9 novembre 2019

*People's Democratic Republic of Algeria*
*Ministry for Higher Education and Scientific Research*
*University of Ghardaia*
*Faculty of Science and Technology*
*Laboratory of Mathematics and Applied Sciences*
*Department of Mathematics and Computer Science*

# MASTER THESIS

*Presented to obtain the **Master diploma** in Computer Science*
*Specialty: Intelligent Systems for Knowledge Extraction*

# THEME

# *Machine Learning-based DDoS Attacks mitigation*

***Presented by:***

BEKKOUCHE Noussaiba

**Jury members:**

| | | | |
|---|---|---|---|
| M. OULAD NAOUI Slimane | MCB | Univ.Ghardaia | President |
| BETKA Messoud | MAA | Univ Ghardaia | Examiner |
| KECHIDA Khaled | MAA | Univ Ghardaia | Examiner |
| KERRACH Chaker Abdelaziz | MCB | Univ Ghardaia | Supervisor |

***University Year: 2018/2019***

# Abstract

Distributed Denial of Service (DDoS) attacks appears in a great significant on the network level, and since the most common users of the network frequently are the institutions and centers to communicate with the other side who is the user, hence that they vulnerable to this attack increasing the number of users and the occurrence phenomenon of the flood that affected the server. This leads to disabling and lack of communication between the users in the network. This causes of malfunction divided into two parts : the first is the normal user, and the second is the attacker It leads to numerous losses, both economic, temporal and Physical. Our study is about the attack that is the biggest cause of this problem, we applied this study to the data site of the University of Ghardaia where it launched an attack type HTTP flood which stopped for a period of time until been relieved. We have proposed a modern technique helps to study the attack and detection it using algorithms is Kmeans algorithms, the result obtained showed with approximate 70% of the cluster that contains the amount of use of the methods of requests, which deduced 34% of the error rate on the dataset.

## Keywords :

DDoS, Attacks Mitigation, Machine Learning.

# الملخص

تظهر هجمات "رفض الخدمة الموزعة (DDoS) "بشكل كبير على مستوى الشبكة ، وبما أن أكثر مستخدمي الشبكة شيوعًا هم المؤسسات والمراكز التي تتواصل مع الجانب الآخر وهو المستخدم ، وبالتالي فهي عرضة لهذا الهجوم زيادة عدد المستخدمين وظاهرة حدوث الفيضان التي أثرت على الخادم. هذا يؤدي إلى تعطيل وعدم وجود اتصال بين المستخدمين في الشبكة. تنقسم أسباب الخلل هذه إلى قسمين: الأول هو المستخدم العادي ، والثاني هو المهاجم وهو يؤدي إلى خسائر عديدة ، اقتصادية ووقتية ومادية. تدور دراستنا حول الهجوم الذي يعد أكبر سبب لهذه المشكلة ، قمنا بتطبيق هذه الدراسة على موقع بيانات جامعة غرداية حيث أطلقت فيضان HTTP من نوع الهجوم الذي توقف لفترة من الوقت حتى تم التخفيف منه. لقد اقترحنا تقنية حديثة تساعد في دراسة الهجوم واكتشافه باستخدام خوارزميات هي خوارزميات Kmeans ، والنتيجة التي تم الحصول عليها أظهرت بنسبة تقريبية 70٪ من الكتلة التي تحتوي على مقدار استخدام طرق الطلبات ، والتي استنتجت 34 ٪ من معدل الخطأ على مجموعة البيانات.

# Contents

# Table of figures

# List of Tables

# List of Algorithms

# DEDICATION

*To the special man who spent his life in the cause of my happiness and made of his eyes a lamp that illuminates my path , my father "Kaddour "*

*To the special woman who care about me and give me everything I asked about, I ask Allah to give her the happiness and the long life my mother "Fatima "*

*To the man who gives me everything I need patience and advice, I ask God to give him a long life, my husband "Khireddine "*

*To My dear sisters: Soumaia ,Bouthaina Roufaida and my dear brother Taha*

*To my dear friends*

# *TANKS*

## *I thank*

*A LLAH who gives us the help to arrive to this day*
*Our parents for their support us throughout our study*

*My supervisor KERRACH Chaker Abdelaziz*
*followed this work*
*My teacher BELAOUAR Slimane for his helps and advices*

*My jurors who have followed this work*
*Special thanks to my teacher OULED NAOUI Slimane*

*All my teachers from the primary school to the university*
*Who helped us in this work even those who prayer for us*

# General Introduction

## Contexte

The objectives of the server is to meet the needs of the users whether in demanding, storing or in extracting the data, which should be very precise and accurate, the server uses the network to meet the requests with one of these means which are the websites and the servers around the network. In recent years, with the rapid evolution of information technologies and provide protection against attacks on the network, it has emerged many techniques to get to the information.

Denial of Service (DoS) is now a prominent issue because they use multiple protocols as a means to launch their attack ; Where specialists dealt with this type of attack in a final and has not appeared since then, With the insistence of attackers to continue to find other ways. Where it was created new method in our time is the DDoS attack, Protocols that provide scope for this attack are wider than DoS. One such protocol Transmission Control Protocol (TCP), User Datagram Protocol (UDP), Hypertext Transfer Protocol (HTTP)[20].

DDoS attack occurs a lot in the institutions and big centres that contains a big amount of data, which rely on help factors such as Botnet to launch its attack on the victim.

## Problem and Objectives

To develop ways to protect the information on the network and find solutions to close down such available hackers gaps, the attacker found another ways to launch their attack with network protocols but with a different advantage over other known attacks, forces the server to stop and not continue its work on the network causing great losses, this is called a DDoS attacks.

Hence, we proposed to solve the problem (Chapter 3) a detection and Attack mitigation it ; but did not arrive them to the final stage of mitigation by Machine Learning.

The main objective of this work is to detect the reason of the attack, and the system or server must benefits the machine learning suggestion to expose and avoid the attack, helping it find ways to protect it.

## Thesis plan

To discuss on subject, organized as follows :

1. Chapter one represent Explain two concepts DoS and DDoS With a method used in the network and how to treat them in various ways.
2. Chapter two represent Machine Learning, algorithms that which allow for massive data processing.
3. Using a statistical method to detect and attack mitigate, this is in Chapter three.

Chapitre 1

# DoS And DDoS

## 1.1 Introduction :

In the field of networks there are many attacks that are exposed to it every day, but this attack is somewhat different for them. Of the nature of the attacks that take place in a network that disrupts the addresses or systems in the computer or equipment of the devices, or reduce the effectiveness of the previous and not give a result of exact or incorrect.

DDoS and DoS attacks have become very widespread for ease of implementation and applied to networks. They cause a significant loss of data due to service interruption directly or indirectly, from there which we expect this danger Based on network level. Take several security measures and one of the biggest security concerns is its inability to address any type denial of service attack.

It is important before entering into a solid subject note that there is a subtle difference between the meaning of DoS and the DoS attack, it is certain there are many ways to do DoS, and it is essential to be classified as an DoS attack.

Although these attacks have been more prevalent in the past but have not been as important before they hit major economically major to such Yahoo !, Amazon.com and CNN[53].

This chapter provides attacks, their motivations, an overview of Denial of service attacks (DoS), Distributed Denial of Service Attacks (DDoS), Reasons for Denial-of-Service, then we explain types of attacks and also present a brief history of DoS and DDoS And some of it is methods used and then we

explain Distributed Denial of Service Attacks and some of it's used methods, eventually some metigation techniques for solution in the network .

## 1.2 Motivated attacks

Computers and websites attack is something that is certain and likely in companies and administrative institutions . Exploit an error in the computer system (operating system, programs, etc.) For reasons that are not pre-defined This type of attack usually occurs against devices, servers, and access to the company so that customers can not access the information [58]. The purpose of attack is not to change data, delete it, or even steal any information. This is damaging the reputation of online businesses by preventing the running of their activities some motivation :

- Ideology : called "hacktivists" People with different ideas.
- business feuds : such as Cyber Monday, use attacks to strategically take down competitor websites.

There is a lot (boredom, cyber warfare (Sites opposed to the government or a political matter).

## 1.3 Definition

### 1.3.1 Denial of service attacks (DoS)

Denial of service (DoS) not a difficult attack to defend it, but it is more effective than others, it can attack any kind of networks, sites, servers and routers . Its effective rate is not less than the 99.7%[28] which affects the level of security gaps with IP and TCP. We have two categories of offensive effects :

- Denial of service by saturation of server from the requests so, that it can not be able to respond.
- Through the security holes that make the server or site unable to receive and send and be invalid to the user.

DoS is a mono attack[28] that takes place between an attacker and a server, and is the first appearance of the attack in the network ; its characteristics are clear and way to addressing them simple and easy, it prevent legitimate users from accessing the desired sites. There are two groups in the DoS attacks.

1. Attack on network level.
2. Host Based Attack.

Network-level attacks disable legitimate users access by attracting power at the network, while Host Based-level attacks disrupt the service by terminating server power. DoS is considered one of the most current network attacks For example email, DNS or HTTP servers and need stages or methods for defense[28] .

The principle of DoS attacks is to send a set of data in a random manner, in order to cause saturation and thereby prevent them from providing network services and communicating with users. Which causes this kind of attack to disrupt the target equipment. Therefore, denial of service (DoS) is classified from attacks  Possibility to control it in enterprises or major sites, which can cause serious problems in the event of such a situation for only a few hours(Figur (1.1)).

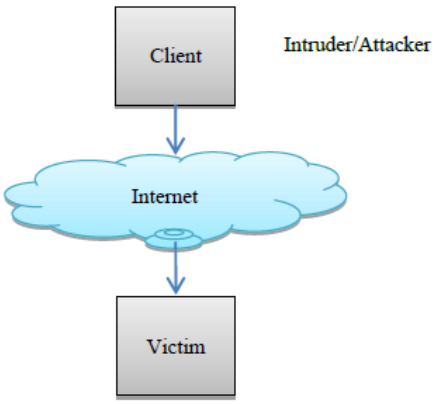



FIGURE 1.1 – Generic DoS model[65].

### Brief History About DoS Attacks

Since the first DoS attack was launched in 1974[36], by student David Dennis, a 13-year-old (Radware Ltd 2018)[11], where he experimented with PLATO devices  Special By CERL, Which makes external conversions. He wanted to discover what would happen if sent "ext" to many PLATO devices at the same time, have been the first instance of a denial of service (DoS) attack on a network computer[36].

In November 3, 1988, launch Robert Morris Jr an attack. leading to the disruption of all institutions across united States of America[50]. In 1988 was first detection happened, where researchers have detected and controlled it[50].
In 1990s, traded Online chat (IRC) Widely Where many users wanted controlling her , Where ordinary users are forced to withdraw Of this program and remains a little controlled by the users  Hence, used denial of service at that time[36] and Then came Botnet.

To prevent an attack leading to the interruption and non-communication with the external user and server of its conditions the time and quantity sent. As we explained earlier with DoS the same thing with DDoS ; but to circumvent the attacker with the server it follows other ways In order to disable it, thus, we will explain what DDoS and methods of attack.

### 1.3.2 Distributed Denial of Service Attacks

The distributed denial of service attacks were called server disruptions because of the concurrent transmission of the capacitor to it simultaneously. It is more complex than DoS attacks because it happens to users who are not harmful by participating in this disruption. Thus, it can not distinguish between ordinary users and attackers (Figur1.2).

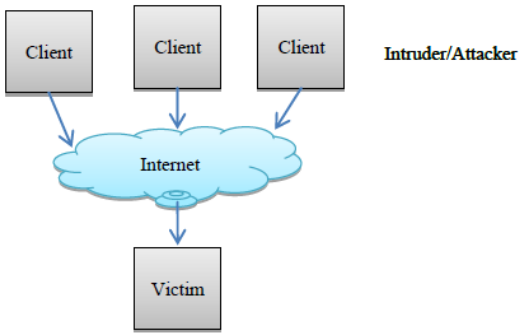



FIGURE 1.2 – Generic DDoS model[65].

The duration of DoS attacks may persist for a long time or for a short time for a target, as in DDoS, but there are changes in the structure of the attack, which is a DoS development. DDoS is a multi-server DoS attack that can control these servers, which are a group of assistants called botnets. It is the means that serves the attacker on victim, Lead to disable it.

The famous example is Amazon[4]. The Mstream tool uses tricking method for attacking the target host, can marsh the information used by routing methods(Figur(1.3)).

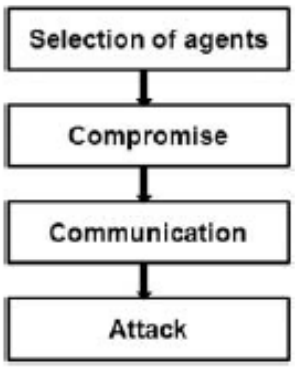



FIGURE 1.3 – Steps to perform a DDoS attack[10].

**Example to illustrate : network communication**

The center distributes the network to all the city's population. If the population increases at the local level, there is an inflation in the telecommunications service resulting in a network malfunction causing temporary disruption until it is repaired[58].

To attack DDoS there are three types it is volume DDoS attacks, application layer DDoS attacks and state-exhausting(protocole)[28].
volumetric DDoS attaks or (flooding attack ) in most widely used about 65% [28] of other attacks, including attacks User Datagram Protocol (UDP), is one of the members of the Internet Protocol, and also Internet Control Message Protocol (ICMP) is protocol helps all existing protocols in communicate among themselves[66].

In most cases, the sender and recipient must agree on the amount of messages or information to be processed with time and commitment to it. Therefore, if these conditions are not facilitated, failure occurs at the recipient level (victim).

**Brief history About attacks DDoS**

DDoS attacks are an extension of DoS attacks, which led to a change in the idea of the type of attacks and their seriousness in the field cyber security[36].

September 1996 : Carnegie Mellon's Computer Emergency Response Team (CERT) published it's first bulletin on SYN flooding[75].

October 1997 : They gave depth about denial of service to show that this attack is as a priority for network administrators[75].

August 1999 : In DDoS attack tool called trinoo was deployed in at least 227 systems, of which at least 114 were on Internet2, to flood a single University of Minnesota computer ; this system was knocked off the air for more than two days.

February 2000 : Yahoo, Buy.com, eBay, CNN, Amazon.com, ZDNet.com, was the victims of a DDoS, the three hours Yahoo was down, it suffered a loss of e-commerce Significantly.

2009 : Arbor Networks reported that the size of the largest reported DDoS attacks had increased steadily, from 400 megabits (Mbps) per second in 2002 to 49 gigabits per second (Gbps)[66].

Feb 2010 : Launch an attack on the website of the Australian Parliament (www.aph.gov.au), Also in the same year attack Botnat was a web serve It was not different sources[66].

11 February 2014 : A major attack on the level of US headquarters servers The attack reached 400 Gbps, Also in the same year Massive distributed denial-of-service (DDoS) attack The DNS has been disabled in all servers of the company an online gaming[66].

### 1.3.3 Goals of Denial-of-Service and Distributed Denial of Service Attacks

**1.The bankruptcy of electronic economic companies (EEC)**

EEC is one of the most profitable institutions in the world, so companies compete to win the largest stocks in different economies. Competitors may take illegal ways to break each other, and one of these methods is to disable them by Denial of service.

**2.Compressing messages**

When the sender sends data to the receiver continuously without interruption or waiting for response, this causes him to be pressured out, either it is an spam or the size of a store in a server is full and can not receive again, this it is called a process denial of service.

**3.Experimental**

This is the biggest frequently traded cause and is a catalyst to launch DDoS attack, at the beginning of the attack experience of many programmers or hackers is the discovery of attack methods to get experience or pleasure only and applied to small centers. This type of attack can be detected and defended from it[33].

### 1.3.4 Types of DoS and DDoS Attacks

DoS attacks are limited between attacker and the victim and the process of amplification occurring at the network level have led to its disabling, to launch a DoS attack used its protocols including TCP. The principle of the methods of evolution of the attack in DoS, did not lost it the attacks of DDoS, but many types were created as needed and as they exploited the protocols of the Security gap which enables them to control the victim temporarily and it is often permanent[2](Figure(1.4)).

1. volume DDoS attacks (Flooding attacks)
2. Protocol based attacks (state-exhausting DDoS attacks)
3. Application layer based attack

In general cases of denial of service attacks take two forms, They either flood by services or crash them. The most used are flooding.

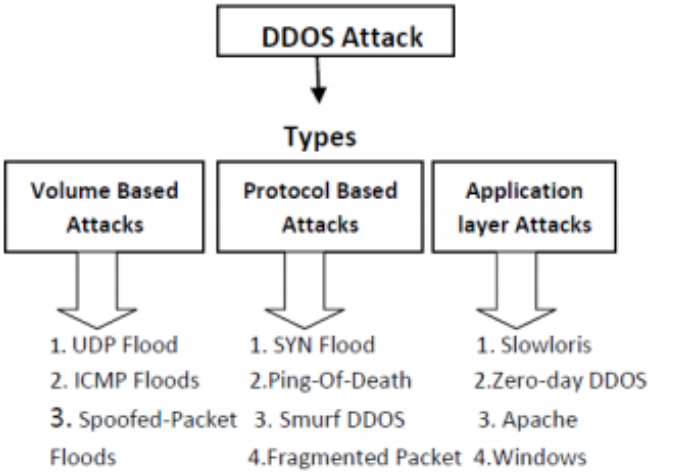



FIGURE 1.4 – Distributed DoS attack classification[42].

**1.Flooding layer attacks :** Floods are the most frequently used and used attack on DoS. This occurs when the attacker system is overloaded with traffic that the server, sends attacker an overwhelming number of messages at system can not handle, where eventually stops.

1. ICMP Flood :

   Internet Control Message Protocol Flood ICMP[66] is a flooding attack Used at several types of attack including Ping of Death, Ping Floods, ICMP DoS Attack.

   Before we explain ICMP Flood We are reminded of a protocol ICMP (Internet Control Message Protocol) is a connection protocol used for diagnostic purposes for managing error information to connected machines. or querying any server, Used to know the two parties in contact with each other. where This This is done by sending a packet attached to the ICMP echo request to the receiver of the two-way communication. [66]

2. UDP Flood :

One of the DoS attacks is User Datagram Protocol (UDP). It is generally similar to TCP protocol, and differs in some characteristics such as it does not require the consent of the receiving party. Is the simplest communication protocol for the transport layer (Layer-4) of the OSI model, [39] When a contact occurs with the victim, the victim is sent "ICMP Destination Unreachable"[39] to attacker.

**2.Protocol layer attacks**

1. SYN Flood :

A SYN flood is a form of denial-of-service attack, in which an attacker sends of SYN requests a victim server [7], it is SYN packets at very high packet rates that can overwhelm the victim by consuming its resources to process these incoming packets. if a server is protected by a firewall, the firewall will become a victim of the SYN flood . starts with negotiation of session parameters, by SYN or SYN/ACK, between the client and the server. Technique for attack is TCP, The basic vulnerability that allows SYN Flooding attack depends on the design and implementation of TCP,precisely in the three-way handshaking which represents the connection setup part of TCP. A SYN-flood DDoS attack takes advantage of the TCP (Transmission Control Protocol) three-way handshake process by flooding multiple TCP ports on the target system with SYN (synchronize) messages to initiate a connection between the source system and the target system.

**TCP SYN Flood :** Before we make clear the process TCP SYN Flood Explain how to connect of regular communication his name TCP three-way handshaking. Typically, when a customer begins a TCP connection with a server, the customer and server trade a progression of messages which regularly runs this way :[16]

(a) The customer asks for a connection by sending a SYN (synchronize) message to the server.

(b) the server will allocates a buffer for the client and replies with SYN and ACK packet.[16]

(c) Client responds with an ACK (acknowledge) message, and the connection is established.

This is known as the TCP three-way handshake, and is the establishment for each connection set up utilizing the TCP protocol(Figure(1.5)).

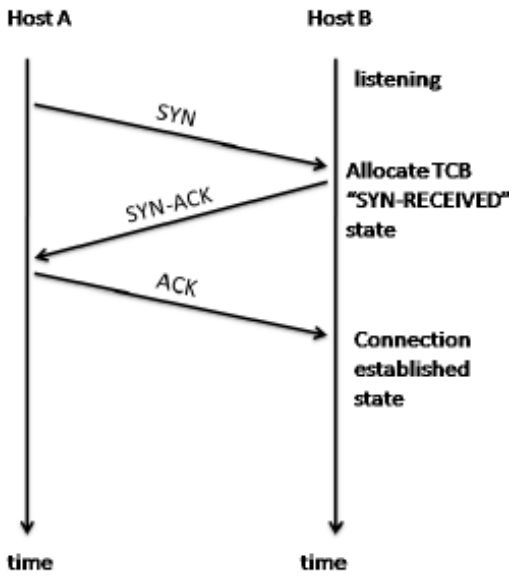



FIGURE 1.5 – TCP three-way handshaking [7].

Allocating resources for the received SYN is the goal of the TCP SYN flooding attack, so the attack aims to exhaust the memory space of the victim for the longest possible time by sending a ood of fake SYN packets, to make the system unresponsive to legitimate traffic .Exploited the The gap by the intruders without being realized by the server, that allows SYN Flooding attack. This attack happensin steps are as below :[16]

(a) The customer asks for a connection by sending a SYN (synchronize) message to the server.

(b) the server will allocates a buffer for the client and replies with SYN and ACK packet. and wait for confirmation or timeout expiration of SYN packets. and leaves an open port ready to receive the response.

(c) If the client does not send back the final ACK packet, which never arrives, Can Easily take exploited of this opportunity and start attacking.

(d) At this time the SYN flood attack and Sends a large package , so that it is full and can not be processed and once all the available ports have been utilized the server is unable to function normally.

there are three types of SYN flood attacks, which are going out in the nowadays Internet networks : Direct Attack, Spoofing Attack and Distributed Direct Attack(Figure(1.6)).

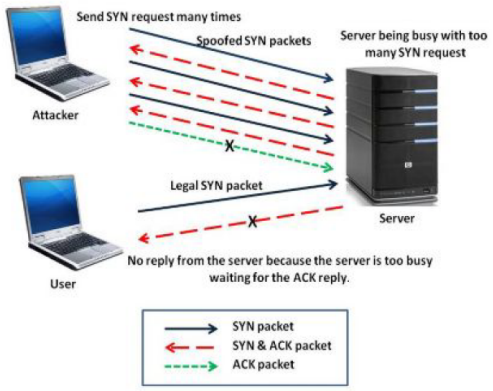



FIGURE 1.6 – TCP SYN Flood[26].

**A SYN flood can occur in three different ways :**

(a) **Direct attack :**

A SYN flood where uses own IP for the attack without spoong we call this a direct attack, the attacker does not mask their IP address at all. As a result of the attacker using a single source device with a real IP address to create the attack, This method of attack is very easy to perform because it does not spoofing into packets [16].

This is often achieved by firewall rules that stop outgoing packets or SYN packets or by filtering out any incoming SYN-ACK packets. first , where the attacker sends a repeated package SYN to the victim for calling the system many times [7]. Second, attackers must prevent their operating system from responding to the SYN-ACKs This is done using firewall because any ACKs, RSTs, or Internet Control Message Protocol (ICMP) messages will allow the listener to move TCB (Transmission Control Block) which maintains information about the local and remote socket numbers, the send and receive buffers, security and priority values, so that any SYN-ACKs are discarded before reaching it. In practice, this method is used rarely just block the IP address of each malicious system [16](Figure(1.7)).

(b) **Spoofed Attack :**

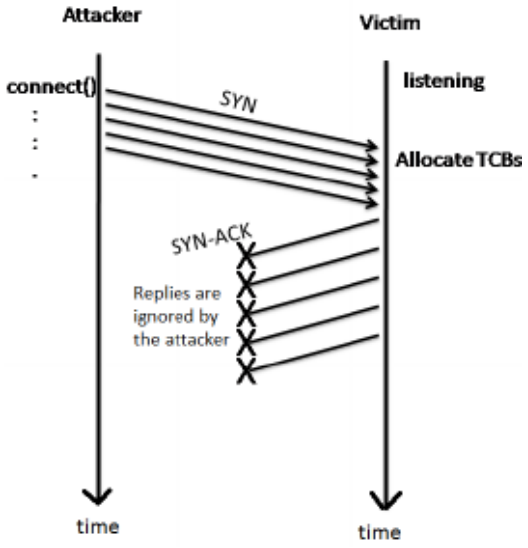



FIGURE 1.7 – Direct attack[7].

Another form of SYN flooding attacks uses IP address spoofing, it call this a spoofing attack.make their identity more difficult to discove, which might be considered more complex than the method used in a direct attack . IP spoong can be done by changing the headersof the packets and in jecting new ones with spoofed IP(s)[7]. For spoofing attacks, a primary consideration is address selection. If the attack is to succeed, Since the attacker spoofs the IP address, then the victim be sent back to the spoofed IP address.

must not respond to the SYN-ACKs that are sent to them in any way. because if that happened, and there were responses from the spoofed IP, then the victim will receive error messages andfree it's resources. the IP address spoofing techniques can be categorized into different types according to what spoofed source addresses are used in the attacking packets (Figure(1.8))

(c) **Distributed attack (DDoS) :**

In order created attack used a botnet there are of thousands of compromised machines that are used by criminals for DoS attacks. is the most dangerous amongst mentioned types of SYN flooding attacks. This attack is known as the Distributed Denial of Service attack (DDoS). In the case, the botnet use direct attacks, but in

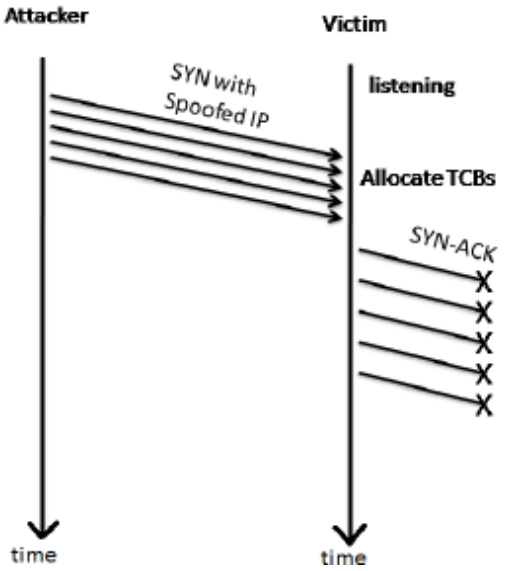



FIGURE 1.8 – Spoofed Attack[7].

order to increase the effectiveness even further[7], an attacker may have each distributed device also spoof the IP addresses from which it, controlled and sends packets through. (Figure(1.9)).

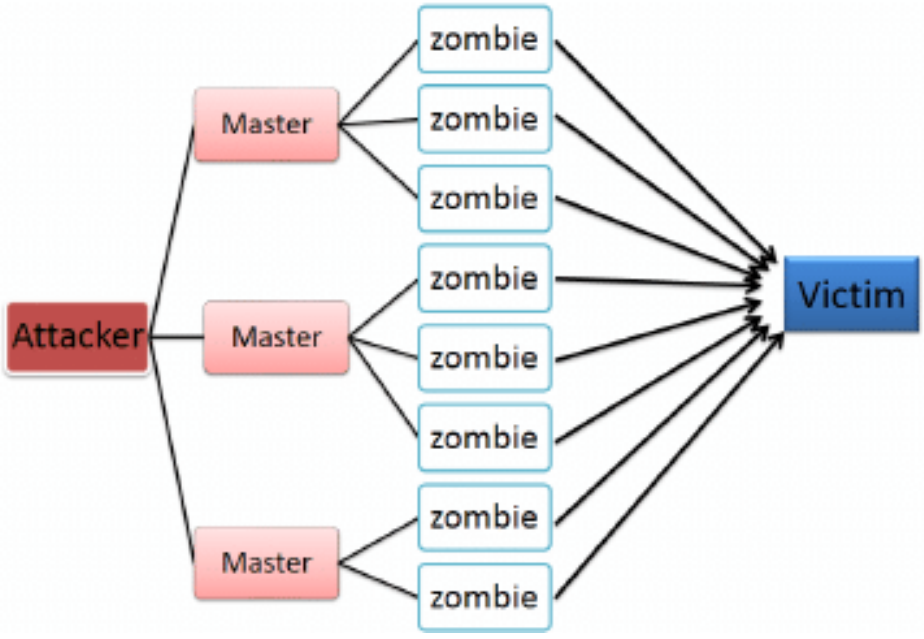



FIGURE 1.9 – Distributed attack (DDoS)[7]

2. Smurf Attack

   A smurf attack is a Distributed Denial of Service (DDoS) network based attack, in which an attacker attempts to flood a targeted server with Internet Control Message Protocol (ICMP) packets. by exploiting vulnerabilities of the Internet Protocol (IP) the malware creates a network packet attached to a false IP address a technique known as "spoofing." and Every host on the intermediary network replies, intermediary site is to be identified by attacker which helps in amplifying attack. flooding the victim and the intermediary network with network traffic. the Smurf attack can quickly cause a complete denial of service[42].

3. Ping of Death

   A Ping of Death attack is a denial-of-service (DoS) attack, A related attack known as an ICMP flood attack is more prevalent. Its attacks are often at the site level is sent out and the “echo”, if the connection is working, the source machine receives a reply from the targeted machine, and his attack begins [17].

**3.Application layer attack :**

1. HTTP Flood :

   Servers are exposed to HTTP (Hypertext Transfer Porotocol) for an extended denial of service attack on the network, where the application package sends the application layer HTTP DDoS [34]when HTTP requests start in large numbers. The process of the attack is that the HTTP protocol connects to protocols TCP and UDP to attack for to be sure to receive a packet requests from a single source or from multiple sources by means of botnet that, it may be an offensive or random attack.

   And the server will not be able to meet the needs, and is a malfunction and can not do any activity and this is what is called the distribution denial of service[34].

   In order to carry out this attack, HTTP uses many methods, some of them it places a trace on the server and some of them have a temporary

effect. We will provide some of them in the most important methods used.

**GET/ POST Flood** This attack occurs when an attacker initiates a vast number of requests in one session. The GET method sends the encoded user information appended to the page request. The POST method transfers information via HTTP headers[29].

2. Slowloris attack : is a highly-targeted attack, works by sending a large amount of simultaneous HTTP requests, be it GET or POST, to a server. It accomplishes this by creating connections to the target server, The targeted server keeps each of these false connections open. This eventually Exceeds the volume of communications required, and leads to denial of additional connections from legitimate clients.[34]
3. Zero-day DDoS : This type depends on the detection of a flaw or defect in a program that enables the collapse of the target completely, even if there is protection or firewall regulating communications, this type does not need a distributed network or botnet network enough one person exploit the gap and the server will collapse.

## 1.4 DoS attacks Methodes

**Bandwidth Attacks :**

**SYN Flooding Attacks** The attacker sends a packet to IP address. When normal attackers of the first class use a Ping attack to falsify addresses. In this experiment, our goal was to show the results of an attack on a device if it is connected to the network, we are sending 65500 packs (-l 65500) via TCP protocol to address (10.100.48.247) :(Figure(1.10))


Figure 1.10 – Attacker Information

To detect this attack, open Task Manager, notice the beginning of the attack on the receiver device level, the curve height increases when sending a large packet(Figure(1.12)).

Stages of the attack with an increase in the number of packets and an increase in the curve.

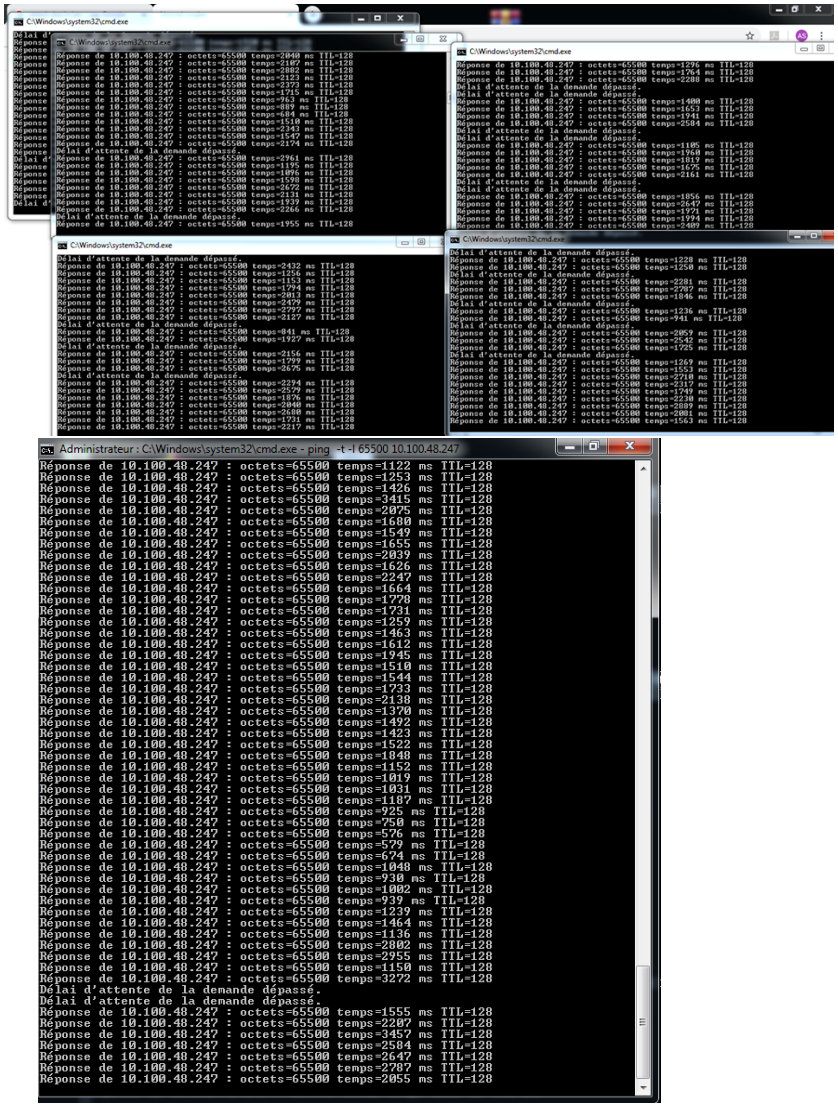



FIGURE 1.11 – In case of attack

## 1.5 Some vectors of attack

**Botnets :**

The variety of attacks that occur on sites often direct, they mean destruction, and there are indirect attacks. All this happens through gaps on the server by agents or assistants called botnet[52].

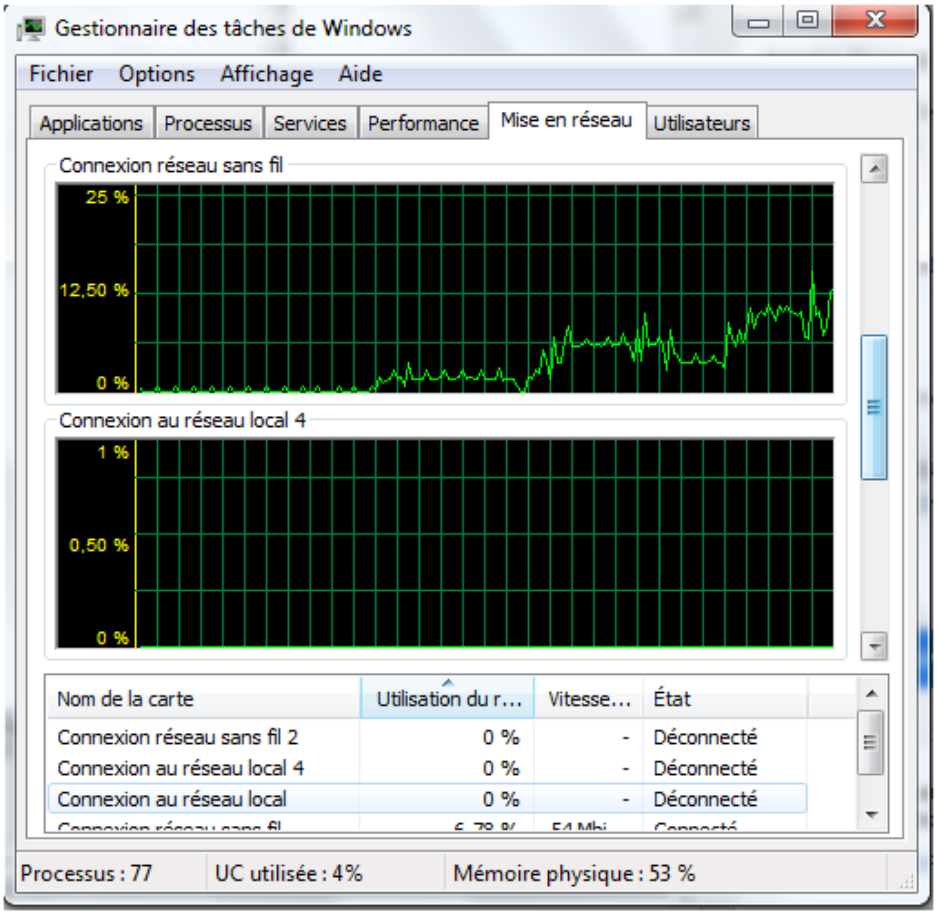



FIGURE 1.12 – Traffic Attack IP/ TCP

The idea about botnet established botnets network by writer developing a program, called a bot or agent, of Botnets network is that instead of attack one person it replaces several people or devices through Cloud Computing (Internet-connected ) and installing the program on compromised computers on the internet using various techniques. This technique is difficult and complex to dismantle. This distinction is the main reason make a difference in the two definitions in DoS and DDoS.

1. **Botnet Communication**

Internet Relay Chat (IRC) model[52] is protocol was initially developed for Internet chat applications, it is mainly designed for group (many-to-many) communication in discussion forums called channels,

but Attackers exploited it and use to communicate between bot malware[69].

Once bot malware is recruited on the victim machines the botmaster has to discover these bot malware infected machines. The botmaster needs to control these victim machines, so IRC is the mean of communication Once installed in the compromised computers, the bot will automatically join a specific IRC channel on an IRC server.

Instead several organized command languages and control protocols called botnet Command and Control (C& C) techniques. [69] However, IRC channels are not the optimal solution for attackers to communicate with the Botnet, with increases in volume of botnets happening inflation occurs at the level of IRC channels and creates a single-point-of-failure.

2. **Botnet Function**

One of the advantages of DDoS is the use of botnets[Honeynet 2005] , each type of botnet software contains a set of flooding mechanisms, such as SYN flood, ICMP flood, and HTTP flood.[52] Requires provision of orders or instructions to control the attack parameters such as sending rate and packet size. For to fix errors botnet, must add new functions into the botnet software. "For example, an attacker can instruct all bots to download a new type of flooding mechanism to defeat a DDoS protection system. Hence, the botnet owner has the capability to design a specific attack for a particular target, and maximize the similarity between attack traffic and legitimate traffic" [73].

## 1.6 DDoS attacks Methodes

Normal traffic at the network of the University of Ghardaia, There is no flood of requests and the network and TCP connections is in a stable state. We note the addresses that are connected at that moment (Figure(1.13)).

**HTTP Flood Attacks :**

**GET :** An unusual flow of communication with the network DDoS at the University of Ghardaia (Figure(1.14)).

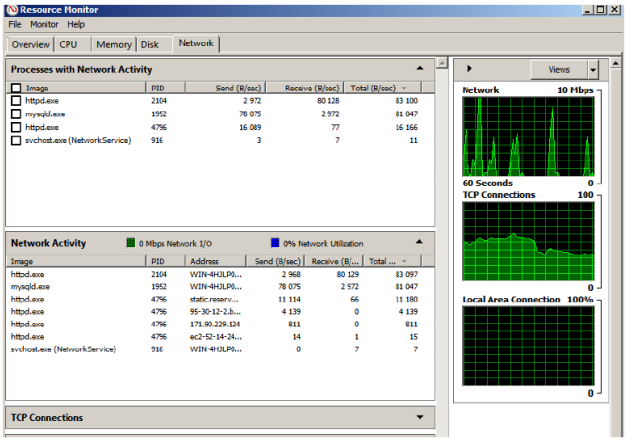



FIGURE 1.13 – Traffic normal at the network of the University Ghardaia(1)

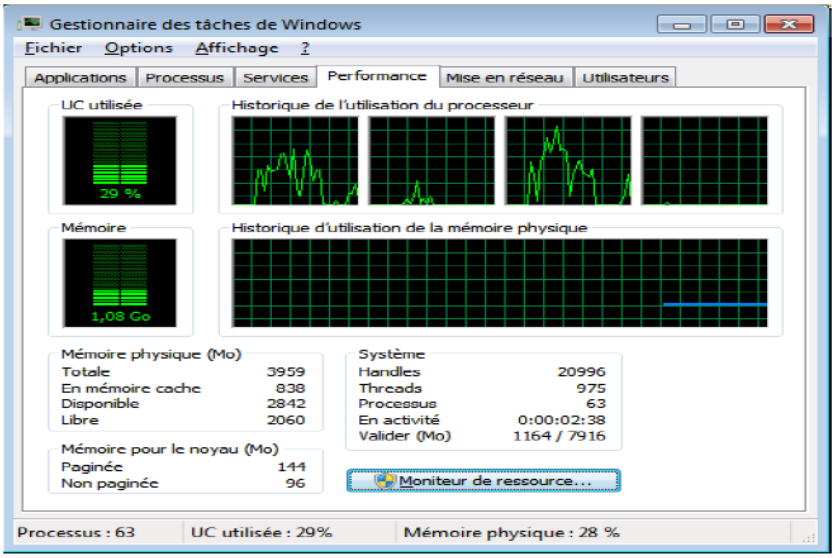



FIGURE 1.14 – Detect an attack DDoS

The first curve shows the flow volume, and the second curve shows the number of connections(Figure(1.15)).

## 1.7 DDoS Attacks on the Internet of Things

The Internet of things (IoT) is Branching and expanding in Internet uses into physical devices and everyday objects. The definition of the Internet of things has evolved due to convergence of multiple technologies, machine learning, and embedded systems. They can be remotely monitored and control-

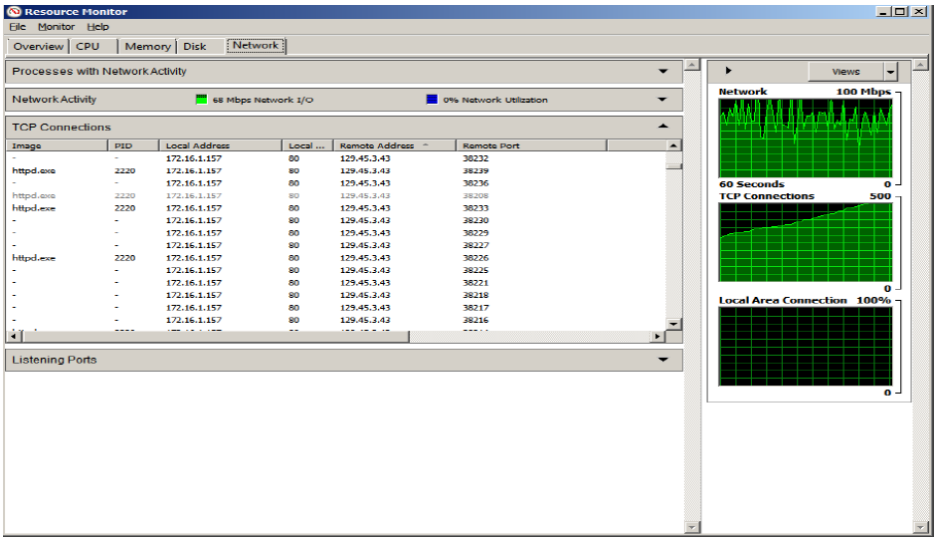



FIGURE 1.15 – The results of the HTTP flood attack

led, IoT technology is most common in the side of medicine [54], has faced world especially in regards to privacy and security concerns related to these devices and their intention of pervasive presence. Internet of Thing (IoT) devices were used to generate a DDoS attack[54].

### 1.7.1 DDoS attack tools

**TFN (tribe flood network) :** It is botnet architecture has Communication with protocol ICMP, he directs his attack on the protocols :IP / TCP / UDP / ICMP[8]. TFN agents(botnet) under the control of the TFN client, carry out the attack distributed on the victim / target and with the type of attack desired by the customer. The addresses of the client and agents are spoofed (fraud), Until an DDoS attack is created that has the capability to deplete both resource and bandwidth of the goal[23].

**Trinoo or Trin00 :** Is a set of computer programs to conduct a DDoS attack, that uses a master host and several broadcast hosts, for compiles a list of machines that can be compromised.[23] Where he communicates with them to compromise them and convert them into the Trinoo Masters that each load a subnet of agents.He directs his attack on the protocols : UDP[33].

**Mstream :** It involves a series of "botnet," programs planted on compromised systems, has ability to forge the source addresses."It creates the TCP

ACK flood and TCP RST flood requests to the target server" [8] mstream uses both TCP and UDP in its functionality and does not exploit a particular service.

### 1.7.2 DDoS Defenses in the internet

#### The techniques for solved problems DDoS by mitigation

It is important to note that some companies legitimately offer services of DoS/DDoS attacks for network testing purposes. The simulation of attacks can help to find weaknesses and test responses and business continuity plans. While this may not be proportionate for small and medium enterprises, there are other defences that may be suitable. While modern operating systems are better equipped to manage resources, which makes it more difficult to overflow connection tables, servers are still vulnerable to SYN flood attacks.

There are a number of common techniques to mitigate SYN flood attacks, including

It is known that exposure to SYN flood has long been known and the great damage it creates, a number of mitigation paths have been used. Some styles include :

#### 1. SYN Cache & SYN cookies

**SYN Cache :** SYN cache uses the properties of TCB for the mitigation process, As we know its role TCB (Transmission Control Block) [7]is storing the incoming communication information such as IP.

SYN Cache idea is postponing the processes of full information storing until the process is completed the three-way handshaking completes. When the attacker sends the SYN, A small amount of information about incoming communication such as IP addresses and port numbers is stored.

SYN cache divides enough space to store in the hash table by each entry in the form of a unique index, the index is calculated using IP and port number pairs. When the three-way hand-shaking is completed and the tick value is calculated using the information contained within the ACK (Acknowledgement), if a match exists, the connection is sent to the TCB (Transmission Control Block). If no match is found in the hash table, the packet is canceled[7](Figure(1.16)).

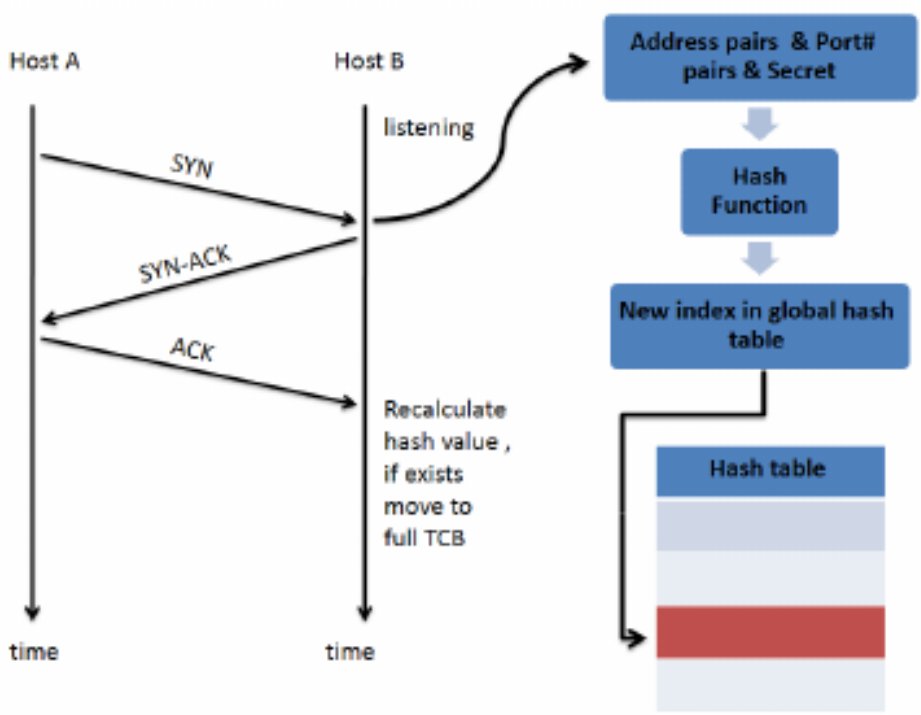



FIGURE 1.16 – SYN-Cache [7]

**SYN cookies :** In SYN cookies have the same SYN CACHE process and differs with it in storage. The cookies stores a zero state around all incoming links and delays in creating and storing information in the TCB (Transmission Control Block), until the three-way handshaking process is complete[9].

That is a sort of defense about the syn flooding where it stores informations which belongs to its sender, in order to maintain the integrity of the communication between the parties, if it sends a cookies containing the informations of the sender and when the recipient responded to their request, the sender must send a cookies in order to investigate about the informations[7](Figure(1.17)).

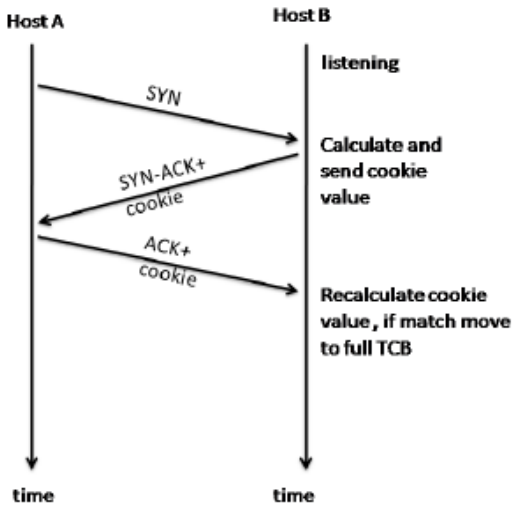



FIGURE 1.17 – SYN-Cookies[7].

**2. MPLS based DDoS mitigation mechanism** MultiProtocol Label Switching (MPLS), based Mitigation Technique. Where are directing data in telecommunications networks Each package gives its own label on entry into the service by a router called Label Switching Router (LSP), it directs its path packet until it reaches the desired address. Used by large enterprises in order to implement quality of service, To do this most protocols are used on the network[57].

On the one hand attacks have a significant impact in mitigating DDoS attack before detecting the causative agent. Where it is through the distribution of flows between the MPLS tracks that are related to the service concerned in order to mitigate DDoS volume attacks[18](Figure(1.18)).

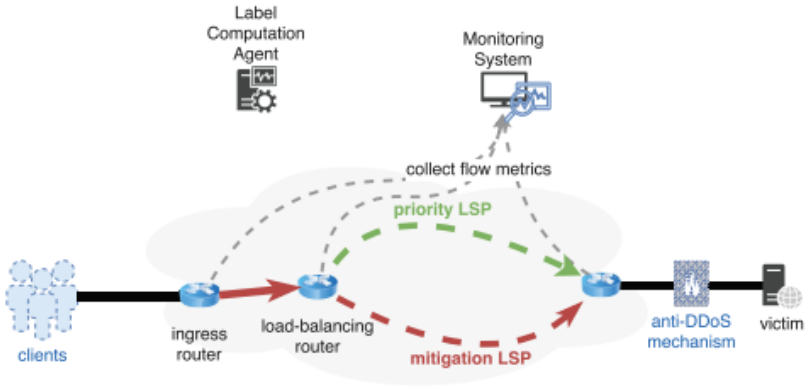



FIGURE 1.18 – Attack mitigation[18].

### 3.Software Defined Networking (SDN)

Software-defined networking (SDN) technology It has many uses in terms of speed to implement and respond to requests, help improve network performance and monitor it. Also has a significant role to mitigate DDoS attacks.

The use of SDN allows us to deal with DDoS attacks, this programming helps us to have greater control over network traffic, (Giotis et al.) proposed a DDoS mitigation scheme across multiple SDN domains or networks. Where the process of mitigation of the victim's network to the source attack, with the use of border gateway protocol (BGP). But to expose it to some problems, they managed suggestion efficient and easy to deploy collaborative DDoS mitigation scheme leveraging SDN it is C-to-C communication protocol for SDN-controllers. Where the SDN controllers are to implement the following at same time firstly, no Harmful flow within the network and Secondly, inform neighboring domains and networks of any ongoing attack.

This not only mitigate the DDoS attack within the victims's network, but also helps the transmission of attack information along the path of an attack, so we can to filter the DDoS attack they also require great resources to learn about all the attacks from various levels. SDN it is of emerging solutions in DDoS attacks (Ng et al.). Development of platform nmeta for classification of traffic SDN in the network. Developed (Hayes.) platform

nmeta into nmeta2 for to address traffic performance and scalability[6]. Used SDN traffic Classification to detect attacks on two levels SYN flooding and web application attacks(Lin et al.) [6].

**4. Pushback and cleaning center**

**Pushback :** Pushback is a mechanism for defending against distributed denial-of-service (DDoS), is a mechanism that allows a router to request adjacent upstream routers to limit the rate of traffic[32].

**Cleaning Center :** Cleaning Center provides this service at ISP (Internet Service Provider) level for multiple networks. These centers perform the necessary traffic, when an attack occurs at the network level, where cleaning when a set of destination hosts is under attack.

**5. Machine learning based Mitigation**

The machine learning of its advantage has many methods to solve problems in the world of technology, they also have solutions in the problems of attack in the network, including mitigation of the attack.

## 1.8 Conclusion

For the Denial of Service attacks and Distributed Denial of Service attacks have many ways of exploiting the protocols accurately to launch this attack, with the use of help tools to communicate and facilitate the process. The experts resort to research in how to defend it and its confrontation and non-continuity, has found many solutions, including the electronic cloud, which was exploited greatly in order to hide and not reveal the identity of the attacker. So it was prepared to dismantle this attack on way. However, the attacker continues to try and establish other ways to implement them.

At present, experts are developing other methods to address Any attack was according to its type and size. They proposed a technique of an executive character and accurate to solve any problem in the world of technology, it is technique the machine learning.

In the second chapter we will studying this technology and what its role in the world, and provide the methods used in such a security situation and others, with some examples to illustrate how to use them.

Chapitre 2

# Machine Learning

## 2.1 Introduction :

The evolution of technology has led to a diversity of science, where researchers create to keep up with these developments. Automated learning plays an important role in this development and is a basic and modern technology, that is used to predict appropriate solutions.

Machine Learning addresses many of the problems found in science, whether medical, engineering, statistics, etc. Researchers have used them extensively in their fields. There are many ways to teach the machine. Where the researchers have categorized each type according to the method used and according to the data used in this study. This chapter has therefore been devoted to clarifications of this new and evolving technique, we have started the concept of machine learning and its history in the age of technology and arithmetic.

What are their importance when most scientific fields, and After that, we clarify the different types used in machine learning, With algorithms that address these problems.

Finally, we present a variety of methods of use, which show us the usefulness of obtaining a good and accurate result(Figure(2.1)).

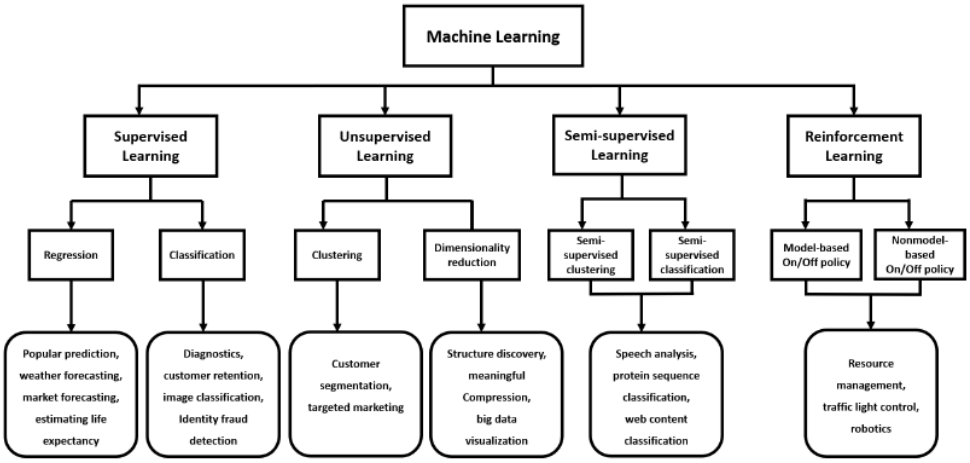



FIGURE 2.1 – Machine learning classes[30]

## 2.2 Machine Learning

Machine Learning is a method used in artificial intelligence that is an algorithm that analyze a set of data to deduce rules that constitute new knowledge to analyze new and appropriate in the following situations :

1. Problems for which existing solutions require or language rule lists.
2. Fluctuating environments : systems based on machine learning adapt to new data.

Machine Learning is about creating programs and electronic algorithms to enhance and develop many applications, such as extracting data from fraudulent credit card transactions and other applications[46].

There are many ways to learn automatically and often require the efforts of human commentators. For example, to obtain an example of a classifier classifying the protein form, one of the main biological challenges is computer science, Which requires months of expensive analysis by an expert, Crystal. The problem of combining unmarked data with data with effective markings is therefore of paramount importance for machine learning.

Before going deeply into machine learning, There are two general dataset types. One is labeled and the other one is unlabeled : [12]

— Labeled dataset D : $X = \{x^{(n)} \in R^d\}_{n=1}^N, Y = \{y^{(n)} \in R^d\}_{n=1}^N$
— Unlabeled dataset D : $X = \{x^{(n)} \in R^d\}_{n=1}^N$ [12]

Basic Technique There are 4 types of machine learning.

1. Supervised Learning
2. Unsupervised Learning
3. Semi-supervised Learning
4. Reinforcement Learning

## 2.3 Brief History a Machine Learning

machine learning has a long history of nearly 70 years, which is traded , and developed during these years was great in the field of technology, it depends on mathematics in the first place, so we will arrange the history of machine learning and how it is now.

Before 1940 : THOMAS Bayes was famous for his theory Which relates to many of the mathematical foundations of machine learning .

In 1812 : Which Pierre-Simon Laplace made definition of theory ; In 1805 : Adrien-Marie developed the method of small squares for Data structure .

In 1913 : Andrey Markov described the techniques of analysis by name Markov chains . In 1940 : the first computer was invented manually ENIAC and it has relation with machine learning because they were aspiring to develop the machine and make it able to learn .

In 1949 : A step in the field of medicine has been suggested HOP has proposed the first step towards cerebral palsy called learning theory Hebbian.

In 1952 : Arthue Samuel developed with a company IBM Program to play a DAMA game ; This is a program that monitors the situations and learning the models within the scoring function using topics on the board, However, written symbols could not be aligned ; Because it uses the mimimax strategy, which was developed at the end and became the mimimax algorithm and designed SAMUEL a number of mechani to make it a better program .

In 1967 : the nearest adjacent algorithm was designed for mapping and methods which were the solution to the problem of a salesman to obtain a lower road. Marcello Pelillo obtained the best in invention .

In 1995 : It was the first appearance of the SVM proposed by Vapmik and logistic regression and it was very useful and at this time ML was advancing any two groups NN(Neural Network) and SVM (Support Vector Machines).

In 2001 : BREIMAN discovered the decision tree, followed by the emergence of deep learning, which was used in NN (Neural Network).

In 2006, the FRGC (Face Recognitition Grand Challenge) program was launched The algorithm evaluates face recognition .

In 2012 : Google XX Lab developed an ML algorithm that can browse videos To locate cats on video[45].

## 2.4 Importance of machine learning

Technology in our time is very important , and with its development led to solving many of the problems. The role of the technique of machine learning it was effective, so we presented some of the importance of machine learning and where to use.

### 2.4.1 Data storage

The importance of large companies and institutions in storing their data and their order precisely ; Regardless of the number and nature of such data. The more larger the data, a more difficult to store and how to extract it. so machine learning it is Facilitate for this operation in the database.

### 2.4.2 Spam filtering

Email is a way to send millions of advertisementse inexpensive , so there are who take exploited of this opportunity by several organizations. As a result, Email boxes for millions of people are cluttered with so-called "unsolicited bulk e-mail also known as “spam” or “junk mail”. Since these spam messages are not useful, they affect the server because of the large amount that delays the delivery of legitimate e-mail. And machine learning algorithms has a solution in classification spam emails and nonspam emails (Figure(2.2)).

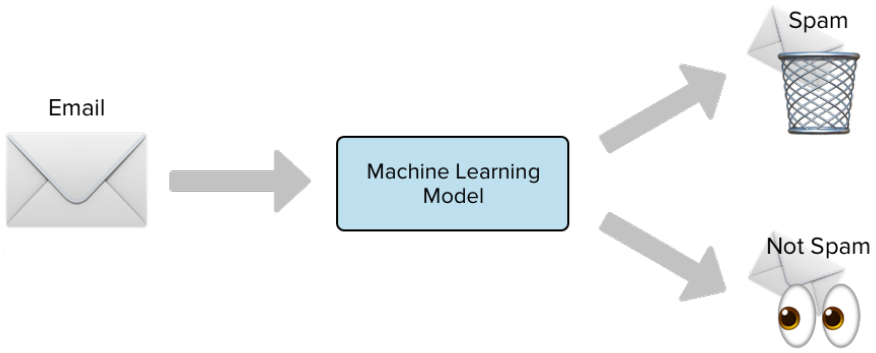



FIGURE 2.2 – classification spam and not spam

### 2.4.3 Search engines

Browser searches were previously difficult and heavy to find out what is required, so some of the developments for strong, continuous and rapid use. Machine Learning uses supervised as a solution to search engines.

### 2.4.4 Handwritten digits recognition

From the diversity of the writing of humanity, it constitutes a collection of numbers that have the same meaning but with different writing. Thus, machine learning has a way of identifying and knows the numbers whatever their differences (Figure(2.3)).

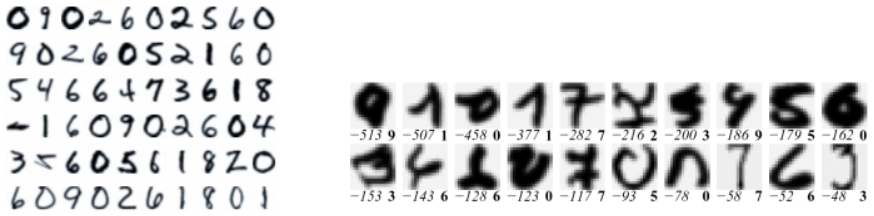


FIGURE 2.3 – Handwritten digits recognition(Mnist database handwritten digits)[63]

### 2.4.5 Forecasting in urgent situations

One of the reasons for Machine Learning is Benefited from the learning power, Many of the fields help them in advance detection of cases. Medicine

today is highly dependent on technology because of its potential to detect and treat the disease accurately.

## 2.5 The various fields in which machine learning exists

Machine learning is also related to other specialties such as big data, statistic, artificial neural networks, artificial intelligence, security, etc. machine learning focus more on why machine can learn and how to model, optimize, organization and order to make the best solution for use it in the projects(Figure(2.4)).

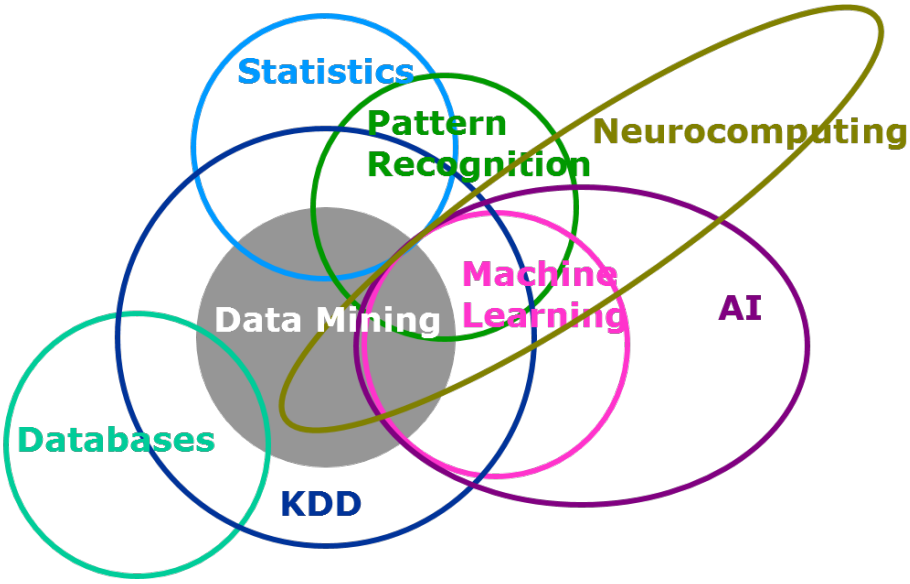



FIGURE 2.4 – Different disciplines of knowledge and the discipline of machine learning.[47]

### 2.5.1 Big data

Before we know the term "big data," we present its characteristics that distinguish it from the previously known database concept : volume , velocity, variety, veracity and value.

To illustrate these five characteristics, it is the amount of information that exists and the large nature in different fields, e.g images, files, tables or even

databases. And the veracity of this information is compatible with its source or where it is sent for storage and processing, and the value of this information which is important to the extent of providing the means to store it.

The important thing is not just the data, but how it is processed and controlled, it differs from traditional data for its high speed. Since the data have grown up so large that the inflation catastrophe, the world resorted to machine learning that works on multiple and diverse data to provide the best result. machine learning has many algorithms to help solve a problem, including big data . Here are some algorithms : linear regression and logistic regression, SVM ans neural network,and decision trees.

### 2.5.2 Artificial Intelligence

There are many areas that use artificial intelligence :

- Database mining : Search engines, Spam filters ,Medical and biological records.
- Big Data is boosting intelligent behavior in machines.
- Intuitive for Humans is recognizing faces in images or spoken words, based to supervision and hard-coded logical inference rules and learning ability.
- play chess (IBM's Deep Blue 1997) based to artificial intelligence it is easy for computers but difficult for humans.

All this the A.I. system needs to acquire its own knowledge, This capability is known as machine learning (ML). Artificial intelligence is part of computer science, it does what man does and is intelligent. The machine learning is the means that gives the ability to learn and improve, including intelligence.

The techniques used by artificial intelligence are learning algorithms, The reasons for learning is that the learning algorithm, implemented by Google or Microsoft for how to order pages and Facebook in order to get to know friends' photos is all this through automated learning.

### 2.5.3 Statistical

Statistical learning has a wide range of tools to understand data, it plays a key role in many fields of science, finance, industry and health.

For example :

— Predict whether the patient, entered the hospital because of a heart attack , that it will have a second heart attack. Prediction is to take the necessary precautions and diet for this patient.

— Forest cover type Prediction prediction project is classifying different forest types from field observations. This is a multi-class classification problem with 7 classes. All this need to machine learning .

The role of machine learning in the areas of statistics, mining data and artificial intelligence, intersecting with other fields of engineering and disciplines.

type used for statistics to controlled their data : Unsupervised learning : the objective is to find something useful in the data, it is more specified than supervised learning. The technique or method used is clustering (grouping the data based on their similarity), density estimation (estimating the probability distribution behind the data), anomaly detection (removing outliers from the data).

### 2.5.4 Security

Cybersecurity is a set of technologies and processes that have an important role to protect computers, networks, programs and data from attacks[59].

Therefore, designers of network security must To prepare to each attack and the method of detection and protection from it. So they resorted to the modern technology of machine learning, which has a variety of ways in this field.

Application of machine learning technology (ML) in cybersecurity more than before, of which :

— IP (Internet Protocol) traffic classification,

— Traffic classification of the hackr.

— Effective against new attack threats (zero-day).

It has also been defined on multiple attacks and the performance of each algorithm in machine learning has been evaluated, used in

cyber-security in three main areas : IDS(intrusion detection system), Anomaly detection module and misuse detection.
— Clustering algorithm works in Anomaly detection, it is a high speed that is easily processed.
— SVM (support vector machines)also works on Anomaly detection and misuse detection, and this technique requires high expectations.
— The decision tree is one of the algorithms that help to create expectations suitable for this area.
— but due to the advent of support vector machines (SVMs) Algorithms ANN that contain hidden nodes are not entirely appropriate.

## 2.6 Different techniques of machine learning

Machine Learning classifies a set of algorithms on multiple techniques according to the way they are used. this the techniques are required to improve the accuracy of predictive models, and With the ever increasing amounts of data becoming available, These techniques help to solve their problems.We will explain the techniques used a lot in the field of machine learning[47].

### 2.6.1 Supervised Learning

Supervised learning typically begins with an established set of data. They are executed when specific targets are defined to access them from a given set of inputs , the aim Of which is often to get the computer to learn a classification system[5]. For example, there could be millions of Various trees and Includes explanation of each tree is and then you can create a machine learning application that distinguishes one tree from another. Is the most common technique for training neural networks Depends on information for classification and determine the error of the network and then adjust the network to minimize it[5], More generally, classification learning is appropriate for any problem where deducing a classification is useful. In our previous example the classification is used for when the data comes from a finite set of values, As for regression use for supervised learning helps you understand the correlation between variables. For example,weather forecasting [30].

1. **Regression :**

   Regression is another prototypical application, the goal is to estimate a real-valued variable $x \in R$[63], the regression learning problem

is to learn a function estimator from examples $(x_i, f(x_i))$ There are many type of Regression, Of which Simple Linear Regression, This is one of the most common and interesting type of Regression technique, and Logistic Regression[63].

**Simple linear regression :** Is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.
— representation model.
— the cost function

(a) function : $h_0(x^i) = \theta_0 + \theta_1 x^i$
(b) the cost : $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_0(x^i) - y^i)^2$
(c) goal : minimization $J(\theta_0, \theta_1)$

**Logistic Regression :** Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
— function of logistic regression :

$$\frac{1}{2}||\theta||^2 + \sum_{i=1}^{m} log(1 + exp(< -y_i \varnothing(x_i), \theta >))[63].$$

2. **Classification :**

Classification is the most common task in machine learning and is a technique for determining class the dependent belongs to based on the one or more independent variables. used for predicting discrete responses. Researchers have identified a large range of templates, which can be used to address a large set of situations. Which makes machine learning easy, One of these Binary Classification and binary classification and 3-class classification[63](Figure(2.5)).

linear classification ; separate stars from diamonds. In this example we are able to do so by drawing a straight line which separates both sets (Figure(2.6)).
Left : binary classification. Right : 3-class classification. Note that in the latter case we have much more degree for ambiguity.

FIGURE 2.5 – linear classification[63]

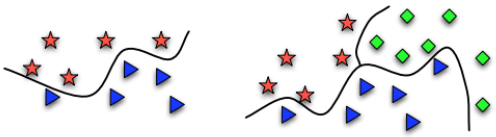


FIGURE 2.6 – binary classification and 3-class classification. [63]

### 2.6.2 Unsupervised Learning

Unsupervised learning algorithms according to Ghahramani (Ghahramani, 2008)[5] are designed to extract structure from data samples. is much harder Because it works with the data that is unlabeled, the goal is to have the computer learn how to do something With undefined information We do not know what to do with it[5]. Unsupervised learning They are used on all have large amounts of unlabeled data, to classify them according to the patterns or clusters to which they are attributed. There are two approaches to unsupervised learning. The first approach Clustering, the goal simply to find similarities in the training data. For example, used with email spam-detecting technology. Unsupervised learning algorithms segment data into groups (clusters) or clusters of features. This technique is based on repetition, the more information the machine learns to classify into similar clusters and are used to make decisions about To the limits of convergence between them. As For the second type Dimensionality reduction This technique is based on a study widely in machine learning community, working data to a low dimensional latent space while preserving some properties of the original data.

1. **Clustering :**

   Clustering is one of the most widely used techniques for exploratory data analysis[60], an unsupervised learning technique data is not tagged, it contains quite a few algorithms to segment an accumulated data collection forms groups according to types of similarities

between both, we can measure the quality of clustering by intra and inter-cluster distances to measure respectively the approximation of the examples of each cluster and the distance of clusters from each other.

2. **Dimensionality Reduction :** Dimensionality reduction has been studied widely in machine learning community. (van der Maaten, Postma, and van den Herik 2007)[51], Action of dimensionality reduction is how extracting low-dimensional structure from high-dimensional data—arises, this technique is very important in machine learning. It is divided into two methods :
   (a) Linear Dimensionality Reduction Methods
   (b) Non-linear Dimensionality Reduction Methods or Graph-Based Methods.

Where Linear Dimensionality Reduction it has many methods and the most common and known ; PCA (Principal Component Analysis) used for the low-dimensional representation of a high dimensional data set. Where input patterns $x_i \in \mathbb{R}^d$ are apply it into the m-dimensional for minimizes the reconstruction error.

$$\text{PCA } \sum_i ||x_i - \sum_{a=1}^{m}(x_i \cdot e_a)e_a||^2.$$

Where the vectors ${e_a}_{a=1}^{m}$ by it are determine a partial or thonormal of the input space[**?**]. There are many methods such as Factor Analysis

The second method of Non-linear Dimensionality Reduction or Graph-Based is Multi-dimensional scaling (MDS), Non-linear are used when the data doesn't lie on a linear subspace. and MDS it is one popular learning methods. A technique used for computing the low-dimensional representation of a high dimensional data set, It keeps the input unchanged. Where The outputs $\varphi_i \in \mathbb{R}^m$ are apply it for minimizes.

$$MDS = \sum_{ij}(x_i \cdot x_j - \varphi_i \cdot \varphi_j)^2$$

the Gram matrix is solution for the minimum error where formul it is :

$G_{ij} = x_i \cdot x_j$.

### 2.6.3 Semi-supervised Learning

Semi-supervised learning (SSL) dates back to the 60s[13], is a type of machine learning Where Semi-supervised learning falls between unsupervised learning and supervised learning, Use Collection Variety Of the data both labeled and unlabeled data to generate an appropriate function or classifier[5]. Semi-supervised learning uses the unlabeled data to gain more understanding of that the situation. Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy over unsupervised learning.

### 2.6.4 Reinforcement Learning

Reinforcement learning is different from supervised learning, The learner is not told the instructions and decisions he makes, but instead must rather than the style of discovery, Which gets great results. Reinforcement learning (RL) is an area of machine learning concerned with how software agents ought to take actions In the situation studied The goal is to make the system respond comprehensively so that it works correctly in situations not in the training group, but alone it is not adequate for learning. For example teaching someone to play a game. The rules and objectives are clearly defined. However, the outcome of any single game depends on the judgment of the player who must adjust his approach in response to the incumbent environment, skill and actions of a given opponent[5].

## 2.7 Application of Machine Learning Algorithms

Where there is Many Machine learning algorithms, we will address the well-known and frequently used algorithms.

### 2.7.1 Classification and Regression

Classification and regression are two ways of Supervised Learning availability of a training set of examples labelled, There are several examples of the application of these two characteristics in order to obtain a convincing result

that helps to study the situation of a situation. for Example of a spam e-mail, we know Spam e-mail filtering is a good example of binary classification.

**1. Support Vector Machines (SVM)**

an iterative algorithm can be designed which scans through the dataset looking for violators.

$$f(x) = \sum_{x_j \in S} \alpha_j y_j K(x_j, x) + b[68].$$

Where $x_i$ denotes the training patterns, $y_i \in (+l, -1)$ denotes the corresponding class labels and S denotes the set of Support Vectors . In order to finding the closest pair of points, we uses an iterative algorithm which scans the existing data for to find a violation point. The algorithm stops when all points are classified within an error bound. To explain this we present simple SVM algorithm[68].

**Algorithm 1** Simple SVM [68]

| | |
|---|---|
| 1 | candidateSV = { closest pair from opposite classes } SV : Support Vectors |
| 2 | **while** there are violating points **do** |
| 3 | Find a violator |
| 4 | candidateSV = candidateSV ∪ violator |
| 5 | **if** any $\alpha_p < 0$ due to addition of c to S **then** |
| 6 | candidateSV = candidateSV $\backslash p$ |
| 7 | repeat till all such points are pruned |
| 8 | **end if** |
| 9 | **end while** |

**Example of a spam e-mail :**

There are several methods in the classification, some of them are mentioned : Example of a spam e-mail.

— $x_1$ :The quick brown fox jumped over the lazy dog.

— $x_2$ : The dog hunts a fox.

and associated labels $y_i$, denoted by $Y = \{y_1, ....y_m\}$ ,the labels satisfy $y_i \in \{spam, ham\}$.

| | the | quick | brown | fox | jumped | over | lazy | dog | hunts | a |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| $x_2$ | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |

TABLE 2.1 – Vector space representation of strings.

**2.Naive Bayes**

Is a simple technique for constructing classifiers, but surprisingly powerful algorithm for predictive .where the class labels are drawn from some finite set. Each value is given x conditional probability, can be used to make predictions for new data using Bayes Theorem. naive Bayes classifiers can be trained very efficiently in a supervised learning setting[63]. This simple algorithm is known to perform surprisingly well, it can be found in most modern spam filters.

The general idea in the naive algorithm is to count on the frequency of a word in a text if it is y= spam or y=ham. search of information if was ham or spam must reads documents x and labels y, and after compare between for find the repeated word. To explain this we present Naive Bayes algorithm[63].

**Algorithm 2** Naive Bayes[63]

| | |
|---|---|
| 1 | **Initialize** $b := logc + logm_{ham}$ to offset the rejection threshold |
| 2 | **Initializer** $P \in \mathbb{R}^{2\times n}$ whith $P_{ij} = 1, w_{spam} = n, w_{ham} = n.$ |
| 3 | {Count occurrence of each word} |
| 4 | {Here $x_j^i$ denotes the number of times word j occurs in document $x_i$} |
| 5 | **for** $i = 1$ to m do |
| 6 | **if** $y_i = spam$ **then** |
| 7 | **for** $j = 1$to n do |
| 8 | $P_{0;j} \leftarrow P_{0;j} + x_i^j$ |
| 9 | $w_{spam} \leftarrow w_{spam} + x_i^j$ |
| 10 | end for |
| 11 | else |
| 12 | **for** $j = 1$ to n do |
| 13 | $P_{1;j} \leftarrow P_{1;j} + x_i^j$ |
| 14 | $w_{ham} \leftarrow w_{ham} + x_i^j$ |
| 15 | end for |
| 16 | end if |
| 17 | end for |
| 18 | {Normalize counts to yield word probabilities} |
| 19 | **for** $j = 1$ to n do |
| 20 | $P_{0;j} \leftarrow p_{0;j}/w_{spam}$ |
| 21 | $P_{1;j} \leftarrow P_{1;j}/w_{ham}$ |
| 22 | and for |
| 23 | Classify($x$) {classify(x) document x} |
| 24 | Initialize score threshold $t = -b$ |
| 25 | **for** $j = 1$ to n do |
| 26 | $t \leftarrow t + x^j(logP_{0;j} - logP_{1;j})$ |
| 27 | end for |
| 28 | **if** t $> 0$ **return** spam else**return** ham |

### 3. k-Nearest Neighbor (K-NN) Classification

The near neighbor approach can easily be used on regression problems by using a real target variable value But if we have a way to compile the examples we use the k-nearest neighbour method, When the $k \geq 1$. An even simpler estimator than Naive Bayes is nearest The kNN classifier is based on a distance function that measures the difference or similarity between two instances. The standard Euclidean distance d(x, y) between two instances x and y is defined as : $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$[72] This simple algorithm

k-Nearest Neighbor Classification[63].

**Algorithm 3** k-Nearest Neighbor Classification [63]

| | |
|---|---|
| 1 | Classify(X,Y,x) {reads documents X, labels Y and query x} |
| 2 | **for** $i = 1$ to m do |
| 3 | Compute distance $d(x_i, x)$ |
| 4 | **end for** |
| 5 | Compute set I containing indices doe the k smallest distances $d(x_i, x)$. |
| 6 | **return** majority label of $\{y_i where i \in I$ |

## 4.Decision trees

A decision tree is a predictor,Where it contains node represents a single input variable (x) and nodes of the tree contain an output variable (y), used to make a prediction$h : X \to Y$ Decision Tree learning is one of the most widely used and practical methods for inductive inference over supervised data.can simply define a set of instances D to be homogenous if they are all from the same class.

For to construct the decision tree :
- Top-bottom algorithm.
- Impurity measure.

In Top-bottom algorithm,find the best split condition and Stops when no improvement possible.

Impurity measure,measures how well are the two classes separated.so we explain that by Impurity functions.

Imp($\{D_1, ....D_l\}$)=$\sum_{j=1}^{l} \frac{|D_j|}{|D|} Imp(D_j)$ ; where D =$D_1 \cup ... \cup D_l$ .

we provide a possible implementation. It is based on a popular decision tree algorithm known as ID3" (short for Iterative Dichotomizer 3")[60].

**Algorithm 4** Decision tree [60]

| | |
|---|---|
| 1 | Input : training set S, feature subset $A \subseteq [d]$ |
| 2 | **if** all examples in S are labeled by 1, return a leaf 1 |
| 3 | **if** all examples in S are labeled by 0, return a leaf 0 |
| 4 | **if** $A = \varnothing$, return a leaf whose value = majority of labels in S |
| 5 | **else :** |
| 6 | Let j= $argmax_{i \in A}$Gain(S, i) |
| 7 | **if** all examples in S have the same label |
| 8 | **Return** a leaf whose value = majority of labels in S |
| 9 | **else** |
| 10 | Let T1 be the tree returned by ID3($\{(x, y) \in S : x_j = 1\}, A \setminus \{j\}$ |
| 11 | Let T2 be the tree returned by ID3($\{(x, y) \in S : x_j = 0\}, A \setminus \{j\}$ |
| 12 | **Return** the tree |

### 5.Neural networks

An Artificial Neuron Network (ANN), popularly known as Neural Network,[27] is a computational model based on the structure and functions of biological neural networks.

Neural networks (Bishop C. M., 1995) can actually perform a number of regression and/or classification tasks at once[5], is composed of several interconnected neurons, Neural networks are a set of algorithms, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception,Neural networks help us cluster and classify. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on. and Neural Network Elements :(Figure(3.26)).

Neural network, is a structure consisting of three layers : input layer ; hidden layer ; output layer, Connected with each other and each entrance gives it weight.

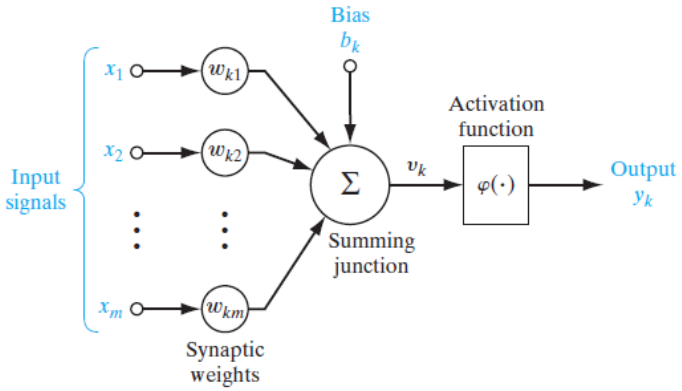



FIGURE 2.7 – Nonlinear model of a neuron, labeled k[27]

.

### 2.7.2 Clustering

The aim of the assembly is to extract information by forming groups separating the data as described earlier. Its famous algorithms are k mean . Where it discover better congestion through the repetition.

#### 1. K-Means Clustering

"K-means clustering (MacQueen, 1967)"[61] where apply algorithm profusely because very easy of understand and implement. Used to automatically divide the data set into k groups. Applicable to data of big sizes, by choosing a good notion of distance, The greater disadvantage of k-means is required of data the number of the cluster (group) as parameter[63].

This algorithm to prove how create cluster by k means : Note : k is a number class. In order to reach the goal to cluster , it must pass through stages, Before this should reduce the data, where it was reduces the spaces between points by : $j(\mu, r) := \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{k} r_{ij}||x_i - \mu_j||^2$ where $r = \{r_{ij}\}, \mu = \{\mu_j\}, and||.||^2$ denotes the usual Euclidean square norm.

**Stage 1 :** In order found point $x_i$. save the $\mu$ fixed and determine r. $r_{ij} = 1if j = argmin||x_i - \mu_j||^2$

**Stage 2 :**. for all j uses operation to minimized by setting the derivative. so, save the r fixed and determine $\mu$. The calculation Which are doing is : $\sum_{i=1}^{m} r_{ij}(x_i - \mu_i) = 0$

**Algorithm 5** k-means.[63]

| | |
|---|---|
| 1 | Cluster(X) {Cluster dataset X} |
| 2 | Initialize cluster centers $\mu_j$ for $j = 1, ...., k$ randomly |
| 3 | **for** $i = 1$ to m do |
| 4 | Compute $j' = argmin_{j=1,.....,k} d(x_i, \mu_j)$ |
| 5 | Set $r_{ij'} = 1 and r_{ij} = 0 for all j' \neq j$ |
| 6 | **end for** |
| 7 | **for** $j = 1$ to k do |
| 8 | Compute $\mu_j = \frac{\sum_i r_{ij} xi}{\sum_i r_{ij}}$ |
| 9 | **end for** |
| 10 | until Cluster assignments$r_{ij}$ are unchanged |
| 11 | return $\{\mu_1, ....., \mu_k\} and r_{ij}$ |

**2.Hierarchical clustering**

Hierarchical clustering it is a structure are divided into two algorithms are [38] top-down(divisible) or bottom-up(agglomerative). finds successive clusters using previously established clusters In this type of clustering can not known the number of groups exist. The system takes the data set as input and outputs cluster tree. There are two classes of this kind of algorithms : divisible algorithms that start from a set of data and subdivide it into subsets then subdivide each subset into smaller ones, and so on, for generate at the end an ordered cluster sequence from the most general to the finest. The second class is that of agglomerative algorithms that consider each record as an independent cluster and then gather the closest ones into larger clusters, so on until they reach a single cluster containing all the data.
There is other technique like, clustering partitions, clustering based density etc..

## 2.8 Conclusion

In the machine learning exist a lot the algorithms used according to their needs and the nature of the data entered, for solving problems.

With the different techniques and how to use them, they are also varied in the way of solving. As the problems develop, machine learning evolves, so we present the problem of this subject in which we use machine learning to study it and how to avoid its risks. They are DDoS attacks .As we explained in this

chapter about kNN and Decision Tree algorithm is the method used to solve our current problem it uses labeled data. whereas k mean is an appropriate solution because it uses unlabeled data. As per our study of machine learning, we consider Semi-supervised Learning every solution to similar problems, Because it uses different data labeled and unlabeled, It needs resources to learn about all attacks and can classify the attack according to the source .

The following chapter provides an explanation of how to use machine learning in attacks in order to mitigate it, and What technique is used and The way to study the attack.

Chapitre 3

# Machine Learning solution for DDoS mitigation

## 3.1 Introduction :

Traditional techniques can not help solve the problem of online attacks permanently. This calls for an effective strategy to reduce network resources along the route of the attack from source to victim. The effects of DDoS can be disastrous for your service, however this must match the mitigation strategy with the type attack.

The options for DDoS mitigation are plentiful, and implementing the right solution against the exact attack Not conclusive. In this chapter we will to study the attack on the site of the University of Ghardaia, we will provide explanations on mitigation and its stages, with the proposal of machine learning technology to study DDoS attacks in to application layer or another layer and input analysis but in this attack we have, we apply a method that will mitigate the attack, to stop an HTTP GET flood, reach the mitigation process, we are going through stages as explained(Figure(3.1)) [67].

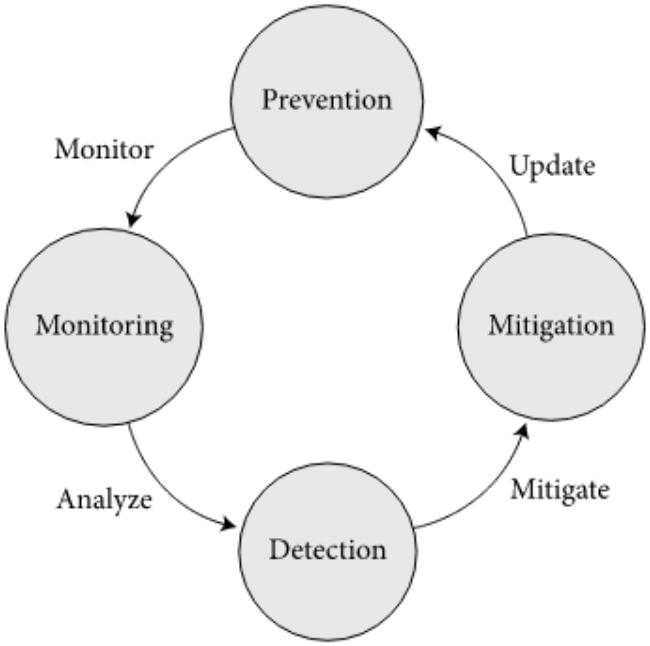



FIGURE 3.1 – DDoS defence life cycle[34].

## 3.2 What intended DDoS Detection and Mitigation ?

Many researchers and network protection specialists on the detection of attacks on the network, used a variety of methods, but did not stop, used machine learning to help them discover the exactness and goodness of these attacks and also to learn and improve its future execution from the results of previous execution and to improve its performance over the time.

Either on the final stage A few class use the attack mitigation, An attack is detected by checking if there is an abnormal attack on the network such as flooding the Requests on sites, for example or sending packets at once. Discovery can occur on the server by monitoring all of the outgoing/incoming. The attack can be mitigated if a malfunction occurs and no service continues, packages are blocked[20]. so, mitigation DDoS is a set of techniques or tools for mitigating the impact of distributed denial-of-service on networks.

**Detection :** Detection is one of the stages to reach the result of mitigation. Experts have tried the detection process on the application layer, Where Application Layer Anomaly Detection (ALAD) can be combined with Packet Header Anomaly Detection (PHAD) to improve detection level but does not currently examine application layer protocols like DNS, HTTP or SMTP[41], and by Application Layer Anomaly Detection Showing communication results TCP connections that are assembled from packets(Figure(3.2)).

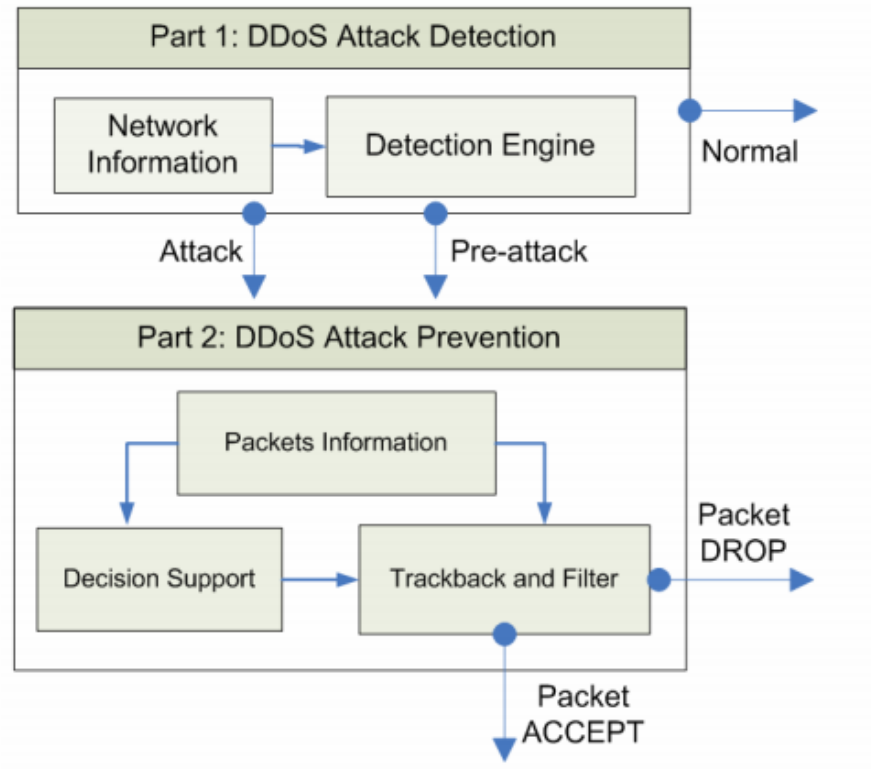



FIGURE 3.2 – A simple anti-DDoS framework[49]

.

The first things to do in DDoS mitigation it is to identify normal conditions for network traffic by knowing the type of protocol used, including IP addresses. After the detection is performed, the filtering process is performed, which can be performed through the anti-DDoS technique by mitigation(Figure(3.3)).

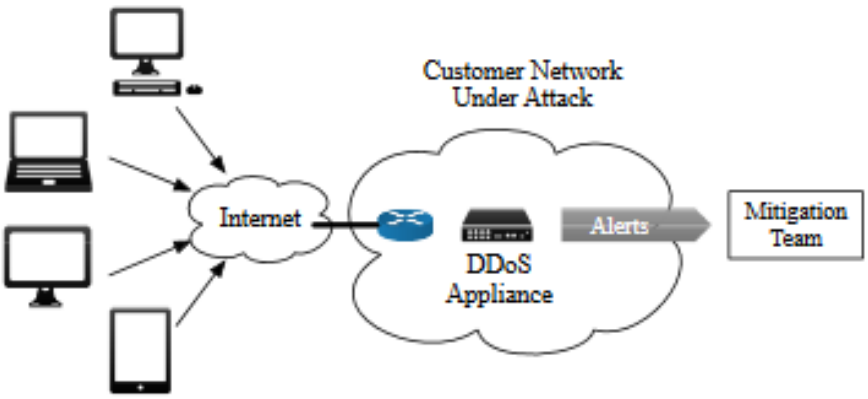



FIGURE 3.3 – A DDoS Internet Defense Network[35].

## 3.3 Algorithms for Detection and technique mitigation of DDoS attacks

DDoS attacks have a great fame in different fields For its malignant impact, and with the machine learning used algorithms in detecting and fighting this attack we offer some of the algorithms used according to their needs in order to learn according to the size and type of data in the detection in the first class(Figure(3.4)).

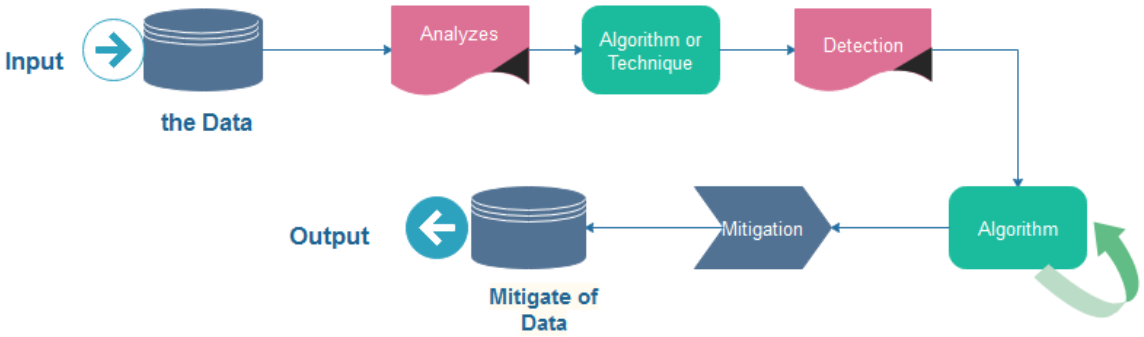



FIGURE 3.4 – Structuring Mitigation Attack

### 3.3.1 K-means ++ and one-class SVM

In this example in order to detect which attack was engineered using Low Orbit Ion Canon (LOIC) tool, Where both were affected TCP and UDP to flooding . The first step in detecting this attack is to capture all imports in traffic, and then apply an algorithm to detect the cause of this attack and to process it, where use two a learning algorithm K-means ++ and one-class SVM.

**Detection :** For Detection attack must pass on two process. In the first pass detection set was clustered using K-means++ algorithm.

In this example they used k-means++ algorithm it is an improvement of k means. In k-means++ algorithm ; Trying K find to shortest distance from a data point x∈ X to the nearest selected point. Are training all existing data, to for labeling IP addresses regular and unusual addresses. The midpoint or centroid was to form training groups ; in two dimensional space, in order to reduce the space into two dimensions without losing much information used PCA (Principal-Component Analysis).[20] PCA is a statistical procedure for dimension-reduction to reduce a large and big set into small set, it retains the contents of the information of the large set [24].

**Algorithm 6** k-means++ [20]

| | |
|---|---|
| 1 | Select a centroid $\mu_1$ , chosen uniformly at random from X ,as first centroid |
| 2 | Choose a new centroid $\mu_i \in X$, such that the probability $\frac{D(\mu_i)^2}{\sum_{x \in X} D(x)^2}$ is highest |
| 3 | **Repeat Step 2 :** until all k centroids are taken. |
| 4 | Proceed with the Lloyd's k-means algorithm, skipping random initialization stage. |

The second pass, When it cluster data label, used one-class SVM classifier [43] to decide the boundaries, Where this algorithm is distinguished in identifying anomalies in a data ; Which trains each address IP. So they compiled addresses IP using unsupervised learning algorithm k-means, then applying one-class SVM on the clusters to decide their boundaries.

One-class SVM, is an extension of SVM where classification is an optimization problem stated as follows :

$$min_{w,b,e^i} \frac{||w||2}{2} + \mathrm{C} \sum_{i=1}^{m} = e^i :$$

where $C > 0$ is the regularization parameter.

thus, One-class SVM where creates binary function which returns +1 for the training, otherwise it returns -1. Following is the optimization function of one-class SVM. They are simpler than the SVM.

using a non-linear function $x \rightarrow \varnothing(x)$ . The hyperplane is represented with the equation ($w\varnothing$ (x) $\geq$) $\rho$, where x $\in R^d$,w $\in$ F ; It separates all data point in feature space F, and maximizes the distance of hyperplane from F. To avoid over-fitting, slack variable $e^i$ is introduced. For separate the data set from the origin, need to solve the following quadratic programming problem :[44]

$$min_{w,\rho,e^i} \frac{||w||^2}{2} + \frac{1}{vm} \sum_{i=1}^{m} e^i \text{ - } \rho$$

Subject to :

$$(w.\varnothing(x^i)) \geq \rho - e^i \textbf{ for all } i = 1, \ldots, m$$

$$e^i \geq 0 \textbf{ for all } i = 1, \ldots .., m$$

where v $\in$ (0 ; 1] introduced in place of C ; is used to set upper bound on outliers/anomalies, and lower bound on the number of training examples[71] and $x^i \in R^d$ is the training.

**Mitigation :** For mitigate in this problem using Network Function Virtualization (NFV)[19], NFV is an advanced technology which allows us to manage network devices remotely, is transferring the network functions that run on standard processors where that network functions are implemented and provided in software, This virtualized network enables us to deploy new network services in a flexible manner. It can know where the attack is originated and how intense it is, can know the nature of the attack ; The control of blocking traffic at a router can also be given to the destination and block all addresses IP which Causing to flooding.

### 3.3.2 Semi-supervised used Co-clustering and Exra-Trees algorithm

In usually there are solutions for attack and detection are done on a supervised and unsupervised level ; in this time used both in one technique it is Semi-supervised, where algorithms which apply by supervised and unsupervised. Thus, (Mohiuddin A.) [1] proposed the co-clustering algorithm an unsupervised approach for detection DDoS. Co-clustering is a set of technique in

Cluster Analysis, its potential to discover latent local patterns, Co-clustering algorithm which is much smaller than that of the traditional Kmeans algorithm, where The co-clustering computational complexity is O(m.k.l + n.k.l) and K means O(m.n.k). Where m is the number of rows, n is the number of columns, k is the number of clusters and l is the number of column clusters. which it help to Dimensionality reduction, it is used in compressed data with preservation of information in the original data[31].

Secondly, an supervised prposed the Extremely randomized trees (Extra-Trees) algorithm [31]. The Extra-Trees algorithm creates the tree for classification and regression problems according to the classic top-down procedure, where it splits nodes by choosing the cut points completely randomly for learning using a sample to complete the tree development[21].

The complexity of the tree growing procedure is on the order of O(M.N.K.logN) with respect to learning sample size. Where, N is the number of samples, K is the number of variables randomly drawn from each node and M represents the number of trees in this ensemble[31].

In order network traffic entropy estimation, are used entropy a measure to helps to distribute network flow ; and allows to reduce the high dimensions of traffic distribution, in coordination with FSD (flow size distribution). Because during an attack, botnet sends a large number of packages to the victim where he creates network flow data. hence, it used FSD to distinguish it from detecting sudden changes in network flow ; Is Called FSD entrpy.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \cdot log(P(x_i))$$

Where X represents a FSD feature $x_i(i, ....., n)$, where $x_i$ the frequencies of items of X ; and n is the total number of items of X, so, for the FSD entropy values was make it natural or together using the formula : $H_0(X) = \frac{H(x)}{logn}$ [31].

The follwing algorithm allows us to use only two, lower ($\delta_l$) and upper ($\delta_u$) for each S in order to detect exception in network traffic[31](Figure(3.5)).

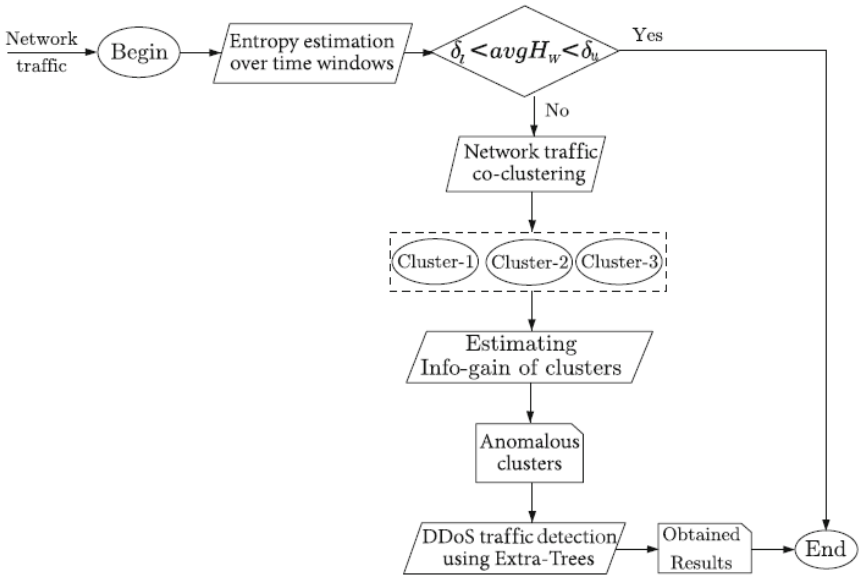



FIGURE 3.5 – Flowchart of the proposed approach [31]

**Algorithm 7** Semi-supervised DDoS detection approach[31].

| | |
|---|---|
| 1 | **Input**network traffic data, time window size w |
| 2 | **Output**Classification Resultas |
| 3 | $S \leftarrow \{src_pkts, dst_pkts, src_bytes, dst_bytes$ |
| 4 | //FSD features set |
| 5 | **while**incoming network traffic **do** |
| 6 | **foreach**W **do** |
| 7 | $dataw \leftarrow$ |
| 8 | {incoming network traffic data} |
| 9 | $avgH^w \leftarrow avgEntropy(dataw, S)$ |
| 10 | **if**$avgH^w \notin ]\delta_u, \delta_l[$**then** |
| 11 | clustrers $\leftarrow Co - Clustering(dataw)$ |
| 12 | **for** cl **IN** clusters **do** |
| 13 | Gain (data w,cl)$\leftarrow$ |
| 14 | avgEntropy(data w,S) - avgEntropy(cl,S)* $\frac{sizeofcl}{sizeofdataw}$ |
| 15 | **end** |
| 16 | AnomalousClusters $\leftarrow$ |
| 17 | {clusters\| min (Gain(data w ,cl))} |
| 18 | Preprocessing (AnomalousClusters) |
| 19 | Results $\leftarrow$ |
| 20 | ExtraTrees(AnomalousClusters) |
| 21 | **end** |
| 22 | **end** |
| 23 | **end** |

### 3.3.3 Random Tree Machine Learning Algorithm

In this attack choose to explain the algorithm applied to web servers, Where it is used the IDS ( intrusion detection system) /IPS( intrusion prevention system). Was estimated it can producing a False Positive (FP) which affects users Web where They block traffic in network.[48] Where Use the random tree learning algorithm to help identify the False Positive and direct them to their intended destination. For this, the random tree algorithm uses rules to help it adjust it is Snort NIPS (Snort Network Intrusion Prevention System) The choice of the decision tree was to classification model to traffic are malicious and normal traffic[48].

The Real server Web hosts, it is Web site where you can listen to all requests received through the Bait server Web Which in turn communicates through it attackers, after The Bait server Web determines what to do with the received traffic, it sends authenticated traffic to the Real server Web and unauthenticated traffic to the Decoy server Web ,Which is illustrated in the following[48](Figure(3.6)).

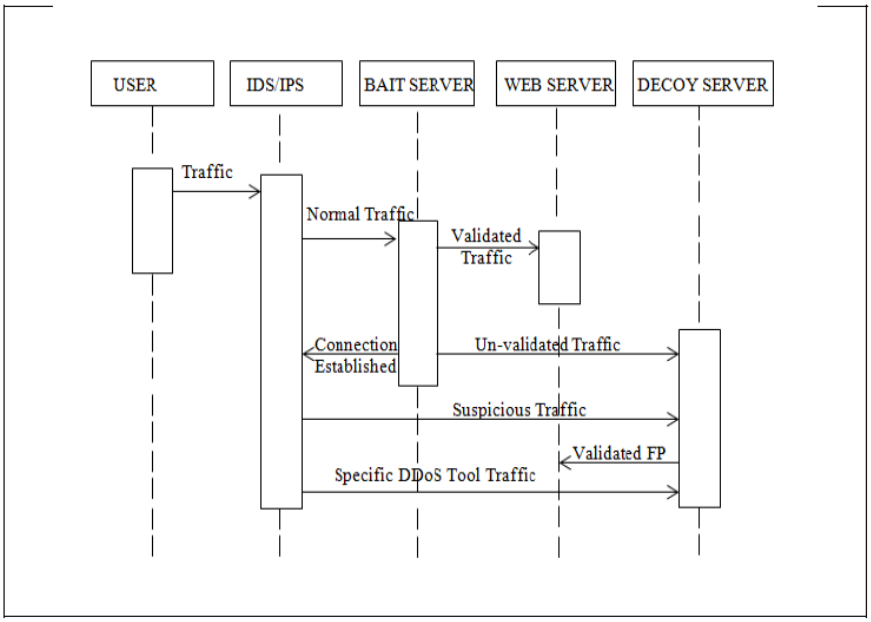



FIGURE 3.6 – Mitigation of DDoS attack traffic sequence.[48]

For this classification use function where x denotes set of possible instances and y denotes set of possible labels. $f : x \rightarrow y$

Where function hypotheses set is $H = \{h|h : x \rightarrow y\}$ which gives the function $f$ training by different data (x,y) to measure impurity degree use entropy-based metric, takes labeled data X and Y as an input for train the classifier and to devlop model the prediction, depends on new data x for output best approximate. As illustrated in the following[48](Figure(3.7)).

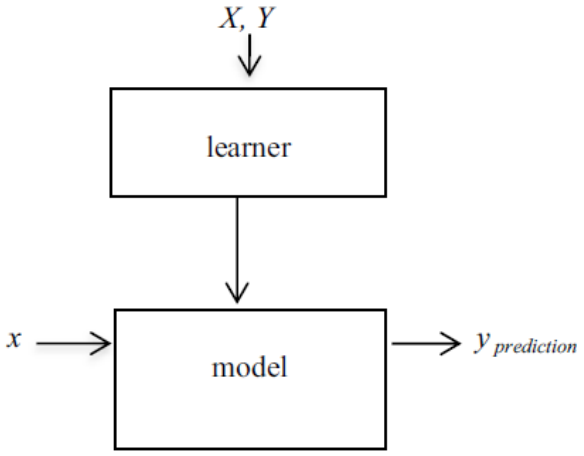



FIGURE 3.7 – Classification model for traffic classes prediction[48].

### 3.3.4 Detection DDoS Attack Using C5.0 Machine Learning Algorithm

Decision Tree algorithm (ID3, C4.5, C5.0). DDoS attack situations to detect it used C5.0 classifier decision trees is better than C4.5 classifiers[25]. taking very less time in the implementation compared to C4.5 classifiers. and gives better accuracy in classification. the C5.0 classifier to detect an attack when it happens, not after it has occurred.and is a best solution for the threats posed by Distributed Denial of Service attacks[25] (Figure(3.8)).

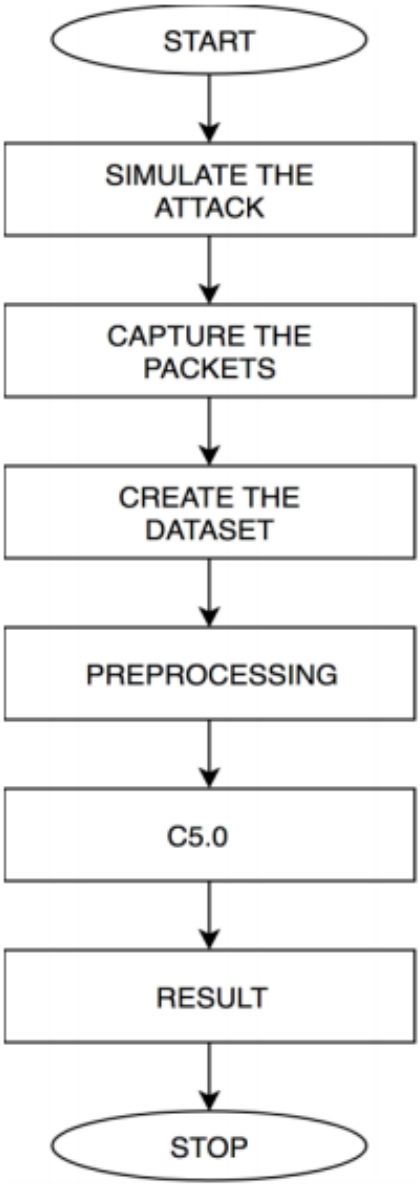



FIGURE 3.8 – Detection Using Decision Tree C5.0 [25]

.

**Analysis of C5.0 :** The following algorithm are explained for each Decision Tree algorithm C5.0[56].

**Algorithm 8** Algorithm to generate C5.0 decision tree. [56]

| | |
|---|---|
| 1 | create a node N |
| 2 | **if** tuples in D are all of the same class, C, **then** |
| 3 | **return** N as a leaf node labelled with the class C |
| 4 | **if** attribute list is empty, **then** |
| 5 | **return** N as a leaf node labelled with the majority class in D |
| 6 | apply attribute selection method(D, attribute list) to find the best splitting criterion |
| 7 | label node N with splitting criterion |
| 8 | **if** splitting attribute is discrete-valued and multi way splits allowed **then** |
| 9 | attribute list ←attribute list- splitting attribute |
| 10 | **For** each outcome j of splitting criterion<br>Let$D_j$ be the set of data tuples in D satisfying outcome j<br>**if** $D_i$ is empty then attach a leaf labelled with majority class in D to node N<br>**else**, attach the node returned by Generate C5.0 decision<br>tree($D_j$, attribute list) to node N |
| 11 | **return N** |

### 3.3.5 Detection of DDoS Attacks using K-NN Classifier

When an attack occurs at the network level, the attacker sends packets to find loopholes to launch a random attack whose address is fixed and is often changed, and when this happens, Lee (2007) suggest a way to identify is :

$$H = \sum_{i=1}^{n} P_i log_2 P_i \text{ [49].}$$

First, we select the features for detecting. Second, we consider the classification of the current network status to one of the classes. There are many well-known methods for classifying documents such as SVM, NN, fuzzy logic, but this situation choose the K-NN method.

After giving K a value for study, for example, we give $k = 5$ , which combines 5 elements of the study fee and begins with two elements as the first study[49].

For the 2 elements $X = x_1, x_2, \ldots, x_n$ and $Y = y_1, y_2, \ldots, y_n$,

$W = w_1, w_2, \ldots, w_n$ is the weighted vector and $w_i$ is the weight of the component i in the general vector.

Then examine the similarities between X and Y.

Similarity(X, Y) = $Cosine(X, Y, W) = \sum_{i=1}^{n} (x_i \times w_i) \times (y_i \times w_i) \,/\, \sqrt{\sum_{i=1}^{n} (x_i \times w_i)^2} \; \sqrt{\sum_{i=1}^{n} (y_i \times w_i)^2}$

Detection of DDoS Attacks Using k-NN Classifier proposed by (Lee et al. (2007))[49]. With normalization, variables become. $z\frac{x-\bar{x}}{\sigma}$ Where x , $\bar{x}$ ,$\sigma$, denotes the value of each feature(Figure(3.9)).

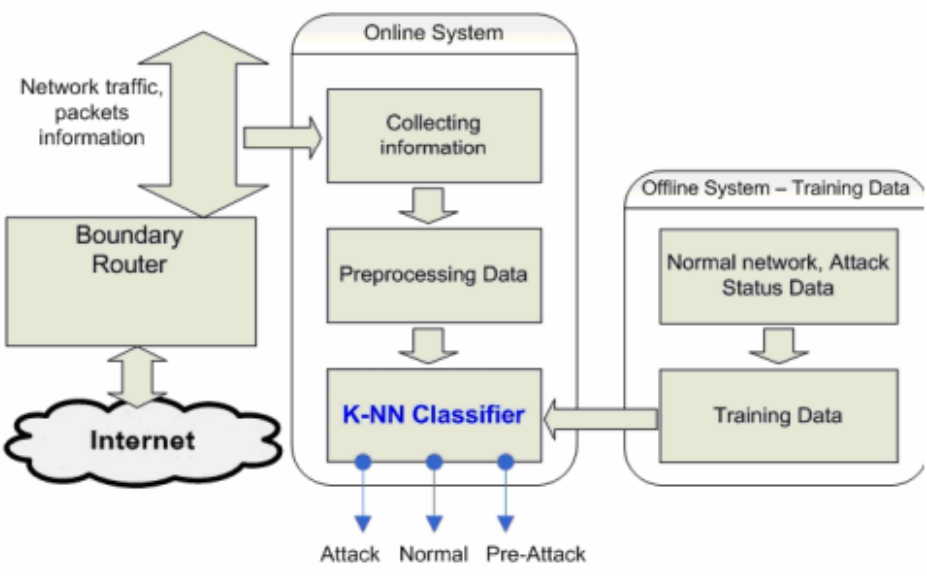



FIGURE 3.9 – General model for detecting precursor of DDoS attacks.[49]

When an attack occurs, classified the data into three groups that are being studied and evaluated [49] :

— Normal class : $N^1, ...., N^L$
— Pre-attack class : $P^1, ...., P^L$
— Attack class : $R^1, ...., r^L$

Where Normal class it is a normal status of network, and Pre-attack class have it two phases :

— Phase 1 of DDoS attack—selection of master and botnet.
— Phase 2 of DDoS attack—communication and compromise .

Last one attack class this it is problem and which needs to solve them or Remove it(Figure(3.10)).

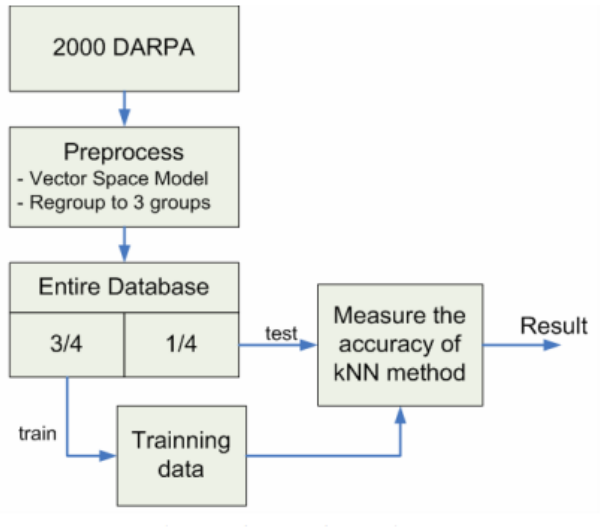



FIGURE 3.10 – Scheme of experiment [49]

### 3.3.6 Detection of DDoS Attacks using RNN algorithm

The spread of DDoS attacks in all ways is for its main purpose, but researchers are proposing algorithms to solve this problem. Therefore, one of these algorithms is Recurrent Neural Network (RNN), which has many features categorized from the algorithm to help this attack. One of its advantages is that when a flow occurs, it can detect flood and its nature if it is strong or slow, so that it can reveal the time of flow and finally, it is the most fundamental attack on the Web[40].

In order to identify DDoS attacks, they use the PCA application, reducing the RNN dimension for data training and obtaining a detection result without difficulty. So researchers developed this work takes a new term which is PCA-RNN, which helps to achieve accuracy[40].
The following figure provides a detection model to attack, which passes through 3 basic phases(Figure(Figure3.11)).

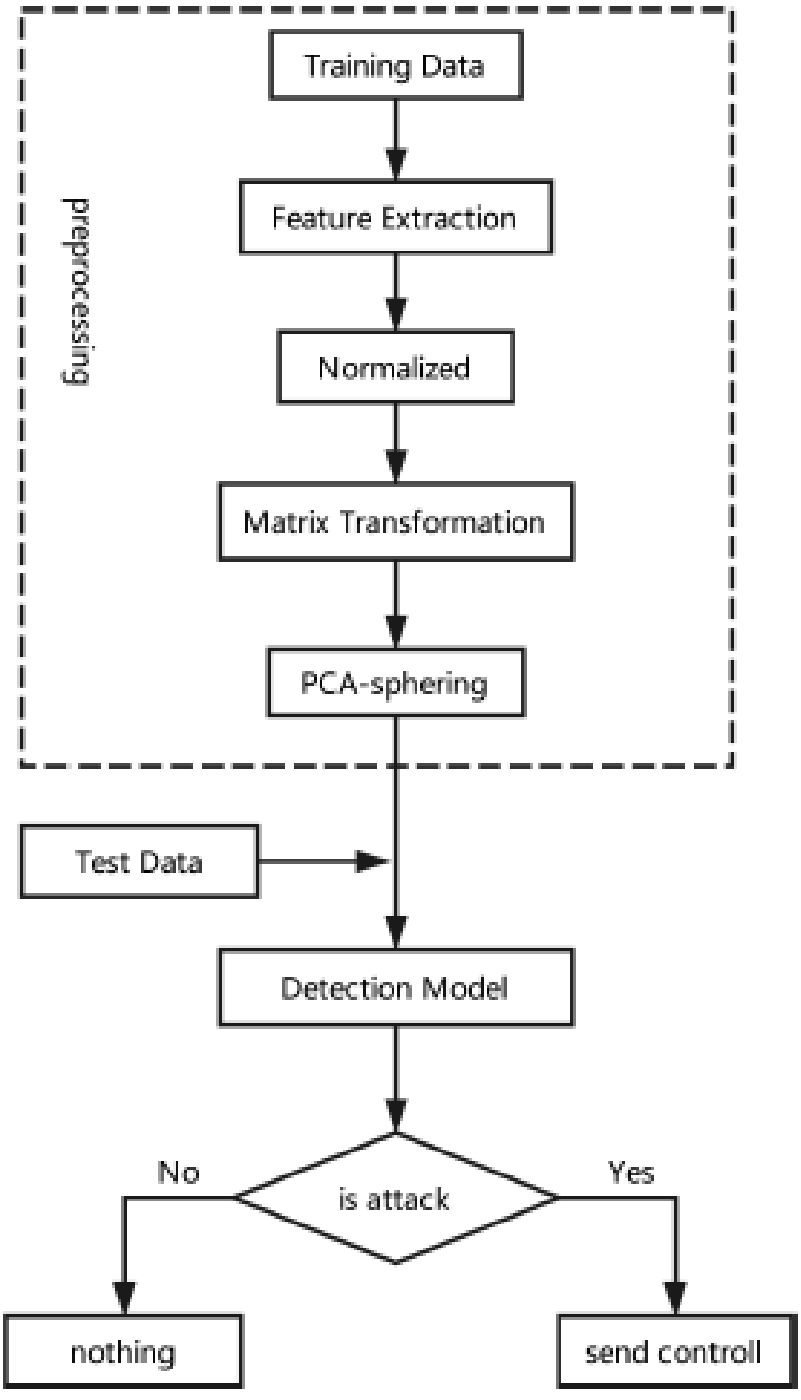



FIGURE 3.11 – Detection model by PCA-RNN [40]

## 3.4 Application Layer Mitigation Techniques

### 3.4.1 Application Layer :

The application layer is a complement to all layer in model TCP/IP . Where its work is very important in the field of communication, it is the means that controls the protocols in order to do their work. With all the positives provided by this class, except that they are exposed to attacks that lose some of their characteristics. In the application layer, there are many protocols to which they belong, including HTTP. They are used in web networks and it is the basis for reviewing a page on a website(3.12).

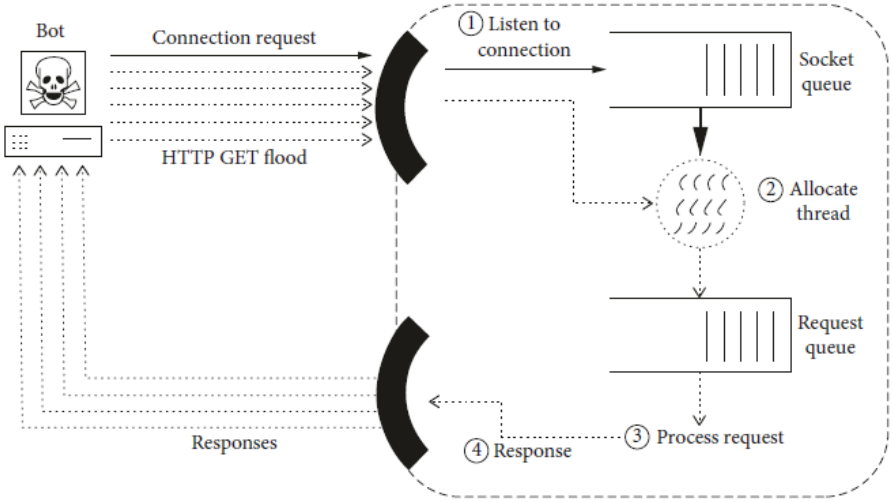



FIGURE 3.12 – Web server architecture[34]

### 3.4.2 Application layer HTTP :

HTTP means HyperText Transfer Protocol used by the World Wide Web. [62] is working on network request-response where client initiates a request by establishing a Transmission Control Protocol (TCP) connection to a particular port 80 for example. The HTTP server waits for the client request message Which receives and is a code and request by a particular method. The server uses two versions : HTTP 1.1 , HTTP 0.1 Each one has its own characteristics.

**The methods for transmission :** GET, POST, CONNECT, HEAD, etc..

| Methode | Difinition |
|---|---|
| GET | demand the content of the resource like file or song or movie. |
| POST | provide data to the resource like php for example wikipedia. |
| CONNECT | Contact through an agent ( proxy). |
| HEAD | just demand the header of the resource. |
| **PUT, DELETE, TRACE** | |

Table 3.1 – Methods Requests HTTP

**Codes Requests :** Are the response codes when HTTP, it is grouped in five classes. Each class has a particular response and the most serious is class 5 An example is a formula : class 2 for example 200 and 201 etc.. with all class

| Codes requets | Difinition |
|---|---|
| 200 | successful operation |
| 201 | sent to the following of a POST when the relevant document has been established (Created). |
| 301 | the document claims to have been moved and is now at another address mentioned in the reply. |
| 304 | response to a GET request when the server does not return a document because there is no new version of it. |
| 400 | syntax error : the server did not understand the request. |
| 404 | the server could not find the request document. |

Table 3.2 – Codes requets

**Example on request GET** Method or comment used in this example is GET, by Protocol HTTP 1.1 to Request WWW.example.com and where replied ok by code 200 'Accept-Encoding :gzip' (3.13).

```
>> Request:

   GET /index HTTP/1.1
   Host: www.example.com
   Accept-Encoding: gzip
```

FIGURE 3.13 – Example request

### 3.4.3 Application Layer Attacks

**HTTP flood :**

It happens through the sites on the server of that institution, thus, we studied around us this attack on the site of the University of Ghardaia. Where a flood occurred on the memory of the server and stopped it for a while until it is mitigate.
According to the study, there was a recent attack at the GET level,

The exploitation of the protocol in the application layer by hackers or attackers on two protocols HTTP, File Transfer Protocol (FTP)[34] happening of GET flood by sending a large amount of GET request at one time and multiple sources to one server (Ni et al. )[34].

**DDoS Attack Strategies at Application Layer**

DDoS attack strategies, multiple of which : proxy.
It use proxy server at the for it's advantage in converting orders to another URL, in order availability many HTTP orders instead of the web server.[34]

**Problem :** With all this explanation, we find a problem in the traffic from a server where it can not connect to the network because of the flood, it occurs on a GET request.

The following figure illustrates the structure of happening process the request going through several stages to reach the desired :(Figure(3.14))

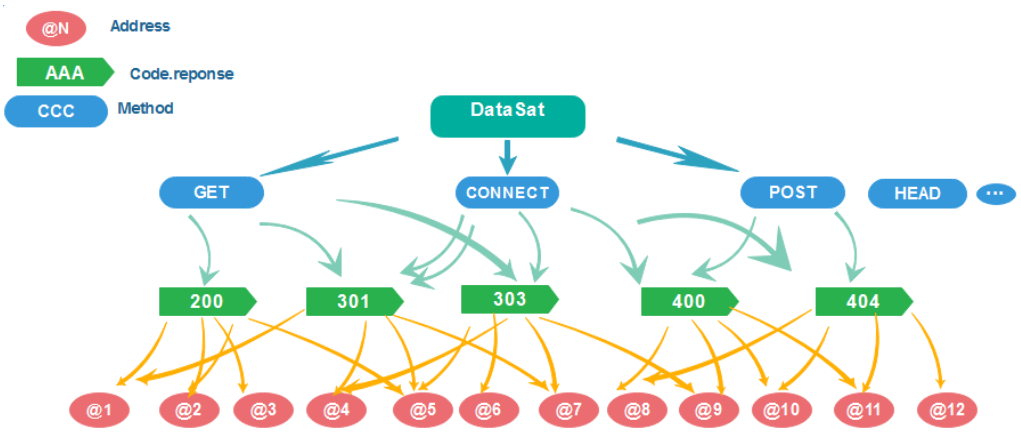



FIGURE 3.14 – Structuring attack in site Ghardaia

**Phases :** To solve this there are two steps first, detect reason of the attack and the server crashes and secondly, mitigation it.

## 3.5 Implementation

**Software tools and Technologies :** Operating Sestem : Windows

**language R :**

**R Programming :** Is a statistical program that helps to study data accurately and give numerical statistical results, it also has many features, such as drawing curves and columns. Provides an open source software environment for classification and clustering data[22]. It also helps in data analysis, in order to do its work must user must install packages as needed[14]. Its use is simple and does not require any complexity and installe on any operating system (Linux or Windows and Macintosh). Some the domains related which uses it :

- Machine Learning & Statistical Learning.
- Cluster Analysis & Finite Mixture Models.
- Time Series Analysis.
- Analysis of Spatial Data.

**Objectives**

- To develop the model to quantify the risk in the system.

— To discover the risk based on the incident reports.

The inclusion of information about the unexpected events in the system helps to discover risk for the unseen event

**1.RStudio :** Is the interface to write instructions for implementation(Figure(3.15)).

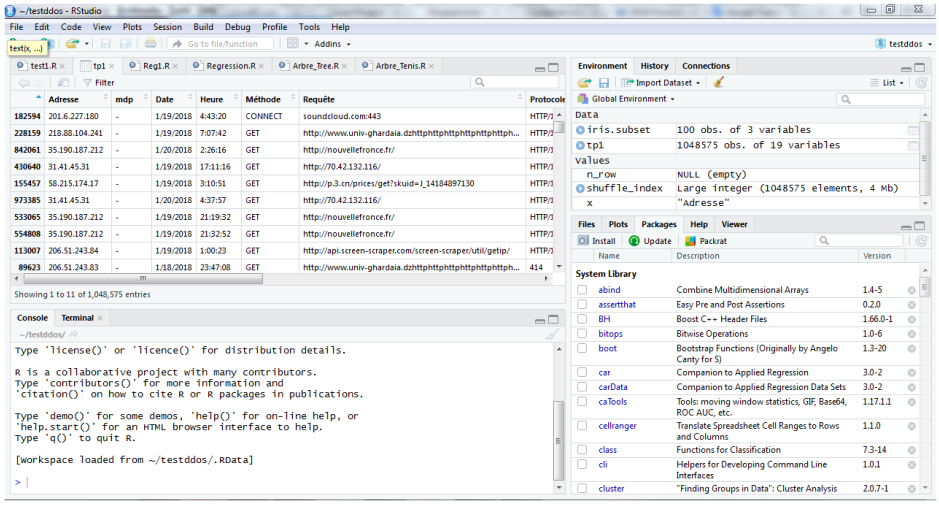



FIGURE 3.15 – Program for implementation

**2.Packages :**

```
#To apply a clustering analysis we need to download the following packages

install.packages("cluster")    # clustering algorithms
install.packages("factoextra") # clustering algorithms & visualization
install.packages("fpc")
install.packages("flexclust")
install.packages("ade4")
```

FIGURE 3.16 – Import Packages

— library("cluster")
— library("ade4")
— library("fpc")
— library("factoextra")
— library("flexclust")

**Dataset :**

Import CSV files 1,048,575 entries but Showing 1 to 14 of 1,048,575 entries.

**Analyse of address :** Teh Address that send to the University of Ghardaia are devided into good addresses and also Malicious addresses (botnet) . Botnet here it is They are addresses that are sent frequently of a fixed period of time in 3 days (Figure(3.17)).

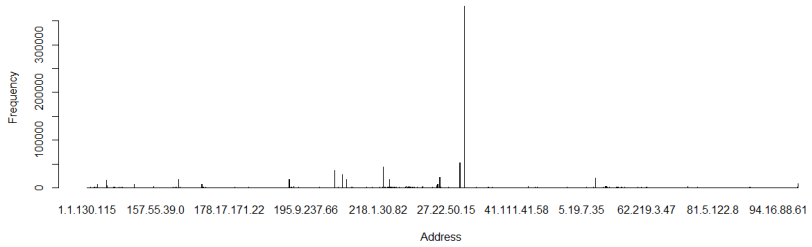



FIGURE 3.17 – Analyse of addresse

**Frequente of method Upon request :** The most used method to make request is GET as it's illustrated by figure (Figure(3.18))

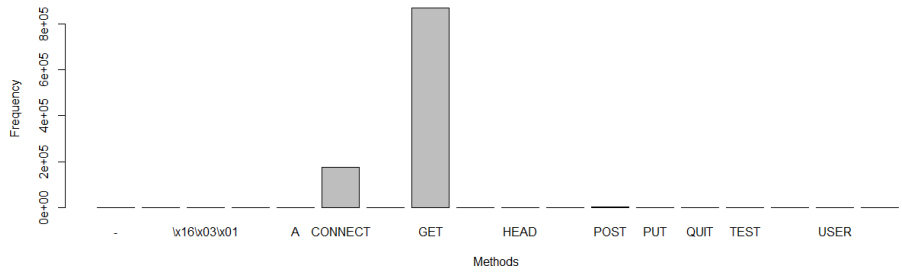



FIGURE 3.18 – Frequente of method Upon request

**Analyse size reponse :** When the user sends a request on the site, each request has a certain size, and the frequent sending of orders, this causes a flood on the level of server memory The following figure illustrate this (Figure(3.19)).

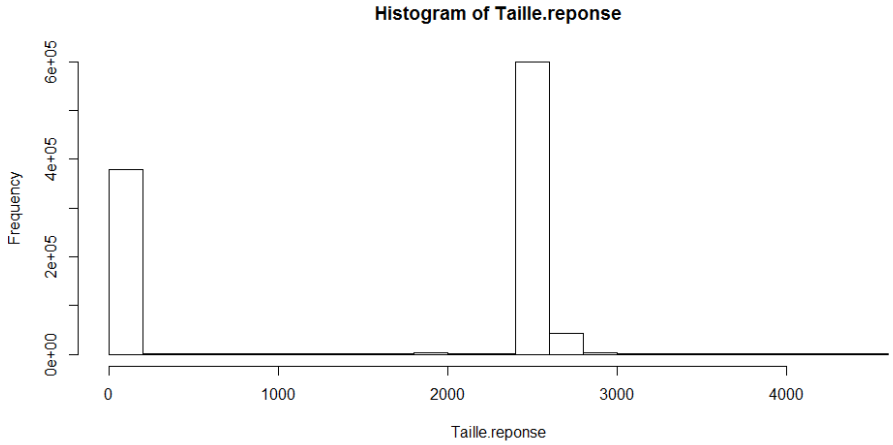



FIGURE 3.19 – Analyse size reponse

**Hours of Attack :** The attack lasted three consecutive days and the second day was heavily attacked in a limited time period as shown in the figure below (Figure(3.20))

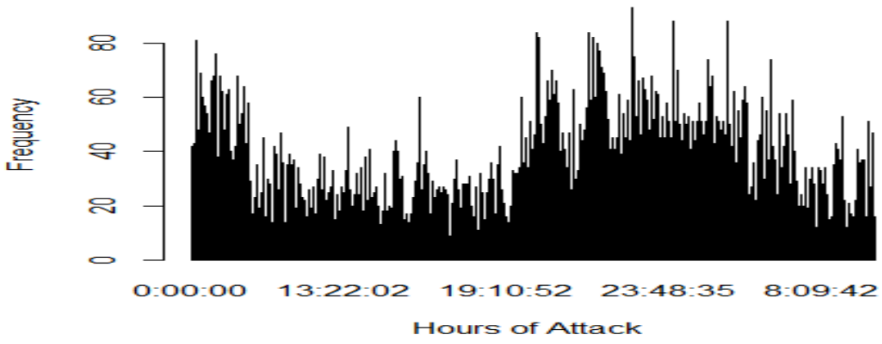



FIGURE 3.20 – Hours of Attack

### 3.5.1 Feature Reduction Using PCA

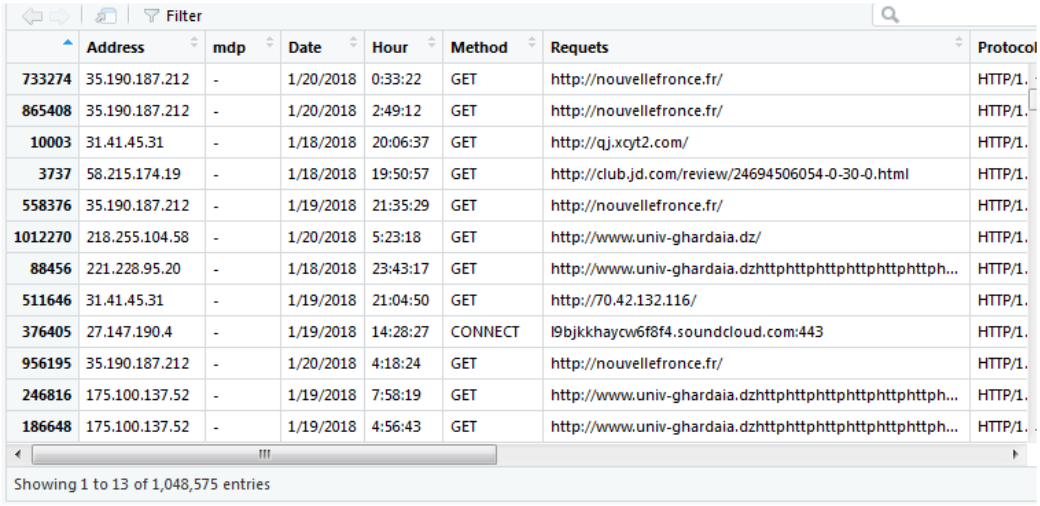
In order to study the data we had to use the Principal Component Analysis (PCA) in order to reduce the dimensions with the knowledge that it retains value as before.

### 3.5.2 Training of dataset by Kmeans

To build cluster by Kmeans, we will proceed as follow :

1. Step 1 : Import the data
2. Step 2 : Clean the dataset
3. Step 2.1 : Remove any missing value (i.e, NA values for not available)
4. Step 2.2 : Replacement space to mean in column Response size (taill reponse)
5. Step 2.3 : Delete redundant and unnecessary columns.
6. Step 3 : Convert type (attribute Method) and (Taille.repons) factor to numeric with dataset
7. Step 4 : scaling / standardization of data
8. Step 5 : Reduce dimensions
9. Step 6 :Choose the number of clusters(K)
10. Step 6 : K-Means Cluster Analysis
11. Step 6.1 : Choose k=2 clusters solution
12. Step 6.2 : Get the averages (mean) of the clusters
13. Step 6.3 : Add the assignment (affectation ) to each cluster
14. Step 6.4 : Improt library cluster
15. Step 7 : Cluster plot

**Step 1 : Import the data :**



Filter

| | Address | mdp | Date | Hour | Method | Requets | Protocol |
|---|---|---|---|---|---|---|---|
| 733274 | 35.190.187.212 | - | 1/20/2018 | 0:33:22 | GET | http://nouvellefronce.fr/ | HTTP/1. |
| 865408 | 35.190.187.212 | - | 1/20/2018 | 2:49:12 | GET | http://nouvellefronce.fr/ | HTTP/1. |
| 10003 | 31.41.45.31 | - | 1/18/2018 | 20:06:37 | GET | http://qj.xcyt2.com/ | HTTP/1. |
| 3737 | 58.215.174.19 | - | 1/18/2018 | 19:50:57 | GET | http://club.jd.com/review/24694506054-0-30-0.html | HTTP/1. |
| 558376 | 35.190.187.212 | - | 1/19/2018 | 21:35:29 | GET | http://nouvellefronce.fr/ | HTTP/1. |
| 1012270 | 218.255.104.58 | - | 1/20/2018 | 5:23:18 | GET | http://www.univ-ghardaia.dz/ | HTTP/1. |
| 88456 | 221.228.95.20 | - | 1/18/2018 | 23:43:17 | GET | http://www.univ-ghardaia.dzhttphttphttphttphttphttph... | HTTP/1. |
| 511646 | 31.41.45.31 | - | 1/19/2018 | 21:04:50 | GET | http://70.42.132.116/ | HTTP/1. |
| 376405 | 27.147.190.4 | - | 1/19/2018 | 14:28:27 | CONNECT | l9bjkkhaycw6f8f4.soundcloud.com:443 | HTTP/1. |
| 956195 | 35.190.187.212 | - | 1/20/2018 | 4:18:24 | GET | http://nouvellefronce.fr/ | HTTP/1. |
| 246816 | 175.100.137.52 | - | 1/19/2018 | 7:58:19 | GET | http://www.univ-ghardaia.dzhttphttphttphttphttphttph... | HTTP/1. |
| 186648 | 175.100.137.52 | - | 1/19/2018 | 4:56:43 | GET | http://www.univ-ghardaia.dzhttphttphttphttphttphttph... | HTTP/1. |

Showing 1 to 13 of 1,048,575 entries

FIGURE 3.21 – Import the data

**Step 2 : Clean the dataset :** Data cleaning is an important stage before starting to implement, in order to facilitate and not fall into trouble to study, to implement this instruction we apply this code (Figure(3.22)).

```
#cleaning dataset: cleaning dataset
AttackDDos[,9] <- as.numeric(as.character(AttackDDos[,9]))

# step1 :Remove any missing value (i.e, NA values for not available)

#1. replacement space to mean
AttackDDos[is.na(AttackDDos[,9]), 9] <- mean(AttackDDos[,9], na.rm = TRUE)
# 2.delet col
AttackDDos$mdp  <- NULL
AttackDDos$X <- NULL
AttackDDos$X.1 <- NULL
AttackDDos$X.2 <- NULL
AttackDDos$X.3 <- NULL
AttackDDos$X.4 <- NULL
AttackDDos$X.5 <- NULL
AttackDDos$X.6 <- NULL
AttackDDos$X.7 <- NULL
AttackDDos$X.8 <- NULL
AttackDDos$X.9 <- NULL

#3.delete missing data
newdata <- na.omit(AttackDDos)
Method <-na.omit(Method)
```

FIGURE 3.22 – Cleaning dataset

**Step 3 : Convert type (attribute Method) and (Taille.repons) factor to numeric with dataset :** The purpose of data conversion type, is to study the data we consider each input as the index, to implement this instruction we apply this code (Figure(3.23)).

```
#Step2: convert type attribut Method and Taille.repons factor to numeric
Method<-as.numeric(Method)
Taille.reponse<-as.numeric(Taille.reponse)
Address<-as.numeric(Address)
Hour<-as.numeric(Hour)
Date<-as.numeric(Date)
Requets<-as.numeric(Requets)
Protocol<-as.numeric(Protocol)
Code.reponse<-as.numeric((Code.reponse))
```

FIGURE 3.23 – Convert type

**Step 4 : scaling / standardization of data :** Scale : a logical value. If TRUE, the data is scaled to the unit variance before the analysis. This standardization of scale prevents certain variables from becoming dominant simply because of their large units of measurement, to implement this instruction we apply this code (Figure(3.24)).

```
#scaling / standardization of data


Method<- scale(Method)
Taille.reponse<-scale(Taille.reponse)

plot(Method,Taille.reponse)
with(AttackDDos,text(Method,Taille.reponse,labels = Address))
```

FIGURE 3.24 – scaling / standardization of data

**Step 6 :Choose the number of clusters(K)** Number of clusters to Methed (Figure(3.25)).

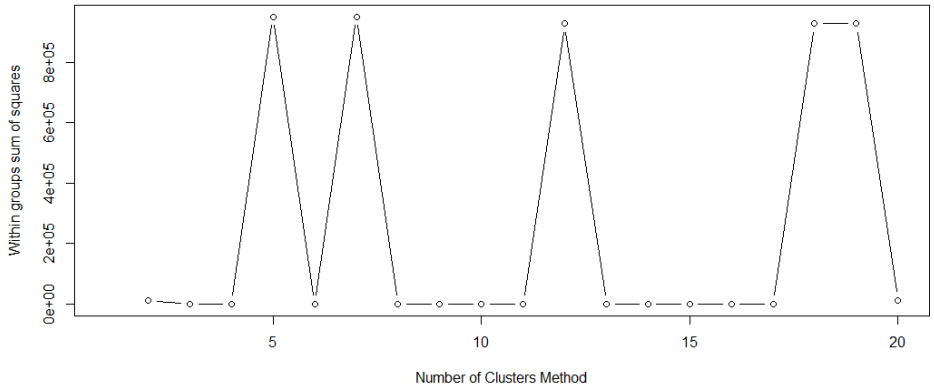



FIGURE 3.25 – Choose the number of clusters(K)

**Resultat :** Cluster the data according to the type of method used for request, where we observe intensification on 2 cluster, indicating that there is pressure in the frequency and convergence of data. According to a study, we deduce the frequently used method of GET, The following figure illustrates this (Figure(3.26)).

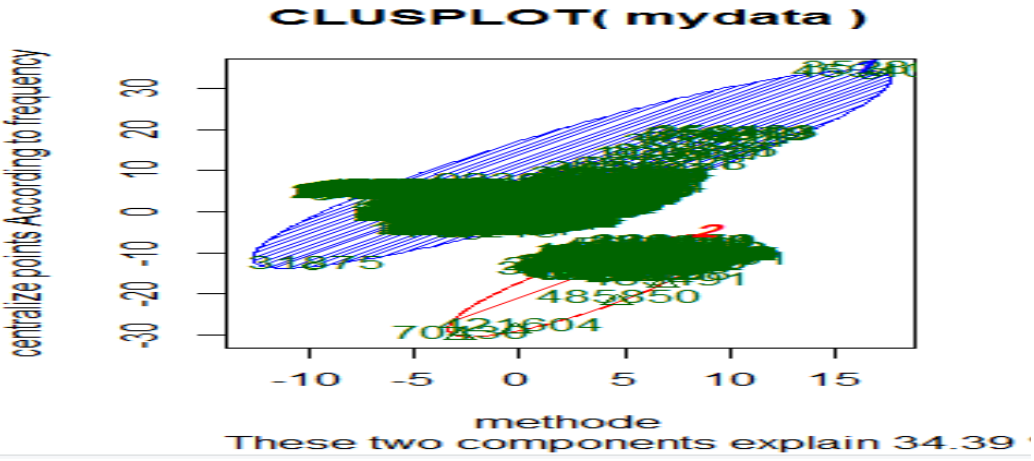



FIGURE 3.26 – Resulatat of clusters

## 3.6 Conclusion

The use of automated learning in the study and detection of DDoS attack was different depending on the data being studied and the type of protocol used to launch the attack. There are several ML techniques such as Bayes, SVM, Decision Tree (ID3, C4.5, C5.0), and K-Nearest Neighbor Classifier Each of these techniques addresses the different aspects of the data set and provides accurate classifications, the cluster technique was also applied when data were unlabel.

In our study the data were not studied before, so we could not know the cause of this flood, we proposed the Kmeans algorithm for groups and we deduce the cause of this flood, which is GET Flood. Where the moderator caused by GET Flood is that the whole site contains many thagarat that allow the attacker to intrude and control the server system. As well as using proxy server to block, but because of these gaps led to control their prexy server and sent many of the phantom orders in order to disable it. Thus, This has led to the appearance of the 301 code frequently, this is considered unusual in sites . In order to study the data concluded 34% of the error percentage in classification.

# General Conclusion

In this work, we have made it clear that the machine learning characteristics of packets received by the server can reveal a DDoS attack and accurately characterize the traffic and type of protocol that caused it. We used a limited set of features to study data, resulting in many factors at University of Ghardaia used proxy server for special purposes, and the strategy that caused the DDoS attack.

Where the attacker took advantage of the gaps he had and increased the proportion of requests for flooding. Thus, the Kmeans algorithm has helped us a lot in detecting the symptoms and deduce the source of the attack. As for mitigation, it has never been applied to machine learning, often using other methods to protect against attack.

Therefore, in this subject access to mitigation on the attack using machine learning, is not specified for the reason that each institution exposed to this attack is competent in the process of mitigation and the selection of stages for this.

# Bibliographie


[1] Mohiuddin Ahmed. Collective anomaly detection techniques for network traffic analysis. *Annals of Data Science*, pages 1–16, 2018.

[2] Esraa Alomari, Selvakumar Manickam, BB Gupta, Shankar Karuppayah, and Rafeef Alfaris. Botnet-based distributed denial of service (ddos) attacks on web servers : classification and art. *arXiv preprint arXiv :1208.0403*, 2012.

[3] Bijalwan Anchit and Singh Harvinder. Investigation of udp bot flooding attack. *Indian Journal of Science and Technology*, 9 :21, 2016.

[4] Ketki Arora, Krishan Kumar, and Monika Sachdeva. Impact analysis of recent ddos attacks. *International Journal on Computer Science and Engineering*, 3(2) :877–883, 2011.

[5] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. In *New advances in machine learning*. IntechOpen, 2010.

[6] Jarrod Bakker. Intelligent traffic classification for detecting ddos attacks using sdn/openflow. 2017.

[7] R Bani-Hani and Zaid Al-Ali. Syn flooding attacks and countermeasures : a survey. In *Proceedings of ICICS*, 2013.

[8] Sunny Behal and Krishan Kumar. Characterization and comparison of ddos attack tools and traffic generators : A review. *IJ Network Security*, 19(3) :383–393, 2017.

[9] Hakem Beitollahi and Geert Deconinck. Analyzing well-known countermeasures against distributed denial of service attacks. *Computer Communications*, 35(11) :1312–1332, 2012.

[10] Monowar H Bhuyan, Hirak Jyoti Kashyap, Dhruba Kumar Bhattacharyya, and Jugal K Kalita. Detecting distributed denial of service attacks : methods, tools and future directions. *The Computer Journal*, 57(4) :537–556, 2013.

[11] callum williams. *The effect of UDP Flood attack on QoS of VoIP and Video Conferencing Services*. SS1628225. 2018.

[12] Wei-Lun Chao. Machine learning tutorial. *National Taiwan University*, 2011.

[13] Alejandro Cholaquidis, Ricardo Fraimand, and Mariela Sued. Semi-supervised learning : When and why it works. *arXiv preprint arXiv :1805.09180*, 2018.

[14] Pierre Lafaye De Micheaux, Rémy Drouilhet, and Benoît Liquet. *Le logiciel R : Maîtriser le langage-Effectuer des analyses statistiques*. Springer Science & Business Media, 2011.

[15] BS Kiruthika Devi, G Preetha, G Selvaram, and S Mercy Shalinie. An impact analysis : Real time ddos attack detection and mitigation using machine learning. In *2014 International Conference on Recent Trends in Information Technology*, pages 1–7. IEEE, 2014.

[16] Wesley M Eddy. Defenses against tcp syn flooding attacks. *The Internet Protocol Journal*, 9(4) :2–16, 2006.

[17] Khaled M Elleithy, Drazen Blagovic, Wang K Cheng, and Paul Sideleau. Denial of service attack techniques : Analysis, implementation and comparison. 2005.

[18] Pierre-Edouard Fabre. *Using network resources to mitigate volumetric DDoS*. PhD thesis, Institut National des Télécommunications, 2018.

[19] Carol J Fung and Bill McCormick. Vguard : A distributed denial of service attack mitigation method using network function virtualization. In *2015 11th International Conference on Network and Service Management (CNSM)*, pages 64–70. IEEE, 2015.

[20] Arpit Ramesh Gawande. *DDoS detection and mitigation using machine learning*. PhD thesis, Rutgers University-Camden Graduate School, 2018.

[21] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1) :3–42, 2006.

[22] EC Grunsky. R : a data analysis and statistical programming environment–an emerging tool for the geosciences. *Computers & Geosciences*, 28(10) :1219–1222, 2002.

[23] BB Gupta, Ramesh Chandra Joshi, and Manoj Misra. Distributed denial of service prevention techniques. *arXiv preprint arXiv :1208.3557*, 2012.

[24] Sergio Armando Gutiérrez and John Willian Branch. Application of machine learning techniques to distributed denial of service (ddos) attack detection : A systematic literature review.

[25] Abhishek H. K Hariharan. M and B. G. Prasad. *DDoS Attack Detection Using C5.0 Machine Learning Algorithm*. Published Online January

2019 in MECS(http ://www.mecs-press.net), January 2019. Department of CSE, BMS College of Engineering, Bengaluru, Karnataka - 560019, India Department of CSE, BMS College of Engineering, Bengaluru, Karnataka - 560019, India.

[26] SHC Haris, RB Ahmad, MAHA Ghani, and Ghossoon M Waleed. Tcp syn flood detection based on payload analysis. In *2010 IEEE Student Conference on Research and Development (SCOReD)*, pages 149–153. IEEE, 2010.

[27] Simon S Haykin, Simon S Haykin, Simon S Haykin, Kanada Elektroingenieur, and Simon S Haykin. *Neural networks and learning machines*, volume 3. Pearson education Upper Saddle River, 2009.

[28] Zecheng He, Tianwei Zhang, and Ruby B Lee. Machine learning based ddos attack detection from source side in cloud. In *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*, pages 114–120. IEEE, 2017.

[29] David Holmes. Mitigating ddos attacks with f5 technology. *F5 Networks, Inc*, pages 2099–2104, 2013.

[30] Fatima Hussain, Rasheed Hussain, Syed Ali Hassan, and Ekram Hossain. Machine learning in iot security : Current solutions and future challenges. *arXiv preprint arXiv :1904.05735*, 2019.

[31] Mohamed Idhammad, Karim Afdel, and Mustapha Belouch. Semi-supervised machine learning approach for ddos detection. *Applied Intelligence*, 48(10) :3193–3208, 2018.

[32] John Ioannidis and Steven Michael Bellovin. Implementing pushback : Router-based defense against ddos attacks. 2002.

[33] N Ch SN Iyengar, Arindam Banerjee, and Gopinath Ganapathy. A fuzzy logic based defense mechanism against distributed denial of service attack in cloud computing environment. *International journal of communication networks and Information security*, 6(3) :233, 2014.

[34] Ghafar A Jaafar, Shahidan M Abdullah, and Saifuladli Ismail. Review of recent detection methods for http ddos attack. *Journal of Computer Networks and Communications*, 2019, 2019.

[35] Ah Reum Kang and Aziz Mohaisen. Automatic alerts annotation for improving ddos mitigation systems. In *2016 IEEE Conference on Communications and Network Security (CNS)*, pages 362–363. IEEE, 2016.

[36] Ronen Kenig, Deborah Manor, Ziv Gadot, and Daniel Trauner. Ddos survival handbook, 2013.

[37] Paul R Kersten, Jong-Sen Lee, and Thomas L Ainsworth. Unsupervised classification of polarimetric synthetic aperture radar images using fuzzy clustering and em clustering. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3) :519–527, 2005.

[38] Hoda Khanali and Babak Vaziri. A survey on clustering algorithms for partitioning method. *Int J Comput Appl*, 155(4) :20–25, 2016.

[39] Sanjeev Kumar and Raja Sekhar Reddy Gade. Experimental evaluation of juniper network's netscreen-5gt security device against layer4 flood attacks. *Journal of Information Security*, 2(01) :50, 2011.

[40] Qian Li, Linhai Meng, Yuan Zhang, and Jinyao Yan. Ddos attacks detection using machine learning algorithms. In *International Forum on Digital TV and Wireless Multimedia Communications*, pages 205–216. Springer, 2018.

[41] Matthew V Mahoney and Philip K Chan. Phad : Packet header anomaly detection for identifying hostile network traffic. Technical report, 2001.

[42] Monika Malik and Yudhvir Singh. A review : Dos and ddos attacks. *International Journal of Computer Science and Mobile Computing*, 4(6) :260–265, 2015.

[43] Larry M Manevitz and Malik Yousef. One-class svms for document classification. *Journal of machine Learning research*, 2(Dec) :139–154, 2001.

[44] Larry M Manevitz and Malik Yousef. One-class svms for document classification. *Journal of machine Learning research*, 2(Dec) :139–154, 2001.

[45] Bernard Marr. A short history of machine learning—every manager should read. *Forbes. http ://tinyurl. com/gslvr6k*, 2016.

[46] Tom Michael Mitchell. *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, Machine Learning . . . , 2006.

[47] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning : algorithms and applications*. Crc Press, 2016.

[48] Jema David Ndibwile, A Govardhan, Kazuya Okada, and Youki Kadobayashi. Web server protection against application layer ddos attacks using machine learning and traffic authentication. In *2015 IEEE 39th Annual Computer Software and Applications Conference*, volume 3, pages 261–267. IEEE, 2015.

[49] Hoai-Vu Nguyen and Yongsun Choi. Proactive detection of ddos attacks utilizing k-nn classifier in an anti-ddos framework. *International Journal of Electrical, Computer, and Systems Engineering*, 4(4) :247–252, 2010.

[50] Raja Azrina Raja Othman. Understanding the various types of denial of service attack. *Business Week Online*, 2000.

[51] Sinno Jialin Pan, James T Kwok, Qiang Yang, et al. Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682, 2008.

[52] Tao Peng, Christopher Leckie, and Kotagiri Ramamohanarao. Survey of network-based defense mechanisms countering the dos and ddos problems. *ACM Computing Surveys (CSUR)*, 39(1) :3, 2007.

[53] Tuomo Penttinen. Distributed denial-of-service attacks in the internet, 2005.

[54] R Porkodi and V Bhuvaneswari. The internet of things (iot) applications and communication enabling technology standards : An overview. In *2014 International Conference on Intelligent Computing Applications*, pages 324–329. IEEE, 2014.

[55] Bahman Rashidi, Carol Fung, and Elisa Bertino. A collaborative ddos defence framework using network function virtualization. *IEEE Transactions on Information Forensics and Security*, 12(10) :2483–2497, 2017.

[56] R Revathy and R Lawrance. Comparative analysis of c4. 5 and c5. 0 algorithms on crop pest data. *Int. J. Innov. Res. Comput. Commun. Eng*, 5(1) :50–58, 2017.

[57] Eric Rosen, Arun Viswanathan, and Ross Callon. Multiprotocol label switching architecture. Technical report, 2000.

[58] Idir Sadaoui, Lamia Hamza, et al. *Les attaques par déni de service distribué dans les systèmes informatiques*. PhD thesis, Universite de bejaia, 2017.

[59] Nilaykumar Kiran Sangani and Haroot Zarger. Machine learning in application security. In *Advances in Security in Computing and Communications*. IntechOpen, 2017.

[60] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning : From theory to algorithms*. Cambridge university press, 2014.

[61] Niharika Sharma, Amit Mahajan, and V Malhotra. Machine learning techniques used in detection of dos attacks : A literature review. *International Journal of Advance Research in Computer Science and Software Engineering*, 6(3) :100–105, 2016.

[62] Karanpreet Singh, Paramvir Singh, and Krishan Kumar. Application layer http-get flood ddos attacks : Research landscape and challenges. *Computers & security*, 65 :344–372, 2017.

[63] Alex Smola and SVN Vishwanathan. Introduction to machine learning. *Cambridge University, UK*, 32 :34, 2008.

[64] Theodoros Spyridopoulos, G Karanikas, Theodore Tryfonas, and Georgios Oikonomou. A game theoretic defence framework against dos/ddos cyber attacks. *Computers & Security*, 38 :39–50, 2013.

[65] Kutub Thakur. Analysis of denial of services (dos) attacks and prevention techniques. *International Journal of Engineering Research & Technology*, 4(7), 2015.

[66] Mrs S Thilagavathi and A Saradha. Impact analysis of dos & ddos attacks. *IOSR Journal of Computer Engineering*, 6(6) :24–33, 2014.

[67] S Umarani and D Sharmila. Predicting application layer ddos attacks using machine learning algorithms. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 8(10) :1912–1917, 2014.

[68] S.V.M. Vishwanathan and M Murty. Ssvm : a simple svm algorithm. volume 3, pages 2393 – 2398, 02 2002.

[69] Gernot Vormayr, Tanja Zseby, and Joachim Fabini. Botnet communication patterns. *IEEE Communications Surveys & Tutorials*, 19(4) :2768–2796, 2017.

[70] Danshi Wang, Min Zhang, Meixia Fu, Zhongle Cai, Ze Li, Huanhuan Han, Yue Cui, and Bin Luo. Nonlinearity mitigation using a machine learning detector based on $k$-nearest neighbors. *IEEE Photonics Technology Letters*, 28(19) :2102–2105, 2016.

[71] Yanxin Wang, Johnny Wong, and Andrew Miner. Anomaly intrusion detection using one class svm. In *Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004.*, pages 358–364. IEEE, 2004.

[72] Yang Xin, Lingshuang Kong, Zhi Liu, Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, and Chunhua Wang. Machine learning and deep learning methods for cybersecurity. *IEEE Access*, 6 :35365–35381, 2018.

[73] Shui Yu. *Distributed denial of service attack and defense*. Springer, 2014.

[74] Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.

[75] Ethan Zuckerman, Hal Roberts, Ryan McGrady, Jillian York, and John Palfrey. Distributed denial of service attacks against independent media and human rights sites. *The Berkman Center*, 2010.