	3.4 Paper 4	11
	3.4.1 Introduction	11
	3.4.2 Objective of Paper, technique/methods	11
	3.4.3 Advantages	12
	3.4.4 Limitations	13

3.5 Paper 5	13
3.5.1 Introduction	13
3.5.2 Objective of Paper, technique/methods	14
3.5.3 Advantages	15
3.5.4 Limitations	15
<b>4. DATASET</b>	<b>16</b>
<b>5. PROJECT REQUIREMENTS SPECIFICATION</b>	<b>20</b>
5.1 Introduction	20
5.1.1 Project scope and motivation	20
5.2 Product perspective	20
5.2.1 Product Features	21
5.2.2 User classes and characteristics	21
5.2.3 Operating Environment	21
5.2.4 General constraint, Assumptions and dependencies	21
5.3 Functional Requirements	22
5.4 External Interface Requirements	22
5.4.1 User Interfaces	22
5.4.2 Hardware Requirements	23
5.4.3 Software Requirements	23
5.4.4 Communication Interfaces	23
5.5 Non-Functional Requirements	23
5.5.1 Performance Requirement	23
5.5.2 Safety Requirements	24
5.5.3 Security Requirements	24
<b>6. SYSTEM DESIGN</b>	<b>25</b>
6.1 Current System	25
6.2 Design Details	25
6.2.1 Interoperability	25
6.2.2 Performance	25
6.2.3 Security	26

6.2.4 Reliability	26
6.2.5 Maintainability	26
6.2.6 Portability	26
6.2.7 Re-usability	26
6.2.8 Application Compatibility	26
<b>6.3 Proposed Methodology/Approach</b>	<b>27</b>
<b>6.4 Simple Architecture of DDOS attack</b>	<b>28</b>
<b>7. IMPLEMENTATION AND PSEUDOCODE</b>	<b>32</b>
<b>8. EXPERIMENTATION RESULTS AND DISCUSSION</b>	<b>36</b>
<b>9. CONCLUSION AND FUTURE WORK</b>	<b>46</b>
<b>REFERENCES/BIBLIOGRAPHY</b>	<b>47</b>
<b>APPENDIX A DEFINITIONS, ACRONYMS AND ABBREVIATIONS</b>	<b>51</b>



### LIST OF FIGURES:

Figure Number	Title	Page Number
4.1.1	Types of DDoS Attacks	17
6.4.1	How a simple DDOS attack takes place	28
6.4.2	Simple DDOS Architecture	29
6.4.3	Data Flow Diagram	30
8.1	RoC Curve for Random Forest	37
8.2	RoC Curve for Decision Tree	38
8.3	RoC Curve for SVM	39
8.4	RoC Curve for Naive Bayes	40
8.5	RoC Curve for Multi Layer Perceptron	41
8.6	RoC Curve for LSTM	42
8.7	RoC Curve for XgBoost	43
8.8	RoC Curve for AdaBoost	44
8.9	RoC Curve for MVC	45

### LIST OF TABLES:

Table Number	Table Title	Page Number
8.1	Accuracy Score of Artificial Intelligence and Machine Learning Models	36
8.2	F1 Score of all Models	38

# CHAPTER -1

## INTRODUCTION

The DDoS attack is one of the most well-known and significant cyber attacks in modern years. The aim of initiating the DDoS attack is to absorb the victim's resources. The attacker sends a massive volume of traffic to the victim's side. As a result, these facilities cannot be used for a certain period of time and hence the service could not be delivered to legitimate consumers.

DOS is like making the victim's system inaccessible from any type of services.

The attack is going to happen through hijacking the servers, overloading the transmission link with larger packet size, port overloading, and overloading the server with multiple connection requests. This attack is performed through a single machine and can be monitored from a single machine. There are many types of dos attacks like.

- Ping of death (send packets with more packet size),
- reflector attack (sending multiple connection requests),
- mail bomb (attacks through emails),
- teardrop attack (it is going to attack fragment offset value).

DDoS attack refers to distributed denial of service attack. In this attack we are going to use multiple systems and monitor from multiple systems to attack victim's systems and make them inaccessible from using websites and make the system overload with multiple connection requests.

The effect of DDoS attacks may last for weeks. The attacker is going to establish a zombie network and attack the victim's systems. This zombie network contains multiple systems in which the attacker has injected malware code to attack the victim's system. Multiple connection requests will be sent from the zombie network and overload the victim's server. After the attack in the victim's system, response to any request or service will be much slower or will not be accepted at all.

The DDoS attack is performed to steal sensitive data. The DDoS attacks are increasing day by day, so we came up with an idea to detect these attacks. Several studies have been conducted on detecting and mitigating DDoS attacks. Study carried out years ago revealed that the ineffectiveness of detecting and mitigating DDoS attacks is directly related to constant configuration errors and wasted time due to the lack of tools that follow the dynamics of the network without constant human interference. It has led researchers to use autonomous solutions that can operate (detect and mitigate) based on the behavior and characteristics of the traffic.

## **CHAPTER - 2**

### **PROBLEM DEFINITION**

In our project, what we are trying to do is we are adopting different solutions with techniques based on artificial intelligence, mainly ML and DL has been distinguished by offering high flexibility in the classification process, consequently improving the detection of malicious traffic. By doing so, we are enhancing the system. This documentation gives a brief description about “DDoS Cyber Threat Detection Using ML, AI and DL models. In this study of the project we present an Ensemble Classifier that combines the performance of top 4 algorithms and compares it with different AI and ML algorithms to effectively detect the different types of DDoS Threats by converting the problem to a multilabel classification problem.

## CHAPTER - 3

### LITERATURE SURVEY

#### **3.1 Paper1: “A Machine Learning Based Classification Technique to Detect DDoS Attack in Cloud Computing Environment” by Abdul Moqeet (2021)**

##### **3.1.1 Introduction:**

One of the most well-known and significant cyber attacks in recent memory is the DDoS threat. It's one of the most popular and aggravating issues that cloud providers and consumers face. In recent years, many well-known cloud providers, such as Amazon EC2 and Rackspace, have been subjected to DDoS attacks, resulting in thousands of dollars in losses. In terms of long duration and high volume of traffic, these attacks are getting more serious and dangerous by the day.

Abdul Moqeet researched DDoS attacks and related security approaches in the cloud. In the literature, 48 methods for detecting and preventing DDoS attacks have been discussed. When opposed to other DDoS defence strategies, machine learning-based techniques are capable of detecting high-rate DDoS attacks with the greatest precision. For the diverse design of traffic, static attribute collection approaches are ineffective. To achieve optimum accuracy and low false detection of DDoS attacks, these techniques must be improved.

The following are a few things to think about when he conducts his studies.

- To reduce the time and space complexity of classification algorithms, identify the most important attributes and the smallest number of attributes.
- DDoS attack classification using various effective classifiers to increase accuracy, precision, recall and F1-score.

### **3.1.2 Objective of paper, Techniques/Methods:**

Based on a published research paper, he addressed the study of regular traffic and DDOS attacks in this report. He looked at how these types of TCP, UDP, and ICMP traffic behaved in both normal and attack situations. The I/O graph is used to analyse the behaviour of each traffic type. On a desktop, the virtual world is developed. Two virtual machines are mounted on the host computer in his experimental environment using Virtual Box 6.1. Microsoft Server 2017 is built on Guest 1's desktop, which serves as the victim computer. The HOIC and LOIC tools are used to perform TCP and UDP flooding attacks on the host (malicious) computer.

The following is a discussion of his analysis methodology.

- Utilize a variety of web tools to learn more about your research subject. Conduct a literature review to determine the benefits and drawbacks of the strategies under consideration.
- To get a better understanding of the system's actions, analyse all regular and DDoS flows.
- DDoS identification technique proposed based on machine learning.
- The proposed machine learning methodology would be tested in an academic setting.

Six classifiers are used against each sorting strategy for detection and prevention: Decision Tree (DT), AdaBoost (AB), Support Vector Machine (SVM), K-nearest neighbour (KNN), Random Forest (RF), and Multilayer Perceptrons (MLP) and Bayesian and J48 are two more best classifiers that have been discussed. Using the NSL-KDD dataset, he performed subsystem preprocessing, attribute extraction, and normalisation.

### **3.1.3 Advantages:**

His suggested machine learning-based classification methodology increased the accuracy of DDoS attack detection. To apply various machine learning methods, all attributes of the incoming traffic are normalised on a regular scale. CFsSubsetEval and Best-SearchFirst produce the most associated features, reducing the number of features. The least time taking classifier is determined by comparing

various classifiers on the basis of the time taken in seconds by each classifier. In comparison to all other classifiers, random forest has the best accuracy, according to this article.

### **3.1.4 Limitations:**

These attacks are difficult to detect and monitor since they come from a variety of sources, making it difficult to distinguish between attack packets and valid packets. In complex settings, static attribute collection methods are incapable of detecting DDoS attacks accurately. According to his findings, Random Forest (RF) is the best DDoS classifier for accurately classifying anomaly and natural classes with high accuracy, precision, recall, and F1-score. However, J48 has been identified as a strong competitor to Random Forest, and it needs less time to identify.

## **3.2 Paper2: “Clustering based semi-supervised machine learning for DDoS attack Classification” by Muhammad Aamir , Syed Mustafa Ali Zaidi (2019)**

### **3.2.1 Introduction:**

They employed semi-supervised machine learning to collect network traffic data using an unsupervised method, then used majority voting method to classify the data points to identify regular, DDoS and suspected traffic. Finally detect the unknown traffic class using a supervised method. They also project data points in low-dimensional space during the unsupervised learning process using the feature extraction approach of Principal Component Analysis (PCA). During the supervised learning process, researchers apply optimization within a given range of values and validation techniques to identify improved parameter configurations of machine learning models. Other semi-supervised machine learning-based research on marking and categorising DDoS traffic has revealed various complicated clustering algorithms as well as supervised models with little optimization required.

### **3.2.2 Objective of paper, Techniques/Methods:**

The basic premise of this work is that clustering methods inevitably produce a large percentage of false positives. As a result, utilising several clustering techniques and then evaluating them through voting will raise the confidence that a data point belongs to a specific class. When many clustering methods vote for a data instance to be placed in the same class, it can increase confidence while lowering the inherent false positive problem of a single clustering approach. When several clustering algorithms disagree on whether a data point belongs in the same class, there is a level of confusion that causes the data instance to be categorised as suspect. The basic premise of this work is that clustering methods will invariably produce a high proportion of false positives. As a result, employing multiple clustering techniques and then evaluating them by majority voting will increase confidence that a data point belongs to a specific class. As a result, an ensemble technique with a Bagging strategy could be utilised to categorise data points because classification in one cluster is independent of classification in another.

The paper's primary goal is to

- Find Gaps between different clustering implementations on the same dataset are observed and they are recorded.
- Trying to suggest a voting mechanism for labelling data points obtained from the clusters of different algorithms.
- It is demonstrated that after voting, the data of 'k' clusters can be labelled with 'k + 1' labels.
- Enhancing and validating learning models, which includes discovering the best parameter combination for a given set of values and increasing effectiveness of the model through K-fold cross validation. The dataset utilised in this study was created using network traffic generated in the OPNET Modeler 14.5 simulator (SteelCentral Riverbed Modeler, xxxx) (now termed "Riverbed Modeler"). Supervised learning algorithms are used to train the models and detect



clusters of unknown cases after the data points have been labelled. In this study, they employ the k-Nearest Neighbors, Support Vector Machine (SVM), and Random Forest methods.

### **3.2.3 Advantages:**

Employing dissimilarity metrics, semi-supervised machine learning can be used to discover subsets of unlabeled and partially labelled datasets. This study describes a clustering-based method for separating data characterizing network traffic patterns, including both conventional and Distributed Denial of Service traffic. In the KNN model, an unknown data point's class is determined by computing its separation from current (trained) points. The capability of SVM to handle multidimensional and dynamic data is well-known. However, it is sensitive to interference (noise) and overfitting.

### **3.2.4 Limitations:**

The paper's key restriction is that they can't use supervised learning approaches to classify unlabeled data, therefore they have to use semi-supervised and unsupervised learning models to label it using dissimilarity measures. When different clustering algorithms disagree about which class to place a data instance in, the data object is placed in the wrong group, causing ambiguity. Furthermore, only application layer protocols are employed in k-means classification.

## **3.3 Paper3: “Ranking of Machine learning Algorithms Based on the Performance in Classifying DDoS Attacks” by Rejimol Robinson R R , Ciza Thomas (2015)**

### **3.3.1 Introduction:**

The main purpose of this study is to rate and test a few standard machine learning algorithms in order to reduce type I and type II classification algorithm faults. DDoS detection systems require the ability to gather large amounts of network traffic data as well as the ability to scan that data for malicious packets or packet flows. Machine learning-based classifiers are aimed at spotting datasets by

identifying data features. By automatically developing rules for network intrusion detection systems, machine learning technologies can assist analysts in making better conclusions. Lowering the false alarm rate and ensuring detection accuracy are crucial aspects for machine learning algorithms' reliability. The algorithms' precision and recall are more essential in this work since they are critical factors in better classification of skewed datasets. The algorithms are reviewed and graded using Visual PROMETHEE, a Multicriteria Decision Aid software suite, because there is a tradeoff between precision and recall. Individual, hybrid and ensemble classifiers are the most common types of methods explored.

### **3.3.2 Objective of paper, Techniques/Methods:**

This study compares and rates a variety of supervised machine learning algorithms in order to reduce type I and type II mistakes, improve recall and precision and maintain detection accuracy. Using Multi Criteria Decision Aid tools dubbed Visual PROMETHEE, this study illustrates the effectiveness of ensemble based classifiers, particularly the Adaboost ensemble algorithm with Random Forest as the foundation classifier. Researchers wanted to evaluate ten different supervised machine learning techniques based on various metrics to see how well they could classify data. False positive rate, false negative rate, precision, recall and detection accuracy are the metrics employed and there is a tradeoff between the selected metrics.

The below are the stages of the experiment.

- a) Network trace packet header parsing
- b) Extraction of features
- c) The process of normalisation.
- d) Metric classification and evaluation
- e) Algorithm rankings

Naïve Bayes classifier, RBF network, Multi Layer Perceptron, Bayesnet, IBK, J48, Voting, Bagging, Random Forest and Adaboost are the algorithms employed in this research. The feature extraction

receives the parsed packet header information from traffic traces. The LLS-DDOS 1.0-Scenario was selected. CAIDA 2007 and CAIDA Conficker as attack traces from the DARPA Intrusion Detection Scenario-Specific datasets from 2000. DARPA is used to choose normal traffic. Three distinct training datasets are produced from these attack traces mixed with regular traffic. The data must first be preprocessed before it can be used by Weka. Weka is a widely used machine learning software package that may be used to execute a wide range of data mining activities. The features that are taken into account are Average Packet Size, Number of Packets, Time Interval Variance, Packet Size Variance, and Number of Bytes. The study by Karimazad and Faraahi looked at the seven features that provide adequate information to detect the presence of a DDoS threat. IBK, BayesNet, and J48 were all outperformed by Random Forest.

### **3.3.3 Advantages:**

Machine learning methods are used to classify DDoS flooding attacks in order to reduce the cost of misclassification due to Type I and Type II errors. With a bigger bias, the flooding assault paths are immediately distinct. The detection accuracy and performance of all of the methods is greater than 95%. The PROMETHEE II full rating process, which combines positive and negative preference flows into a single performance, considers the net result of two preference flows. Researchers have also employed ensemble classifiers because single classifiers produce errors on a variety of training examples. As a result, the overall errors in classification could be reduced by assembling together an ensemble of classifiers and integrating their outputs. Maintaining precision in identification, hence lowering false alarm rate.

### **3.3.4 Limitations:**

The output degradation is primarily due to the smaller variation of results when compared to other datasets. Integrity and availability are impossible to guarantee.

### **3.4 Paper4: “The hybrid technique for DDoS detection with supervised learning algorithms” by Soodeh Hosseini ,Mehrdad Azizi (2019)**

#### **3.4.1 Introduction:**

One of the most serious threats to web providers is DDoS (distributed denial of service). DDoS attacks may be used by attackers to prevent or slow down consumers' access to services by employing rapid steps and high performance. In this work, they propose an unique hybrid method based on data stream methodology for identifying DDoS attacks with gradual learning. They utilise an approach that divides the computing weight between the client and proxy sides based on their resources to coordinate the job at a high speed. As a result, the assault is recognised if the deviation exceeds a particular threshold. Otherwise, data is transferred to the proxy side.

Distinct assaults have different behaviours and different features picked for each algorithm result in acceptable output for detecting attacks and greater capacity to discern new attack types.

Based on the supplied goal and algorithm outcome, the researchers employed a combination of algorithms to identify the optimum course of action. There are two important contributions in this study. Due to the abundance of resources, the first major task is to divide and align the interaction between the client and the proxy in order to improve the outcome in a specific time frame. Then both parties, especially on the client's side, the goal is to achieve no overweight. Second, the assault must be recognised as soon as feasible to prevent the client from continuing the attack in the first place, either as a client or as a proxy.

#### **3.4.2 Objectives:**

A new hybrid architecture based on a data stream approach is provided in this paper for detecting DDoS assaults using a strategy that divides the computational weight between the client and proxy sides. Furthermore, by delegating part of the procedures to the customer, the provided work reduces

the time and cost of the intrusion detection system (IDS) process. Many algorithms combined will overcome the flaws of the others and enhance IDS detection rates.

This architecture can handle new attack styles better than any other framework currently available since algorithms have distinct processing ways. KNIME is an open source analytics framework that provides a scalable environment for dynamic process execution and easy visual assembly. This approach makes it simple to run an algorithm, read and edit data in various forms, simulate results and conduct a range of other tasks. They employed naive Bayes, random forest, decision tree, multilayer perceptron (MLP) and k-nearest neighbours (K-NN) to improve results on the proxy line. They examined the three primary works of the target, deployment and remarks. The objective with the exception of one work is the primary goal of each job, which is usually attack detection. The four types of deployment discussed are source side, network foundation, victim side and mixed deployments. The notes included a description of each work's main points. Data cannot be loaded into memory because its size grows exponentially. This issue uses a unique approach. Here are two strategies for dealing with this issue: Parallel processing: in order to reduce calculation time, the algorithm is divided into small, discrete pieces. Incremental processing is a one-pass approach for building and updating models that is used whenever a single data source is examined.

### **3.4.3 Advantages:**

The speed with which work may be organised can be boosted with this approach. The conclusions produced by KNIME are accurate because it presented its success in comparison to other works. The attack profile is saved in a database that can match data features to a particular profile to avoid over-processing on stream data. The a-divergence test was used to test the database.

An early detection approach for DDoS flooding attacks that allows you to take action quickly.

And to be gradually implemented in real-world networks like internet service providers (ISPs).

1. Complementary Processing: To reduce calculation time, the algorithm is broken down into discrete and independent pieces.

2. Gradual processing: To build a model, a one-pass technique is utilised and it is refreshed every time a new data point is retrieved. Handle massive amounts of data, while ensuring the safety of internet resources. The research presented here reduces the intrusion detection system's processing time and costs. Every other platform presented performs better in terms of attack styles.

### **3.4.4 Limitations:**

The following are some challenges with stream data learning:

1. Managing a huge number of streaming tasks such as forecasting, performing, filtering and so forth.
2. As data amount and complexity increase, scalability and performance become more important.
3. Analytics such as real-time data discovery and monitoring, continuous query processing, automatic alarms, responses and so on. Furthermore, KNIME users lack sufficient subject knowledge to communicate. It can be tough to employ memory at times.

## **3.5 Paper5: “An Efficient DDoS TCP Flood Attack Detection and Prevention System in a Cloud Environment” by Aqeel Sahi, David Lai, Yan Li, Member, IEEE, and Mohammed Diykh (2017)**

### **3.5.1 Introduction:**

DDoS TCP flood assaults degrade the cloud's resources quickly, drain the majority of its bandwidth and kill a whole cloud project. As a result, prompt detection and prevention of such assaults in cloud projects, especially in eHealth clouds is crucial. In this work, they present a new classifier approach for identifying and combating DDoS TCP flood attacks in public clouds (CS DDoS). The suggested CS DDoS approach provides a solution for preserving storage information by categorising incoming packets and making a choice based on the classification data. The CS DDoS identifies and determines whether a packet is natural or malicious during the detection phase. Malicious packets will be denied

access to the cloud provider during the prevention phase and the source IP will be blacklisted. A general strategy for categorising, classifying and discriminating diverse objects is classification. Classifiers such as least squares support vector machines (LS-SVM), naive Bayes, K-nearest neighbour and multilayer perceptron are employed in this study.

The structure of this article is as follows: They build on earlier work by including a simulation framework that may be used with or without DDoS TCP flood assaults. The proposed CS DDoS approach is described and the results are tested and validated. Finally, they bring this investigation to a close and look forward to future endeavours.

### **3.5.2 Objective:**

In this research, they present a novel classifier system for identifying and countering DDoS TCP flood attacks in public clouds. The proposed CS DDoS strategy protects stored information by classifying arriving packets and making a choice depending on the classification findings. CS DDoS identifies and decides whether a packet is natural or malicious during the identification phase. The least squares support vector machine (LS-SVM), naive Bayes, K-nearest, and multilayer perceptron classifiers are used to examine the efficiency of the CS DDoS approach. Several techniques for detecting and mitigating DDoS flood assaults have been described in recent years.

A TCP flood attack was carried out utilising software on a virtual cloud network.

Both before and after the attack, Wireshark Network Analyzer 2.0.0 was used to capture and evaluate traffic.

The rank correlation-based detection (RCD) approach was employed. RCD was confident in their abilities. The technique could identify if the requests coming in were from real users or attackers. The ALPi algorithm was built by broadening the principle of packet scoring, which reduced complications in packet flows and boosted functionality.

The findings show that employing LS-SVM, the CS DDoS framework can consistently identify attacks. The system has a 97 percent accuracy and a Kappa coefficient of approximately 0.89 when

subjected to a single attack, and a 94 percent accuracy and a Kappa coefficient of around 0.9 when subjected to several attacks.

### **3.5.3 Advantages:**

The K-fold cross-validation validation model is used to see if the results of a numerical test can be reduced to a single dataset. In general, it's used to check the correctness of a quantitative model's output forecast while it's running. The system is categorised to protect the secrecy and availability of stored data, which is especially important in emergency situations for eHealth data. Kappa coefficients are widely used to bind categorical variables as consistency or validity coefficients. There are three ways to use the proposed CS DDoS framework.

A normal service request envelope is the first method. The requested service will be provided in the usual way. When the source IP address is not blacklisted, but the quantity of service-request packets exceeds a predefined threshold in a given time period, the following situation occurs.

Finally, if the source address of a packet is blacklisted, the packet is dropped without being processed. CS DDoS allows for quick and accurate detection. RCD felt their system could determine if incoming requests were from real users or attackers. Both large-scale cloud projects, such as a health cloud, and smaller projects, such as a private cloud for a medium-sized firm, can benefit from the CS DDoS framework.

### **3.5.4 Limitations:**

Many of the DDoS attack security techniques presented here are unable to scale as networks become larger and faster. In the event of a flash crowd, all packets must be placed in a line and served in order.



## CHAPTER - 4

### DATASET

Our Project mainly depends on the Data, so it is important that our data is most up to date and dataset should be most recent in the contexts of these recent attacks. So our dataset collected consists of various important attributes that will be good for determining attacks efficiently and because if models will be trained with these attributes, they will be in better position for testing data and hence give better results for metrics like Accuracy, F1\_score, etc..

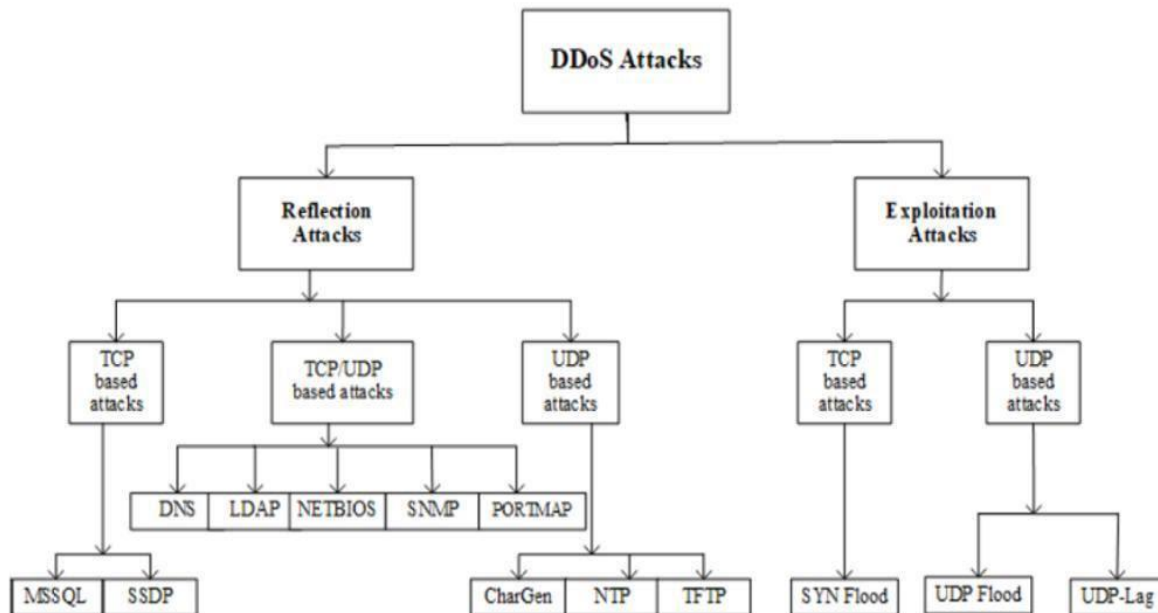
#### 4.1 Overview

Distributed Denial of Service is an attack where an attacker attacks from multiple systems and hides his identity and floods the victim system with constant flood of traffic.

There are basically 2 types of DDos attacks, One is Reflection based and the other is Exploitation based.

In the Reflection based DDos attack, the attacker hides his identity by using third party tools and components. This attack is initiated by sending data packets from the server with source and Ip address of target. These attacks are executed by either UDP or TCP or both. The UDP based attack includes NTP, TFTP and TCP based attacks include MSSQL, SSDP and TCP/UDP based combined attacks include DNS, LDAP, NETBIOS and SNMP.

In the Exploitation based DDos attacks, these attacks also use third party components and software to remain unidentified, these are similar to Reflection based . It has TCP based attacks and UDP based attacks. TCP based attack consists of SYN Flood whereas UDP based attacks consist of UDP-Flood(UDP) and UDP-lag. In a UDP flood attack, the target machine is flooded with a huge amount of UDP messages and thus the target system crashes. In UDP lag is carried out by disrupting the client server connection by lag switch or software that reduces bandwidth of the network. While in SYN flood attack, using 3 way handshake mechanism, sending SYN packets to target until it crashes.



**Fig 4.1.1 Types of DDoS Attacks**

There are many ways in which an attacker can attack a victim system. Attacker attacks victim through various several protocols as shown in the above Picture such as..

**DDOS DNS:** Here attacker attacks victim through Domain name Server of the victim. Exploits vulnerabilities in the DNS server.

**DDOS LDAP:** Here attacker attacks victim through Lightweight Directory Access Protocol Server of the victim.

**DDOS MSSQL:** In this attack the attacker uses new MSSQL reflection techniques for the attack. Exploitation of the Microsoft SQL Server Resolution Protocol (MC-SQLR) which is listening on UDP port 1434.

**DDOS NetBIOS:** DDOS attack happens through NetBIOS name server (Network Basic Input and Output system), NetBIOS is a programme that allows Application to communicate within the Local area Network (LAN).

**DDOS NTP:**Attacker here Exploits vulnerabilities in Network Time Protocol servers to overwhelm the target system with UDP traffic.

**DDOS SNMP:**Attacker exploits Simple Network management Protocol,here attacker sends SNMP queries with spoofed IP address to victim,and in return victim reply to this spoofed address.

**DDOS SSDP:**Attacker here exploits Simple Service Discovery Protocol (SSDP)server.It Exploits Universal Plug and Play (UPnP) Protocols to overwhelm the victim.

**DDOS UDP:** Attacker overwhelms victim network or server with constant flood of UDP messages.

**SYN:** Exploiting vulnerabilities in TCP/IP handshake,3 way handshake. floods victim system with a SYN acknowledgement message.

**TFTP:** Attack on TFTP(Trivial Time server File protocol) Server.

**UDP Lag:** Attack that disrupts Client and server Connection.

So these are the various ways through which an attacker may attack the victim.So our dataset consisting of DDOS attacks happened with various protocols ,eah saved as a CSV file.

## 4.2 Dataset:

The dataset we are using for this project is the CICDDoS2019 dataset,which consists of the most important and common DDoS attacks that have happened recently, which resemble true real world data.

It consists of labeled flows based on Source IP,Source port,Destination IP,Destination port,timestamp and protocol used during attack etc..which are included from the result of network traffic analysis. The main advantage of using this data set is :

1. As we have seen above there are many ways in which attacker may attack the victim using various protocols such as PortMap, NetBIOS, LDAP, MSSQL, UDP,

UDP-Lag, SYN, NTP, DNS, and SNMP. Attack with different protocols are stored as separate file.

2. The collected data is the most recent one.
3. The attributes present in the dataset are Source IP, Source port number, Destination IP, Destination Port, Total length of forwarded packet, Number of fwd Packets Timestamp ...etc these attributes will be good for determining attacks efficiently and because if models will be trained with these attributes, they will be in a better position for testing data and hence give better results for metrics like Accuracy, F1\_score.

The dataset has been organized per day and recorded in such a way that it includes event logs and network traffic. And from the raw data, Feature Extraction is done and more than 80 traffic features(some of them mentioned above) have been extracted.

Since the dataset is large, which is almost 26 GB and there are various csv file(based on protocol used), the important thing before applying or training this dataset to Artificial intelligence, Machine learning and Deep learning Algorithms is to reduce the size of dataset without losing the integrity of the data.

## **CHAPTER - 5**

# **SYSTEM REQUIREMENTS SPECIFICATION**

### **5.1 Introduction**

System Requirements Specification is basically a detailed document that explains the characteristics or features of the software application or the system. The characteristics of a good SRS are Correctness, Consistency, Completeness, Modifiability, Verifiability. The main reason for SRS is it helps provide precise instructions for the constructions of the Project. It will increase the quality and process of the project. This SRS gives a sufficient explanation about our topic. We are going to find the best machine learning and artificial intelligence model that detects Ddos attacks with highest accuracy. We will also specify required steps to protect the website from attack.

#### **5.1.1 Project Scope and Motivation**

DDOS attacks have been affecting the internet for a long time. They cause economic losses due to unavailability of services and serious security problems. So finding a countermeasure is required. The threats are very expensive in terms of bandwidth and power and loss of Confidential data as well. So better algorithms should be devised to detect these attacks. Early monitoring can prevent these attacks. Hence, it has motivated us to do this project.

### **5.2 Product Perspective**

The external view of our project is, which is the best machine learning model that detects the ddos attack. The main characteristic of our product is we have several datasets. And the datasets contain the users information and that information is already attacked with ddos attacks. And now we are going to

apply different machine learning models on the dataset by merging all datasets and we are going to detect which machine learning model has detected an attack with high accuracy.

### **5.2.1 Product Features**

The main feature of our product is that it is going to find which is the best machine learning model that detected the ddos attacks. By this feature the users are going to protect their confidential information from attack.

### **5.2.2 User Classes and Characteristics**

Many Organizations, Companies, Banks and those who want to protect their data use this type of product and this product is going to provide security to the data.

### **5.2.3 Operating Environment**

Any operating system such as Windows, Linux or Mac can be used for implementation.

### **5.2.4 General Constraints, Assumptions and Dependencies**

#### **Regulatory policies**

We should follow proper policies and rules while doing the work. Since we are going to attack the website that we have created ,we should not take that advantage and attack another website which has confidential information that may lead to some criminal cases.

#### **Hardware limitations**

Good Hardware is most important for a good executable project. Signal timing requirements is one of the important aspects. Attacking the website requires a large network. Badly designed hardware have high probability of getting attacks. If the websites are in secure less connection (HTTP) the chances of getting attacks are high. Hence good hardware requirements are required.

---

### **Limitations of simulation programs**

In many cases lack of precision leads to difficulty in measure. If any user system is affected from ddos attack there will be large traffic on his server and he is unable to access any information from the server.

### **Safety and security consideration**

A lot of precautions and details must be followed. The attacker should not misuse the information to attack another website. The user should maintain high quality hardware from preventing attacks. The user should maintain high quality passwords for a safe environment.

## **5.3 Functional Requirements**

The Data is collected from a large Dataset that is CIC dataset. We are going to pre-process, train and test the dataset using different machine learning models like SVM, decision trees and many ensemble learning and deep learning algorithms and we are going to find best machine learning model that detected the attack by finding F1\_Score and Accuracy rate to all the algorithms that applied. The consequences are if the attack happens loss of confidentiality happens. At-last We are going to find the best algorithm that detected the attack to the dataset.

## **5.4 External Interface Requirements**

### **5.4.1 User Interfaces**

Simple browsing system, the user network which was attacked by the attacker, The users confidential information is protected.

---

### 5.4.2 Hardware Requirements

GPUs are optimized for training artificial intelligence and deep learning models as they can process multiple computations simultaneously. GPU will decrease the training time. Basic system supporting a simple browsing system.

### 5.4.3 Software Requirements

The primary software on which all the programs for each of the Machine Learning algorithms are implemented in Python 3.7 along with the use of popular libraries such as NumPy and Pandas Python modules. In addition, for implementing the Artificial Intelligence models Keras is used as the application layer and TensorFlow library is used as backend support on Python 3.7. Jupyter notebook. Data Base.

### 5.4.4 Communication Interfaces

Browsers must allow access rights to examine the user information entered. The recent version of the browser would be best if the user wants a great experience from our product. We use the internet to communicate between the systems.

## 5.5 Non-Functional Requirements

### 5.5.1 Performance Requirement

- Highest Accuracy feature selection method for the dataset must be applied.
- Highest Accuracy train-test split for the dataset must be applied.
- Must display only the model combinations that have highest accuracy.

**Reliability:** Project is dependent on the dataset extracted from CIC dataset and performs best on moderately powerful systems.



**Robustness:** The Model we are building is capable of classifying the input correctly with acceptable minor errors.

**Availability:** Since our model is a freely available open-source model, it can be used by any individual.

**Accuracy:** We expect or intend our model to achieve a higher accuracy compared to other previous works by at least 2% to 4% thus having a significant improvement over the previous one.

### 5.5.2 Safety Requirements

- Buy more bandwidth.
- Build redundancy into your infrastructure.
- Configure your network hardware against DDos attacks.
- Deploy anti-DDos hardware and software modules.
- Deploy a DDos protection appliance.
- Protection of DNS servers.

### 5.5.3 Security Requirements

- Securing the network infrastructure is important.
- Practicing basic network cyber security is appreciable.
- Maintenance of strong network architecture is required.
- Leveraging the cloud is required too.

## **CHAPTER - 6**

### **HIGH LEVEL SYSTEM DESIGN**

#### **6.1 Current System:**

Several studies have been conducted on detecting and mitigating DDos attacks, a study carried out years ago revealed that there is ineffectiveness of detecting and mitigating DDoS attacks that are directly related to constant configuration errors and wasted time due to the lack of tools that follow the dynamics of the network without constant human interference. This has led researchers to use autonomous solutions that can operate (detect and mitigate) based on the behavior and characteristics of the traffic.

In this sense, the adoption of solutions with techniques based on Ensemble classifier that combines the best artificial intelligence, machine learning (ML) and Deep learning models based on the metrics such as Accuracy ,F1\_score,ROC(Receiver Operating Curve) this solution has lead to offering high flexibility in the classification process, consequently improving the detection of malicious traffic.By doing so we are enhancing the system.

#### **6.2 Design Details:**

##### **6.2.1 Interoperability**

This application is interoperable with different operating systems.If the requirements of the applications are met then the application can run in that environment without any problem. Our library will be interoperable with models trained using tensor flow and keras.

##### **6.2.2 Performance**

Using learned models reduces the testing time and increases the performance. We store the dataset on the system and aim to improve the accuracy of the same. This will increase the

support-ability of the model.

### **6.2.3 Security**

The model depends on the data for learning purposes, hence it should not be corrupted otherwise it could bring the system to failure .The computer system should be protected with an antivirus system to avoid the malware attacks.

### **6.2.4 Reliability**

Since we are determining models based on highest accuracy ,the selected models will be reliable in determining the attack. If internet traffic is not monitored regularly then ddos attacks may take place.

### **6.2.5 Maintainability**

Ddos detection needs to use better algorithms to make sure it shows the right result. These algorithms have to be tested and updated frequently to ensure good working of the application. Maintenance would be easy with new efficient algorithms compared to existing algorithms.

### **6.2.6 Portability**

This application is python based .So it can run on any OS, provided it has python installed in it. Also the System should support various Frameworks which have been implemented in the Application Frameworks like TensorFlow ,Keras should be installed.

### **6.2.7 Re-usability**

The main focus of the library is to make the task of Modelling Structured Data reusable for many applications depending on the user's need .We can also reuse the individual modules to different types of attacks.

### **6.2.8 Application compatibility**

All modules and functions in the application will be coded in Python v3.7. The application should be compatible with different versions of PC OS ,so if the version of OS is upgraded, the application should be upgraded accordingly.

## 6.3 Proposed Methodology/Approach:

The DDoS threats are very costly in terms of bandwidth and power and they result in loss of confidential data as well. Therefore it has become of much importance to devise better algorithms to detect different types of DDoS Cyber Threats with higher accuracy while considering the computation cost in detecting these threats.

By Detecting these types of attacks We can protect the users from losing confidential information. Early monitoring/detecting these types of attacks we can prevent this attack early so that no huge data will be lost.

By early detection many organizations and companies avoid huge loss of confidential information. Ddos attacks target websites and online services. The aim is to overwhelm them with more traffic than the server or network accommodates. The goal is to render the website or service inoperable. The traffic can consist of incoming messages, requests for connections, or fake packets. So Our study presents an Ensemble Classifier that combines the performance of top performing algorithms and compares it with different Artificial Intelligence and Machine Learning (AI and ML) algorithms to effectively Detect the different types of DDoS threats by converting the problem to a multilabel classification problem. For example we will be using Algorithms like Decision Tree, Random forest, SVM etc.. and decide the best model based on accuracy of the model to use while detecting DDoS attack.

## 6.4 Simple Architecture of DDos Attack:

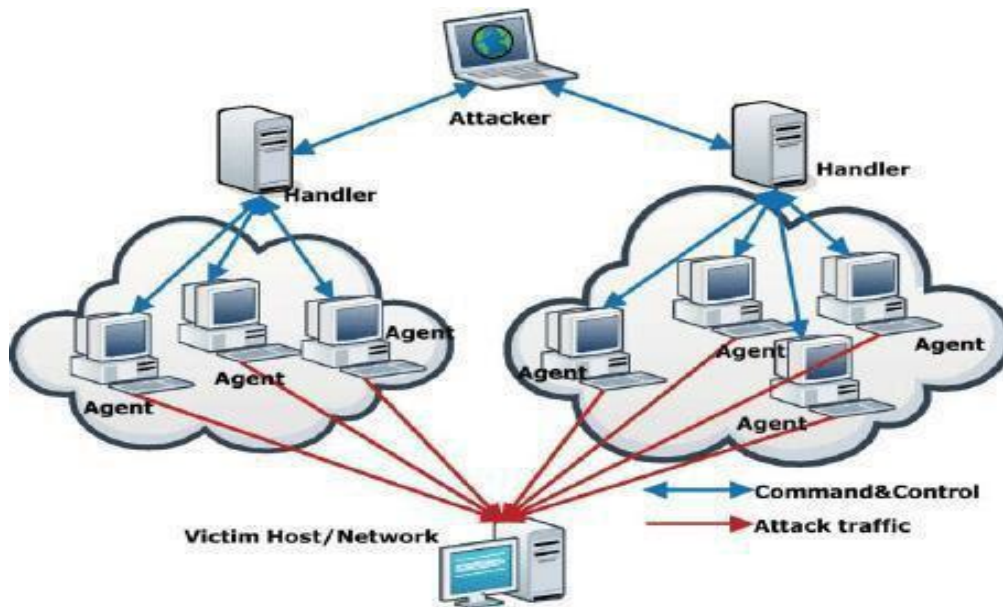
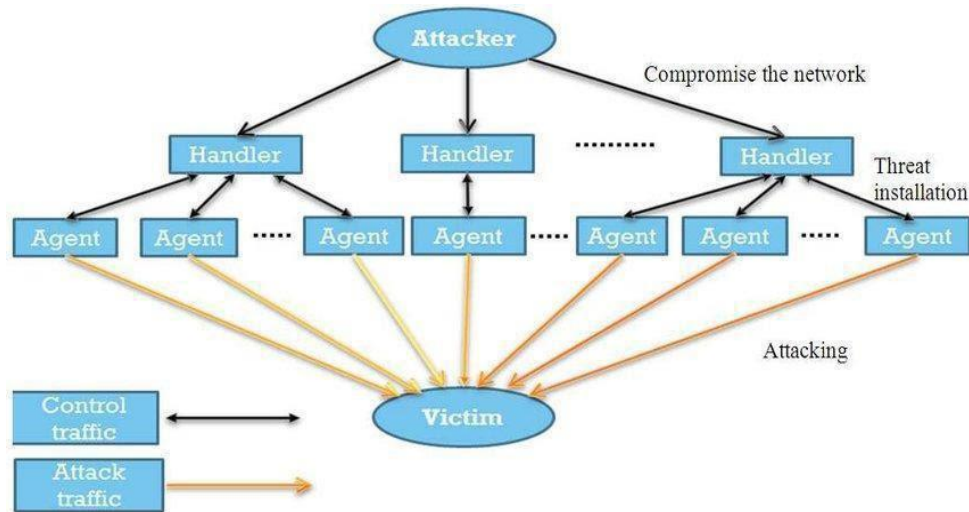


Fig 6.4.1 The above diagram indicates how a simple DDOS attack takes place.

- **Attacker Machine:** One who is attacking the victim's machine
- **Handler/master:** Server which is handled by Attacker
- **Slave/Agent:** Is a single or multiple machine managed by attacker through handler
- **Victim:** Is the machine on which attacker is trying to attack

In a **Simple DOS(Denial of Service)**, Attacker attacks the victim server with a single system, And floods the victim system with flood of traffic. Since Attacker attacks from single system it can easily be mitigated from closing the connection to the server where the attack is coming from.

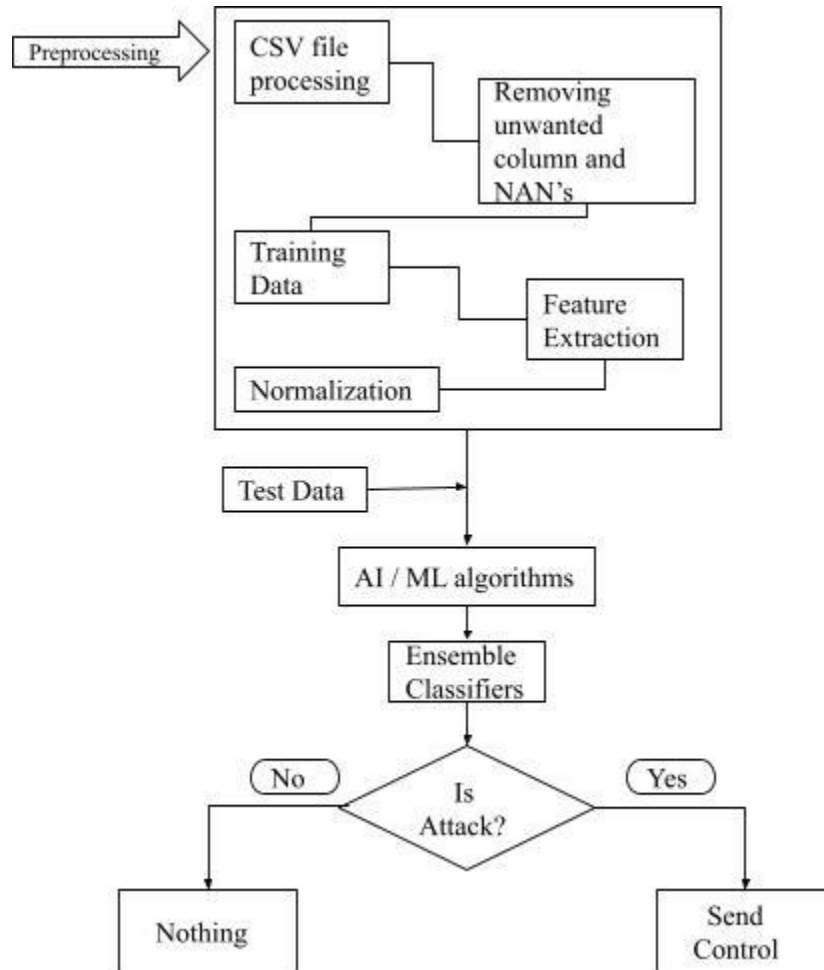
In **Distributed denial of service(DDOS)**, Flooding the target server/Network with constant flood of traffic which overwhelms the target system causing a denial of service to its intended traffic.



**Fig 6.4.2 Simple DDOS attack architecture.**

Since, attack with single system can be mitigated easily , Attacker attacks from Multiple system, Attacker take control of these multiple system by installing malicious software or malware into these system through websites,EMail attachments,internet,the normal system that open these attachments or visits these websites attacker will get hold of these system and start controlling them and form these system into Zombie network or Botnet,from these network attacker will flood the victim with flood of messages and thus hiding his true identity.Attacker will handle all the system in the system ,these army of the system will be waiting to receive instruction from the attacker,for example what system to attack, when to attack the system etc..

## Data Flow Diagram:



**Fig 6.4.3 Data Flow Diagram.**

Steps in how we are going to proceed with the project.

- Data set collected from CIC 2019 Dataset ,has to be preprocessed
- Data Pre Processing consists of CSV file processing, and removing the unwanted attributes(which does not help in Analysis Process) .

- 
- Then Conversion of Dataset into Training and Testing Dataset,from the training data feature extraction is done,which is followed by normalization.
  - Then we are considering different models for project
    - 1. Machine Learning Algorithms.**
      - Support Vector Machine
      - Decision Tree
      - Naive Bayes
    - 2. Ensemble Learning Algorithms.**
      - Random Forest
      - Extreme Gradient Boosting
      - Adaptive Boosting
      - Majority Vote Classifier
    - 3. Deep Learning Algorithms.**
      - Multi layer perceptron
      - Long Short Term Memory.
  - After training these models, Testing for these models is done.
  - Ensemble Classifiers then combines best of these models to detect attack
  - And Lastly if it is an attack then Send Control or notify the Victim saying that attack is being done on the system,so that it will be better prepared for that attack.And if it does not detect any attack, it does nothing.



## CHAPTER - 7

### IMPLEMENTATION AND PSEUDOCODE

#### 7.1 Software used for Implementation

Python 3.7 is the major platform on which all of the programmes for each of the Machine Learning methods are written, utilising popular libraries like Numpy and Pandas Python modules. Furthermore, Keras was used as the application layer and the Tensor flow library was used as the backend support on Python 3.7 to create the Artificial Intelligence models.

#### 7.2 Evaluation Metrics

##### 7.2.1 Accuracy Score

Accuracy score is a popular metric in machine learning for determining the correctness of models. It measures the number of properly predicted data points out of all data points. The accuracy score indicates how near a value is to another.

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y} = y_i) . \quad \longrightarrow \text{Equation (1)}$$

##### 7.2.2 F1 Score

**F1 score** is another metric that is used in Machine Learning. It is a weighted average of Precision (P) and Recall (R). Precision in simple words is What proportion of positives identified were actually correct. While Recall is What proportion of actual proportion were identified correctly. F1-score has a maximum value of 1 and is also known as the Dice similarity coefficient.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} \rightarrow \text{Equation (2)}$$

### 7.2.3 Receiver Operating characteristic curve - ROC

The Receiver Operating Characteristic curve (RoC) is a graph or curve that depicts the relationship between the True Positive Rate and the False Positive Rate. It's a tool for assessing the effectiveness of classification models. TPR is on the ROC curve's Y axis, while FPR is on the X axis. When FPR is zero and TPR is one, the classifiers are predicting all the data points correctly, there are no data points that are falsely predicted. The bigger the area under the curve, the higher the classifier's performance.

## 7.3 Algorithms and Pseudocode

### 7.3.1 SVMs (Support Vector Machines)

SVMs are a family of supervised learning algorithms that are commonly used for classification and regression. SVM a hyperplane that splits various classes and creates a classifier that works. It generalises well because it is based on unseen examples. The hyper variables included were fine-tuned to avoid overfitting. LinearSVC is employed in this work. Sci-Kit to effectively utilise a multiclass parameter as one-vs-rest "ovr," learn how to use it as one-vs-rest "ovr." For the classification of several labels, the loss function was chosen to be square-hinge. Function makes advantage of elementary mathematics to provide conclusions that are computationally efficient. The cost parameter or regularisation parameter was set to 1, which implies that samples within the margin will be penalised more, so they will try to correctly classify with a higher C value. To avoid overfitting, a big value of C was not used and if C was set too low, less than 1, it resulted in a soft margin. Because l1 normalisation produced excessively sparse coefficients, l2 normalisation was employed in penalization.

### **7.3.2 Decision Tree**

Decision Trees are supervised learning techniques that sort a tree from root to leaf nodes. The categorization, which is the label name, is given by the leaf node. Hyperplanes/axis-parallel rectangles are used in decision trees to partition the feature space into classifications. Because Decision Trees are less prone to outliers, they require less data processing. It serves as a starting point for various algorithms. We utilised the Gini Index as the splitting criteria and 3 as the sample split value when implementing Decision Trees. At each node, the best split is chosen as the splitting technique. Pruning was not carried out in order to keep the cost complexity. The number of features analysed to find the optimal split was set to the highest number possible.

### **7.3.3 Random Forest**

A random based classifier is a set of decision trees that are randomly chosen from a subset of the training set, and then the votes are randomly aggregated from all the decision trees, yielding the final class of the object tested. This classifier is mostly utilised because it is highly efficient with large datasets, but it can also be used for other purposes. Without deleting any variables, handle a huge number of input variables. Furthermore, it Additionally, it minimises overfitting by increasing the accuracy score while training. Furthermore, as the forest grows, it creates unbiased data. Estimates of generalisation error The parameters that give this classifier the best accuracy score are: 100 estimators, minimum sample leaves of 1, minimum sample split of 2, and the Gini criterion is used to determine the quality of the split.

### **7.3.4 Extreme Gradient Boosting (XGBoost)**

Extreme Gradient Boosting (XGBoost) is a sophisticated technique that works best with unstructured data and uses the gradient Boosting method (GBM). The gradient descent algorithm is used in this type of boosting strategy and it is used to reduce the number of errors. XGBoost, on the other hand, improves the GBM by Parallelization, tree pruning, hardware optimization, regularisation, and sparsity are some of the techniques used. In all circumstances,

this technique is very scalable. On many memory-constrained systems, it has faster computational performance. As a measure of the quality of a split, "friedman mse" is used. Friedman's improvement score is calculated using Mean Squared Error (MSE). This criterion provides the best approximation in this study. Furthermore, it classifies using probabilistic outputs utilising deviance as a loss function, which is equal to logistic regression.

### **7.3.5 Adaptive Boosting**

Adaptive Boosting, often known as Adaboost, is an ensemble method that uses an iterative methodology to learn from the mistakes of weak classifiers and convert them to strong classifiers. As a result, this algorithm outperforms other learning algorithms. Algorithms that make predictions based on random guesses. The base estimate that was taken into consideration for Decision Trees is the name of this classifier.

As a result, Adaboost aids in the improvement of accuracy. the fundamental estimator However, when compared to other methods, it is computationally slow. and it is highly susceptible to outliers and noise. In addition, Gini is responsible for this algorithm to assess the split's quality.

### **7.3.6 Majority Vote Classifier**

If a class label receives more than 50% of the votes, Majority Voting is an Ensemble learning technique that selects the class label that has been predicted by the majority of the classifiers. In this research, we've combined the best approximation for each class using the top four performing classifiers. Different DDoS threats have different labels. The MV-4 classifier is the name given to the combination of the top four classifiers in this study to distinguish it from other classifiers. RandomForest, AdaBoost, Decision Tree, and XGBoost are the top four performing models, as shown in Table. As a result, we use the Majority Voting Technique to combine these four classifiers to create the MV-4 classifier.

The code was written from the ground up in Python3.7 to achieve the combination of four separate techniques.

## CHAPTER - 8

### EXPERIMENTATION RESULTS AND DISCUSSION

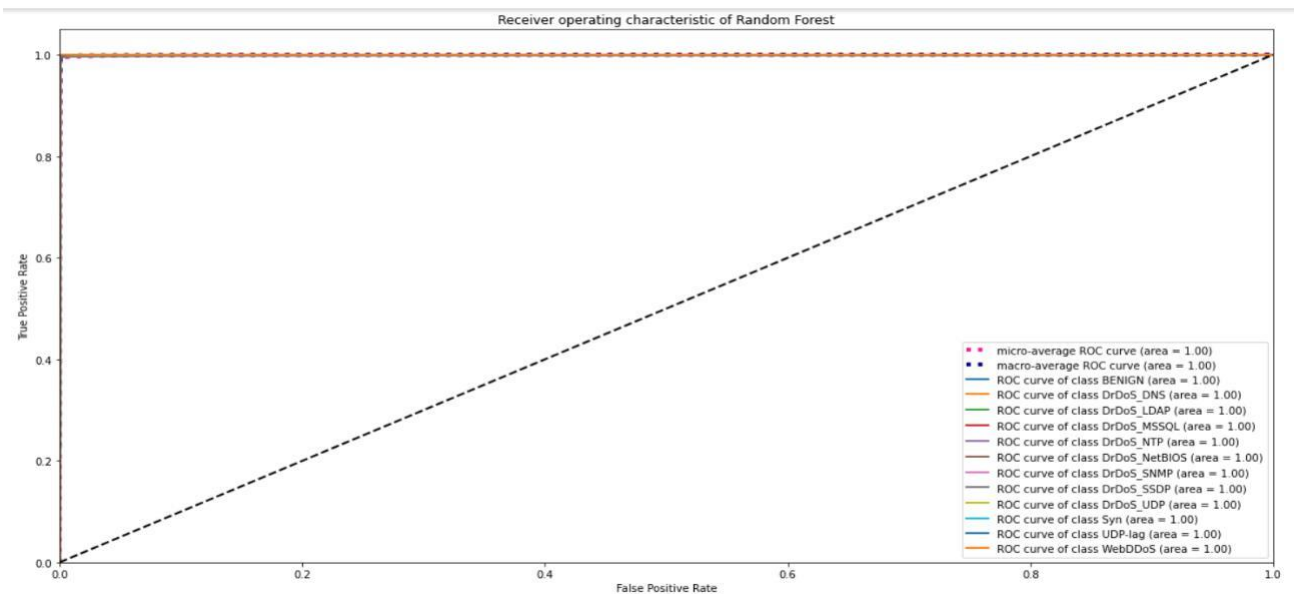
In this section, we discuss and analyze the results obtained from each of the algorithms on the CICDDoS2019 dataset.

**Table 8.1 Accuracy Score of Artificial Intelligence and Machine Learning Models.**

Algorithms	Accuracy Score
Random Forest	99.35
Decision Tree	99.12
SVM	92.25
Naive Bayes	81.38
MLP	88.37
LSTM	95.38
<b>XGBoost</b>	<b>99.51</b>
AdaBoost	99.11
MVC	99.11

**Random Forest** has a 99.35% accuracy rate. This accuracy value is approximately 7.08 percent higher than the SVM algorithm. Table 8.2 shows the F1-Score for Random Forest, which clearly shows that it has the highest F1-Score of all the algorithms in Table 8.2. It has six threats that are perfectly identified, as evidenced by the F1-Score of 1.00: MSSQL, NetBios, SNMP, SSDP, SYN, and UDP..

Furthermore, Benign trac has an F1-Score of 0.96, which is higher than all other algorithms, indicating that it can efficiently distinguish between normal and pathological traffic. Figure 8.1 shows the RoC Curve for Random Forest, which appears to be an ideal classifier for this dataset because the area under the curve for detecting all attack types is 1.00. In addition, the micro and macro averages both have a 1.00 score.



**Figure 8.1** RoC Curve for Random Forest

**Decision Tree** has a multilabel accuracy score of 99.12 percent, which is around 6.867 percent higher than SVM and slightly less than Decision Tree. Furthermore, the F1-Score for Decision Tree is significantly higher for all attack categories except Benign. Based on the metrics study, it is clear that it outperforms the most models for this dataset. By the analysis of RoC Curve for Decision Tree as Shown in Figure 8.2, it can be seen that area under the curve for all the attacks is higher than 1 and for some attacks like BENIGN area under the curve is very low, it is 0.38. So it does not effectively detect all the type of attacks hence can't be chosen for detecting Benign Threats.

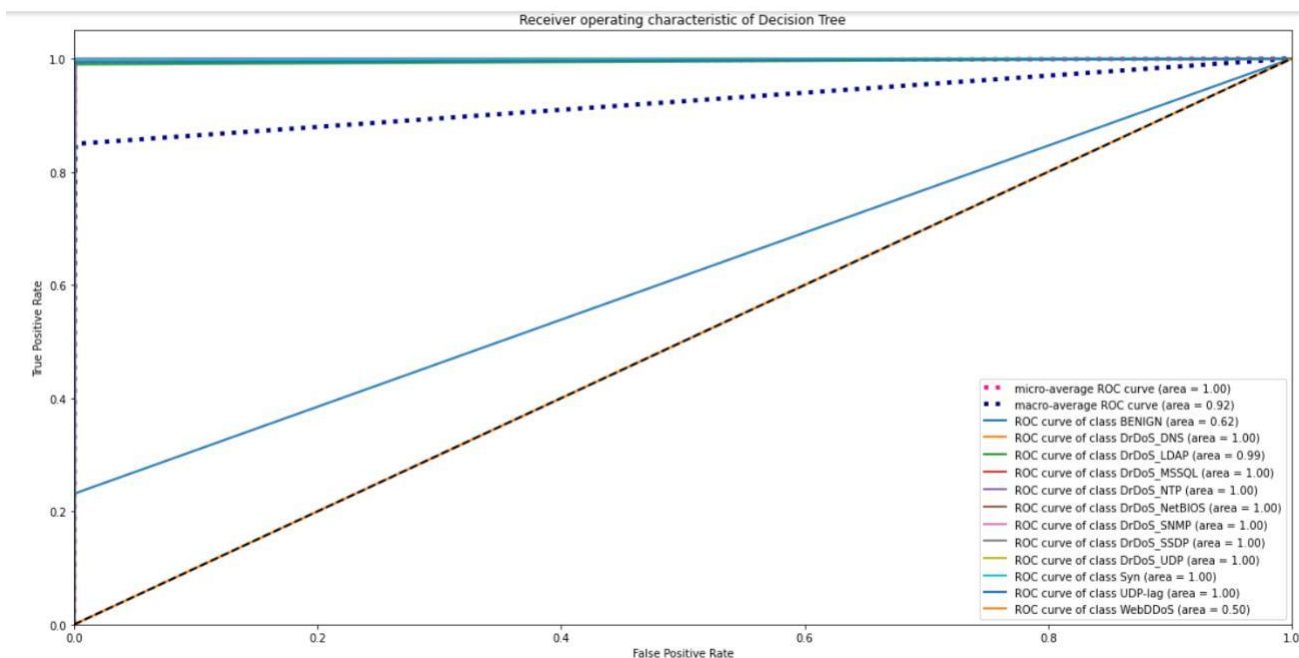


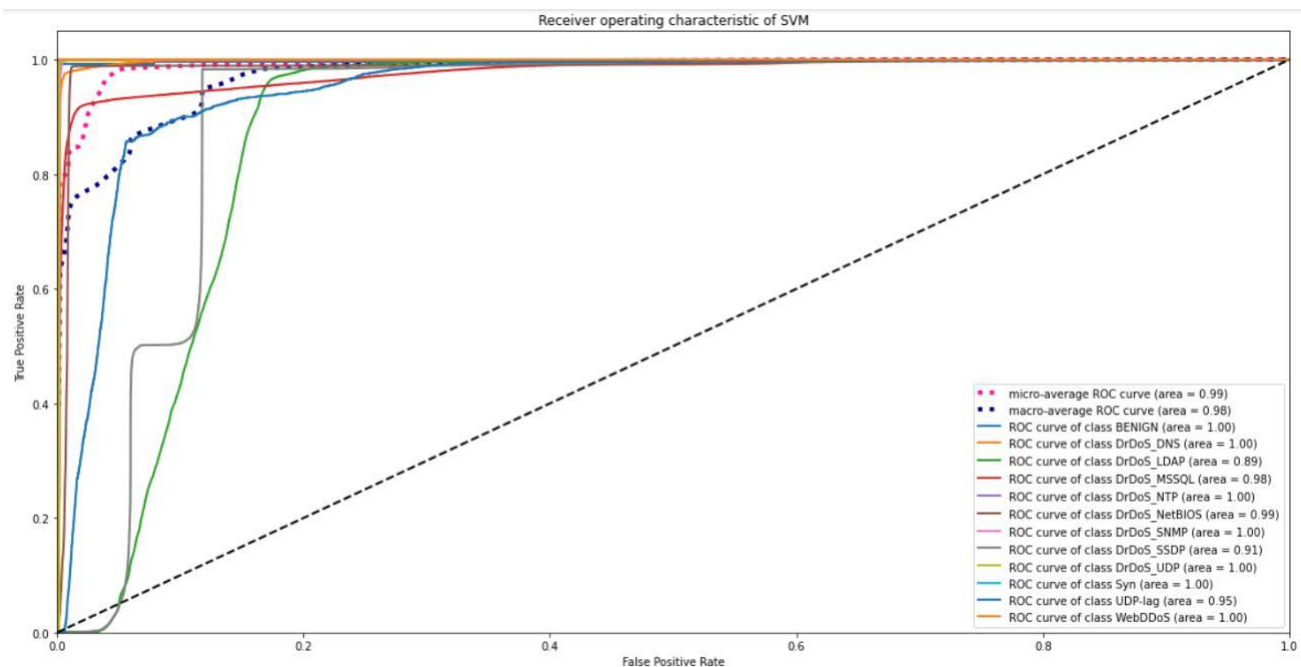
Figure 8.2 RoC Curve for Decision Tree

Table 8.2 F1 score of all the Models.

Threats	Random forest	Decision Tree	SVM	Naive bayes	MLP	LSTM	XG Boost	Ada boost	MVC
BENIGN	0.96	0.38	0.84	0.56	0	0	<b>0.32</b>	0.1	0.1
DNS	0.99	1	0.90	0.52	0.88	0.96	<b>1</b>	1	1
LDAP	0.99	0.99	0.73	0.6	0.73	0.92	<b>0.99</b>	0.99	0.99
MSSQL	1	1	0.94	0.95	0.95	0.99	<b>1</b>	1	1
NTP	0.96	0.99	0.97	0.97	0.95	0.97	<b>0.99</b>	0.99	0.99
NetBIOS	1	1	0.93	0.96	0.93	0.97	<b>1</b>	0.98	0.98
SNMP	1	1	0.97	0.97	0.97	0.97	<b>1</b>	1	1
SSDP	1	1	0.92	0.59	0.68	0.94	<b>1</b>	1	1
UDP	1	1	0.96	0.8	0.83	0.96	<b>1</b>	1	1

SYN	1	1	0.94	0.91	0.91	0.9	1	1	1
UDP-Lag	0.98	0.99	0.41	0	0	0	<b>0.99</b>	0.99	0.99

The accuracy score of the **Support Vector Machine (SVM)** is 92.25 percent, as shown in Table 8.1. And also Table 8.2 reveals that the F1-Score for each DDoS is not properly detected by this model. The F1-Score for DNS Traffic is 0.90, indicating that the system can distinguish between normal and anomalous traffic. By the analysis of RoC Curve for SVM as Shown in Figure 8.3, it is clear that for some attacks such as DNS, NTP, Syn etc. have a higher area under the curve while for other attacks such as LDAP, SSDP, etc have a lower area under the curve which is consistent with the F1-Score obtained in Table 8.2. So the Model has higher area under curve for only some type of attacks, so It is not suitable for detecting all the types of attack.



**Figure 8.3** RoC Curve for SVM



The **Naive Bayes** accuracy Score is 81.38 percent as you can see from table 1. This method has lowest accuracy score. From table 8.2, we can conclude that the F1-Score for MSSQL, SNMP and NetBios are more so these threats can be identified by this algorithm. Figure 8.4 shows that RoC Curve for this algorithm and the area under the curve for NTP and UDP is 1. Macro average is lower than the SVM and Random forest.

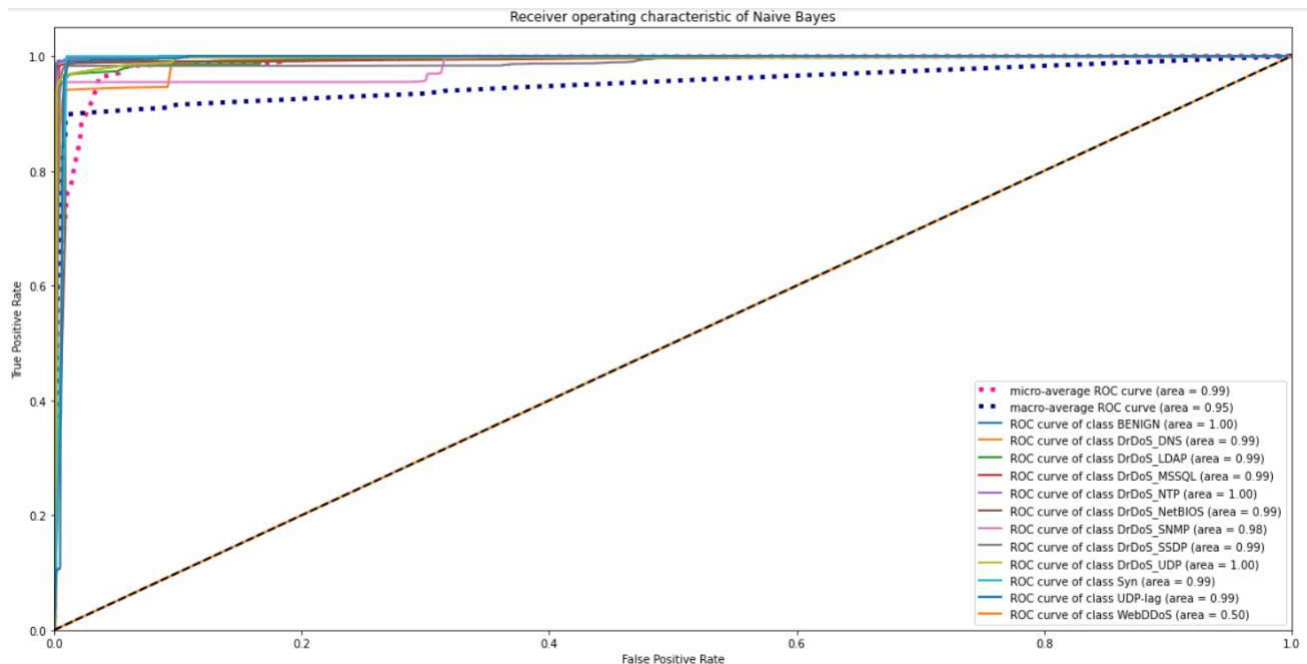
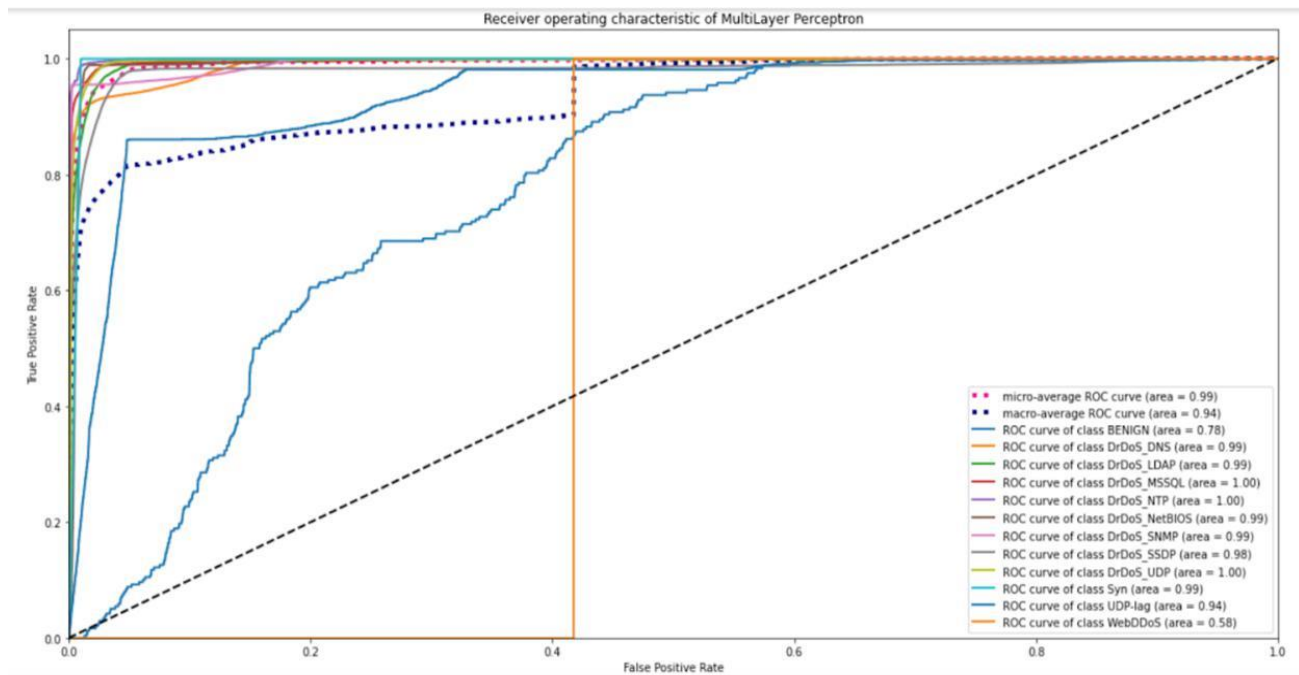


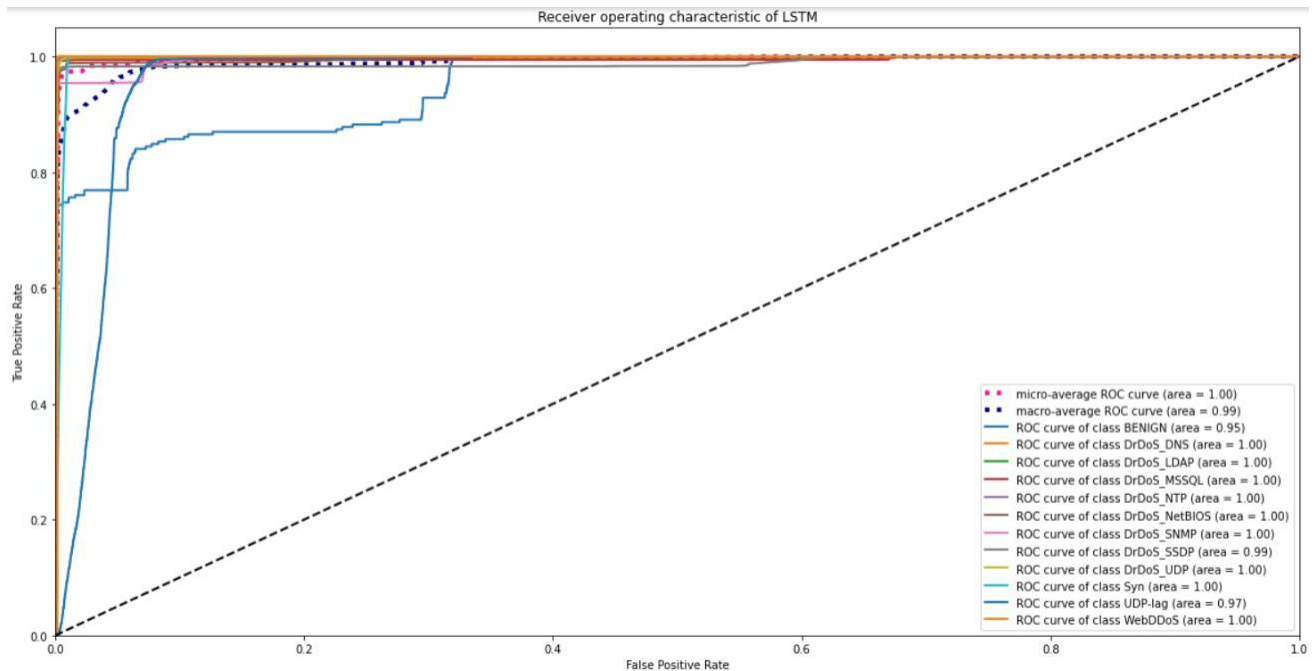
Figure 8.4 RoC Curve for Naive Bayes

The **MultiLayer Perceptron** has an accuracy of 88.37. And, across all of the methods listed in Table 8.1, MultiLayer Perceptron has the second lowest accuracy score. Table 8.2 reveals that the F1-Score for each DDoS is not properly detected by this model. The F1-Score for DNS Traffic is 0.88, indicating that the system can distinguish between normal and anomalous traffic. By the analysis of RoC Curve for MultiLayer Perceptron as Shown in Figure 8.5, it is clear that for some attacks such as DNS, NTP, Syn, SNP, SSDP etc. have a higher area under the curve while for other attacks such as UDP\_Lag, WebDDoS, etc have a lower area under the curve.



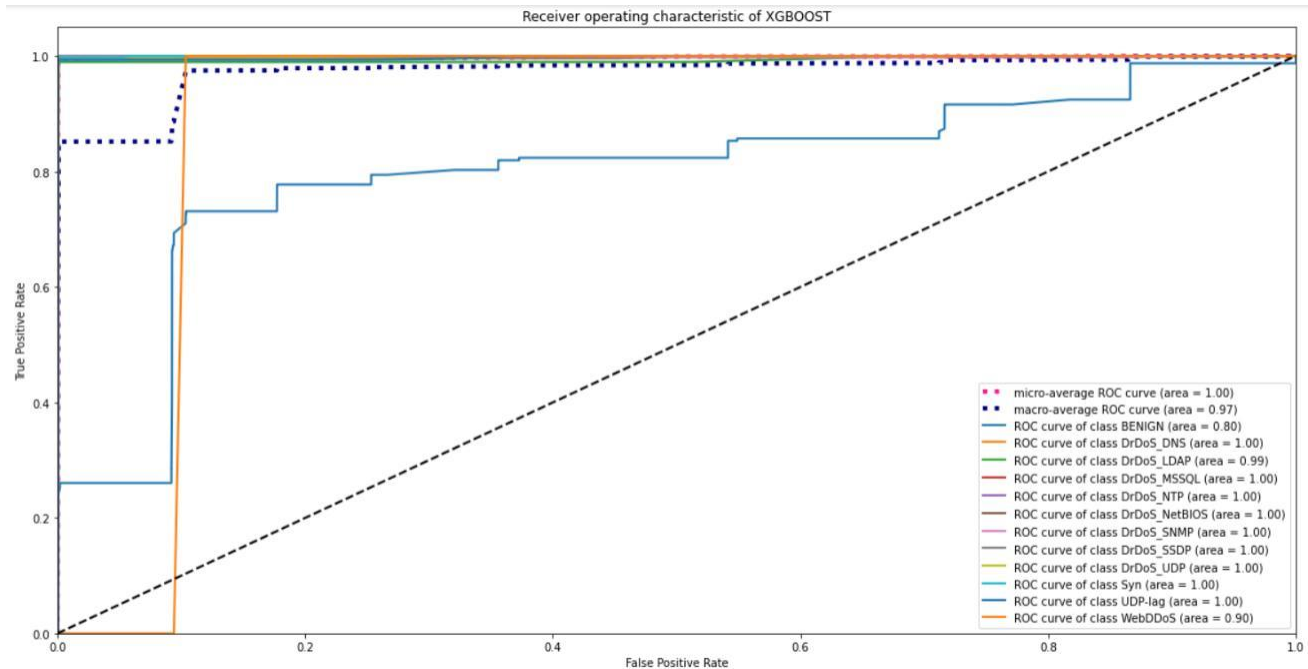
**Figure 8.5** RoC Curve for MultiLayer Perceptron

The **Long Term Short Term Memory** has an accuracy score of 95.58%, which is greater than SVM, Naive Bayes, MLP. F-1 Score for Benign is 0, so it is not detecting this threat and below in figure 8.6, we can say that the area under the curve for DNS, LDAP, MSSQL etc. is 1.



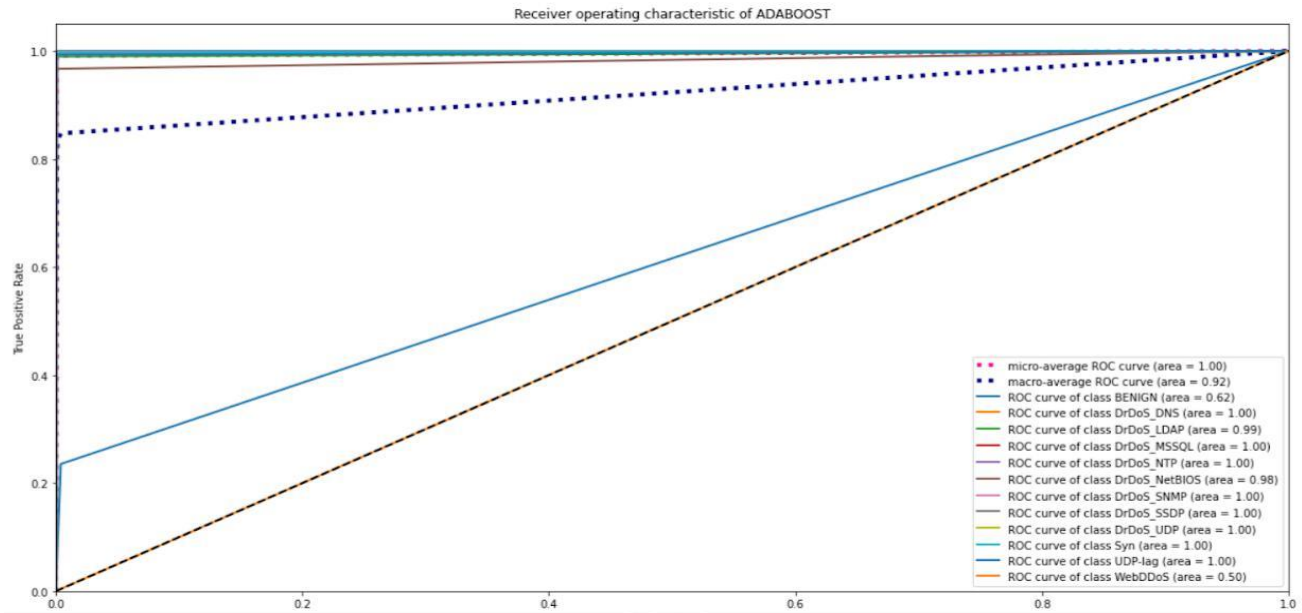
**Figure 8.6** RoC Curve for LSTM

The **XGBoost** method has the highest accuracy score of all the AI and ML systems in this study, at 99.51 percent. Table 8.2 shows the F1-Score for this method, which shows that it is not effective in detecting Benign threats. By the analysis of RoC Curve for XG Boost as Shown in Figure 8.7, it can be seen that area under the curve for all the attacks is higher than 1 and only for webDDoS it is 0.90. So XGBoost is best in predicting these attacks and it is the best model we have implemented. So it is very effective and detect all the type of attacks hence can be chosen for detecting.



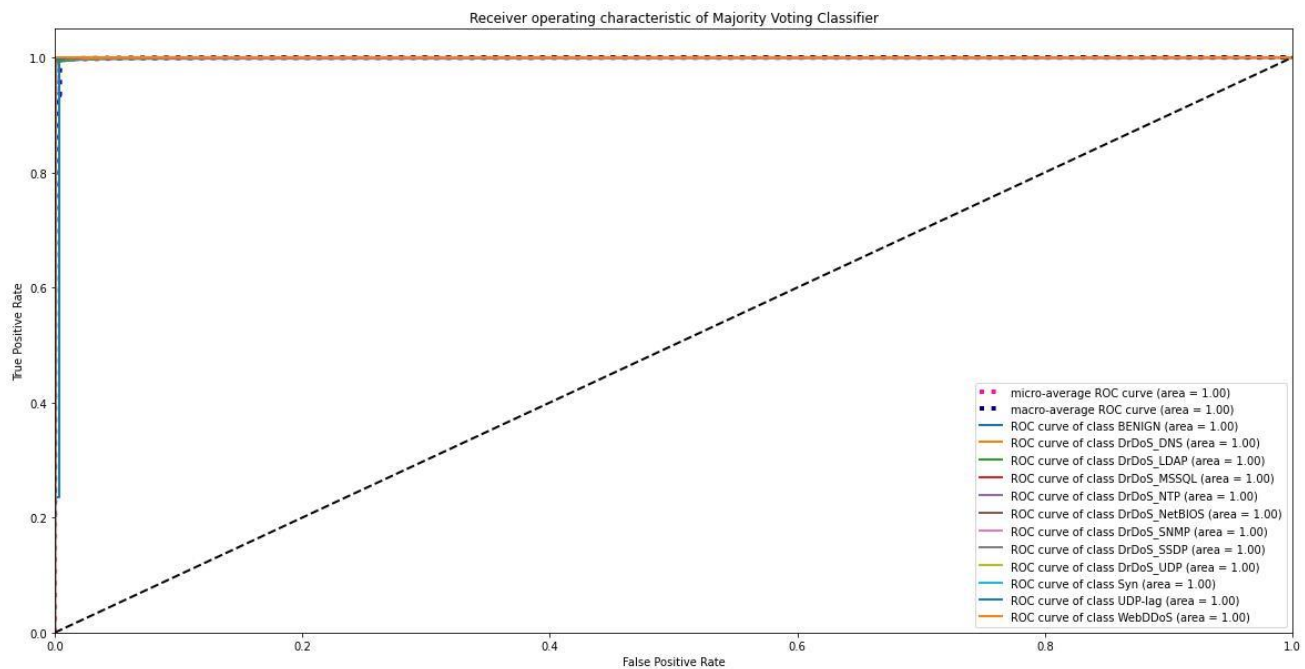
**Figure 8.7** RoC Curve for XGBOOST

The **Adaptive Boosting** technique is an Ensemble learning method that uses Decision Tree as the basic estimator in this study. This algorithm has a 99.11 percent accuracy score, which is greater than SVM, MLP, and almost near to Decision Tree. Table 8.2 also has the F1-Scores for all of the distinct threats. It has six threats that are perfectly identified, as evidenced by the F1-Score of 1.00: MSSQL,DNS,SNMP,SSDP, SYN, and UDP.And by the analysis of ROC curve for Adaboost as shown in Figure 8.8,it can be said that area under curve is higher for all the types of attacks but too low for some attacks.



**Figure 8.8** RoC Curve for ADABOOST

The **Majority Voting Classier** (MV-4) combines the performance of the Random Forest, Adaboost, Decision Tree, and XGBoost algorithms, yielding a 99.11 percent accuracy score for MV-4, which is similar to AdaBoost but slightly lower than XGBoost. The F1-Score in Table 8.2 clearly shows that the majority of the threats have an F1-Score of 1.00. Furthermore, it has an F1-Score of 0.00 for WebDDoS, indicating that it does not identify this form of attack. Figure 8.9 Shows that the RoC curve for MV-4, the area under the curve is 1 for all the threats.



**Figure 8.9** RoC Curve for Majority Voting Classifier

## CHAPTER - 9

### CONCLUSION AND FUTURE WORK

The Multiclass Classification For DDoS Cyber Threat was performed by using different machine learning and artificial intelligence algorithms and each of the threats was individually identified and validated by using different metrics.

We have referred to many different papers and contemplated over their drawbacks and limitations and tried to overcome many of them so an Ensemble Classifier MV-4 was presented for multiclass DDoS Cyber Threat detection which has an high accuracy score.

In addition a comprehensive study of different ML and AL and DL algorithms was performed for DDoS multiclass Cyber Threat Detection. We are using models like decision tree, SVM, Random Forest, Adaptive Boosting, Multi-layer perceptron, Long term short memory. We have studied different ML and AI and DL models and we are going to apply these models on datasets based on some metrics like accuracy, F1\_Score We are going to detect the best model that detected the cyber threat.

The major goal of this work is to successfully implement the detection of various types of DDoS threats. Our goal for future work is to use AI and ML algorithms to deploy multiple solutions for each of the attack types in order to defend the network from such attacks. This work will be expanded upon to create a system that can detect DDoS Cyberthreats and deploy countermeasures to prevent critical CyberSecurity dangers.

## REFERENCES/BIBLIOGRAPHY

- [1] Dr.A.Pasumpon pandian, Dr.S.Smays, “DDOS ATTACK DETECTION IN TELECOMMUNICATION NETWORK USING MACHINE LEARNING” Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2019). <https://www.irojournals.com/jucct/> DOI: <https://doi.org/10.36548/jucct.2019.1.004>.
- [2] Amardeep Chopra, Sunny Behal, Vishal Sharma, “Evaluating Machine Learning Algorithms to detect and classify DDoS attacks in IoT” Proceedings of the 15 th INDIACom; INDIACom-2021; IEEE Conference ID: 51348 2021 8 th International Conference on “Computing for Sustainable Global Development”, 17th - 19th March,2021 Bharati Vidyapeeth's Institute of Computer Applications and Management(BVICAM),New Delhi (INDIA).  
<https://www.researchgate.net/publication/349999101>. IEEE 2021.
- [3] Aween Abubakr Saeed , Noor Ghazi Mohammed Jameel, “Intelligent feature selection using particle swarm optimization algorithm with a decision tree for DDoS attack detection” International Journal of Advances in Intelligent Informatics Vol. 7, No. 1, March 2021, pp. 37-48. ISSN 2442-6571. <http://ijain.org> [ijain@uad.ac.id](mailto:ijain@uad.ac.id). DOI: <https://doi.org/10.26555/ijain.v7i1.553>.
- [4] Abdul Moqet, “A Machine Learning Based Classification Technique to Detect DDoS Attack in Cloud Computing Environment” Capital University of Science and Technology,2021.  
<https://thesis.cust.edu.pk/UploadedFiles/Abdul%20Moqet-MCS183056.pdf>.
- [5] K.R.W.V.Bandara, T.S.Abeysinghe, A.J.M.Hijaz, D.G.T.Darshana, H.Aneez, S.J.Kaluarachchi, K.V.D.L.Sulochana and Mr.DhishanDhammearatchi, “Preventing DDoS attack using Data mining Algorithms” Sri Lanka Institute of Information Technology Computing (Pvt) Ltd. International Journal of



---

Scientific and Research Publications, Volume 6, Issue 10, October 2016. ISSN 2250-3153.<https://www.academia.edu/download/50248537/ijserp-p5857.pdf>.

[6] Muhammad Aamir, Syed Mustafa Ali Zaidi, “Clustering based semi-supervised machine learning for DDoS attack classification” Journal of King Saud University - Computer and Information Sciences Available online 5 February 2019. DOI: <https://doi.org/10.1016/j.jksuci.2019.02.003>.

[7] Rejimol Robinson R R, Ciza Thomas, “Ranking of Machine learning Algorithms Based on the Performance in Classifying DDoS Attacks” 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10-12 December 2015 | Trivandrum, India. INSPEC Accession Number: 16072233 ,DOI: 10.1109/RAICS.2015.7488411.IEEE 2015.

[8] JESÚS ARTURO PÉREZ- DÍAZ, ISMAEL AMEZCUA VALDOVINOS, KIM-KWANG RAYMOND CHOO, AND DAKAI ZHU, “A Flexible SDN-based Architecture for Identifying and Mitigating Low-Rate DDoS Attacks using Machine Learning” The University of Texas at San Antonio and Tecnológico de Monterrey. INSPEC Accession Number : 19975410, DOI: 10.1109/ACCESS.2020.3019330 .IEEE 2020.

[9] Boyang Zhang, Tao Zhang. Zhijian Yu, “DDoS Detection and Prevention Based on Artificial Intelligence Techniques” 2017 3rd IEEE International Conference on Computer and Communications, Chengdu, China. INSPEC Accession Number: 17651878,DOI: 10.1109/CompComm.2017.8322748. IEEE 2017.

[10] 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” Hindawi Security and Communication Networks ,Volume 2019, Article ID 1574749,DOI: <https://doi.org/10.1155/2019/1574749> .

- 
- [11] Fahd A. Alhaidari, Ezaz Mohammed AL-Dahasi, “New Approach to Determine DDoS Attack Patterns on SCADA System Using Machine Learning”, 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia . INSPEC Accession Number: 18674218 , DOI:10.1109/ICCISci.2019.8716432 .IEEE 2019.
- [12] Soodeh Hosseini , Mehrdad Azizi , “The hybrid technique for DDoS detection with supervised learning algorithms” 2019, Kerman, Iran .DOI: <https://doi.org/10.1016/j.comnet.2019.04.027>.
- [13] Afsaneh Banitalebi Dehkordi, Mohammad Reza Soltanaghaei, Farsad Zamani Boroujeni, “The DDoS attacks detection through machine learning and statistical methods in SDN” 2020, Islamic Azad University, Isfahan, Iran. The Journal of Supercomputing.  
DOI: <https://doi.org/10.1007/s11227-020-03323-w>.
- [14] Aqeel Sahi, David Lai, Yan Li and Mohammed Diyykh, “An Efficient DDoS TCP Flood Attack Detection and Prevention System in a Cloud Environment” 2017, University of Southern Queensland, Toowoomba, QLD 4350, Australia . Electronic ISSN: 2169-3536, INSPEC Accession Number : 16870808, DOI:10.1109/ACCESS.2017.2688460.IEEE 2017.
- [15] Afsaneh Banitalebi Dehkordi, Mohammad Reza Soltanaghaei, Farsad Zamani Boroujeni, “A Hybrid Mechanism to Detect DDoS Attacks in Software Defined Networks” 2021, Iran. DOI: <https://doi.org/10.29252/mjee.15.1.1>.
- [16] Saman Taghavi Zargar; James Joshi; David Tipper, “A Survey of Defense Mechanisms Against Distributed Denial of Service (DDoS) Flooding Attacks” IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 15, NO. 4, FOURTH QUARTER 2013. INSPEC Accession Number: 13913972 , DOI: 10.1109/SURV.2013.031413.00127.IEEE 2013
-

[17] Petar Radanliev , David De Roure , Kevin Page, Max Van Kleeck, Omar Santos,La'Treall Maddox,Pete Burnap,Eirini Anthi,Carsten Maple, “Design of a dynamic and self-adapting system, supported with artificial intelligence, machine learning and real-time intelligence for predictive cyber risk analytics in extreme environments – cyber risk in the colonisation of Mars”, Published: 10 February 2021,DOI: <https://doi.org/10.1007/s42797-021-00025-1>.

[18] Noman Haider, Zeeshan Baig, Muhammad Imran, “Artificial Intelligence and Machine Learning in 5G Network Security: Opportunities, advantages, and future research trends”, Submitted on 9 Jul 2020,Riyadh, Saudi Arabia. arXiv:2007.04490 [cs.CR], (or arXiv:2007.04490v1 [cs.CR] for this version).

[19] S.Brahanyaa, L.Jani Anbarasi, “Classification of SNMP Network Dataset for DDoS attack prevention”, Published in: 2018 IEEE ICCIC Madurai, India. INSPEC Accession Number: 18882177 DOI: 10.1109/ICCIC.2018.8782319. IEEE2018.

[20] Amit Praseed and P. Santhi Thilagam, “DDoS Attacks at the Application Layer :Challenges and Research Perspectives for Safeguarding Web Applications IEEE Communications Surveys & Tutorials ( Volume: 21, Issue: 1, First Quarter 2019). INSPEC Accession Number: 18486080 DOI: 10.1109/COMST.2018.2870658.IEEE 2018.

## **Appendix A: Definitions, Acronyms and Abbreviations**

Ddos (Distributed denial of service): Distributed Denial of Service (DDoS) is a type of Cybersecurity threat which is one of many versions of Denial of Service(DoS) that uses IP addresses to attack a particular server/victim. Distributed Denial of Service (DDoS) are flooding threats that deny a legitimate user from accessing its intended service. It is one of the most prevalent threats that the Cybersecurity industry is facing as per the recent market research.

GPU (Graphic Processing Unit): GPU, a specialized processor originally designed to accelerate graphics rendering. GPUs can process many pieces of data simultaneously, making them useful for machine learning, video editing, and gaming applications.

DNS (Domain Name Server): Domain name server is a server responsible for keeping the file that contains information about the domain name(s) and corresponding IP addresses (zone file) as well as for providing the above-mentioned information during DNS queries. Domain name servers are a fundamental part of the Domain Name System.