

Chapter 3

Feature selection and machine learning model optimization for DDoS detection

3.0 Project Implementation

3.1 Introduction

In this chapter, the Implementation design is divided into two sections, the system setup design, and the machine learning flow process design. Both are explained in detail in the section.

The dataset used for the experiment was obtained from the open-source database CSE-CIC-IDS2018 ('IDS 2018' 2022) and the generated data set.

The open-source dataset was used to train a large data set that contained 300967 instances of benign and DDoS datasets while the generated data set contained 28, 972 instances of benign and DDoS datasets. The dataset contained many fields in which "32 out of 80" features were used for the open data set and "31 out of 84" was used for the newly generated dataset.

A comparative study of supervised machine learning algorithms which will be used to predict the accuracy and clarity of how DDoS attacks are detected in the cloud will be presented and evaluated in terms of performance and accuracy.

3.2 System Setup

Figure 8 depicts the system set-up of this experiment comprising of host pc, kali Linux VM, Owncloud Platform and Oracle Virtual Box. Table 4 shows the components needed to set up the system for the experiment.

S/N	Components	Description
1	Hardware	Intel Core i7 8th Gen PC
2	Operating System	Windows 10
3	Memory	34 GB RAM
4	Oracle VirtualBox	Virtual Machine v 6.0
5	Owncloud VirtualBox	UCS 4.4-with-owncloud 1
6	Kali Linux VirtualBox	Kali-Linux-2021.4a-VirtualBox-amd64
7	Slowloris Tool	Slowloris DDoS attack script

Table 4: Requirements for system setup design



Figure 8: System Setup Design.

Brief description of the System setup elements as shown in Figure 8.

3.2.1 Network Type

A cloud network was deployed, specifically, Owncloud VirtualBox. This application was downloaded as an ova file from the internet and imported as a VirtualBox machine

in the oracle VM installed in the host machine (researchers Computer). Other resources were configured such as setting up a graphical user interface, setting the application package, setting up an email server, and setting up two systems shown in Figure 8 for the collection of benign and DDoS traffics.

3.2.2 Kali Linux VM

This is a free and open-source operating system, its source code is freely accessible online, allowing you to examine it and modify it to suit your needs.

As shown in Figure 8 the system was used as an attacker PC which was used to conduct DDoS attacks on the targeted PC (attack pc) and for sending normal traffic to the own cloud benign PC.

3.2.3 Oracle VirtualBox

This is a system where the Owncloud VM and kali Linux VM were hosted which helps to experiment with a tight connection.

3.2.4 Host PC

This is the researcher's PC where all the VMs were hosted to perform the experiment.

3.2.5 Performing of DDoS attack in Owncloud VM

The DDoS attacks were conducted on the Owncloud VirtualBox environment which is a free and open-source platform for the generation of attacks. The attack was created in a safe environment using the slowloris tool for eight hours daily for three days. The attacking machine comprises Kali-Linux-2021.4a-VirtualBox-amd64, Debian 64 operating system with an IP 10.0.2.15 hosted on Oracle VirtualBox. Several terminals were opened representing several systems as shown in Figure 9 on each of the terminals, the slowloris script known as the bot was used with the IP address (192.168.1.36) of the attack PC (Target system) to send attack traffic to the attack PC seen in Figure 9.

The slowloris tool created traffic at the sink nodes during the execution of the DDoS attack, and the Wireshark packet analyser captured strange traffic during this time. The system then received the recorded traffic.



Figure 9: Performing DDoS attacks on Owncloud.

Attack tool	Duration	Attacker	IP	Victim	IP
Slowloris	8 hours in a day. Total of 3 days	Kali Linux VM	10.0.2.15	Owncloud	192.168.1.36

Table 5: Specification of tools and duration of DDoS attack traffic

3.2.6 Performing of Benign Traffic

Another system was set up in the Owncloud VM named benign PC with IP address 192.168.1.36 as shown in Figure 9 which was used to capture normal traffic. To build normal traffic, an email service was set up in the Owncloud benign PC such that there will be mail communication in and out of the system to create an SMTP protocol. Other

tools like Nmap and ping were used to send traffic to the benign system to create varieties of protocols for normal traffic captured by Wireshark.

Method	Duration	System	IP	Platform	IP
Nmap, Email, Ping, web browsing	6 hours a day, a total of 3 days	Kali Linux VM	10.0.2.15	Owncloud	192.168.1.22

Table 6: Methods for collection and duration of benign traffic

3.2.7 Collection of Pcap Files

The DDoS and benign traffic captured by Wireshark was saved on the desktop of the Owncloud and collected as a pcap file. The pcap file was transferred to the host system through email configured in the Owncloud, the ones with large sizes were saved to one drive and then transported to the host PC. Furthermore, a filtering method was used to select real DDoS traffic from the pcap generated to ascertain that only DDoS traffic was collected. On the other hand, Benign traffic saved as a pcap file was transferred through email to the host PC and the large sizes were saved to one drive in the Owncloud and finally transported to the host PC. These two pcap files (DDoS and benign) were stored in the host PC for further processing. Machine learning process design as the second phase will be used to explain other processes involved in the experiment.

📕 D	DoS.Data.p	кар																-		×
File	Edit	View	Go	Capture	Analyze	Statistics	Telephony	Wireless	Tools	Help										
		0		2 🗙	¢	> 🗟 🕅 🤞	1 I I I I I I I I I I I I I I I I I I I	⊕, ⊖, ⊜,												
	Apply a d	isplay filt	er <	Ctrl-/>															•)+
No.		Time				Source	De	stination		Protocol	Lenç	Info								
	2196	2022-0	7-02	11:09:22	2.248667	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → 5	50473	[RST]	Seq=1	Win=0	Len=0		-	
	2197	2022-0	7-02	11:09:22	2.248797	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → 5	50474	[RST]	Seq=1	Win=0	Len=0			
	2198	2022-0	7-02	11:09:22	2.248827	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → 5	50475	[RST]	Seq=1	Win=0	Len=0			
	2199	2022-0	7-02	11:09:22	2.248952	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → 5	50476	[RST]	Seq=1	Win=0	Len=0			
	2200	2022-0	7-02	11:09:22	2.248991	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → 5	50478	[RST]	Seq=1	Win=0	Len=0			
	2201	2022-0	7-02	11:09:22	2.249104	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → 5	50479	[RST]	Seq=1	Win=0	Len=0		į.	
	2202	2022-0	7-02	11:09:22	2.249140	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → <u>5</u>	50480	[RST]	Seq=1	Win=0	Len=0			
	2203	2022-0	7-02	11:09:22	2.249207	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → <u>5</u>	50481	[RST]	Seq=1	Win=0	Len=0			
	2204	2022-0	7-02	11:09:22	2.249222	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → <u>5</u>	50482	[RST]	Seq=1	Win=0	Len=0		[
	2205	2022-0	7-02	11:09:2	2.249340	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → <u>5</u>	50483	[RST]	Seq=1	Win=0	Len=0			
	2206	2022-0	7-02	11:09:2	2.249355	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → <u>5</u>	50484	[RST]	Seq=1	Win=0	Len=0			
	2207	2022-0	7-02	11:09:2	2.249476	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → <u>5</u>	50485	[RST]	Seq=1	Win=0	Len=0		ŀ	
	2208	2022-0	7-02	11:09:2	2.249493	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → <u>5</u>	50486	[RST]	Seq=1	Win=0	Len=0			
	2209	2022-0	7-02	11:09:22	2.249635	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → <u>5</u>	50487	[RST]	Seq=1	Win=0	Len=0			
	2210	2022-0	7-02	11:09:2	2.249651	192.168.1	.36 19	2.168.1.22	2	тср	54	80 → <u>5</u>	50488	[RST]	Seq=1	Win=0	Len=0			
	2211	2022-0	7-02	11:09:2	2.249721	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → 5	50489	[RST]	Seq=1	Win=0	Len=0			
	2212	2022-0	7-02	11:09:22	2.249737	192.168.1	.36 19	2.168.1.22	2	ТСР	54	80 → 5	50490	[RST]	Seq=1	Win=0	Len=0			
	2213	2022-0	7-02	11:09:2	2.249802	192.168.1	.36 19	2.168.1.22	2	TCP	54	80 → 5	50491	[RST]	Seq=1	Win=0	Len=0			

Figure 10: Sample DDoS pcap file.

File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Heip Ipaddr = 0	NOrmal Tr	off Data.pcap													× c
Deader Deader Deader Deader Protocol Lens, Info 57129 2022-07-22 14:59:133.1096202 122.168.1.36 122.168.1.36 127.168.1.32 120.168.1.36 127.168.1.36 <t< th=""><th>File Edit</th><th>View Go</th><th>Capture</th><th>Analyze</th><th>Statistics Telepho</th><th>ony Wireless Tools</th><th>Help</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></t<>	File Edit	View Go	Capture	Analyze	Statistics Telepho	ony Wireless Tools	Help								
Deck Description Descrescription Description	A = 2		XO	२ ०० ००		<u> </u>									
No. Time Gourde Destination Protocol Len; Info 57169 2022-07-22 14:59:13.169620 192.168.1.36 122.168.1.36 1CMP 98 Echo (p1ng) reply 1d=bx0001.seq=41166/52. 57170 2022-07-22 14:59:13.1696.1.22 192.168.1.36 ICMP 98 Echo (p1ng) reply 1d=bx0001.seq=41166/52. 57170 2022-07-22 14:59:13.12721 192.168.1.36 ICMP 98 Echo (p1ng) reply 1d=bx0001.seq=41166/52. 57173 2022-07-22 14:59:15.217251 192.168.1.36 ICMP 98 Echo (p1ng) reply 1d=bx0001.seq=4116/52. 57173 2022-07-22 14:59:15.217255 192.168.1.36 ICMP 98 Echo (p1ng) reply 1d=bx0001.seq=4117/54. 57174 2022-07-22 14:59:37.26806 192.168.1.36 ICMP 98 Echo (p1ng) reply 1d=bx0001.seq=4117/54. 57174 2022-07-22 14:59:37.26806 192.168.1.36 ICMP 98 Echo (p1ng) reply 1d=bx0001.seq=4117/54. 57174 2022-07-22 14:59:37.26806	ip.addr	== 192.168.1.22	2												- +
57109 2022-07-22 14:59:33.169620 192.168.1.36 192.168.1.22 102 57170 2022-07-22 14:59:33.169620 192.168.1.36 102.168.1.32 102 57171 2022-07-22 14:59:33.199420 192.168.1.36 102.168.1.32 102 57171 2022-07-22 14:59:33.199420 192.168.1.36 102.168.1.32 104 1040001, seq-41166/52. 57171 2022-07-22 14:59:33.27221 192.168.1.36 1122 104 104 104 104 104 104 104 104 104 104 104 104 104 104 104 104 104 104 104 102 104 </td <td>No.</td> <td>Time</td> <td></td> <td>5</td> <td>Source</td> <td>Destination</td> <td>Protocol</td> <td>Lenç</td> <td>Info</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td>-</td>	No.	Time		5	Source	Destination	Protocol	Lenç	Info						-
57170 2022-07-22 14:59:34.193463 192.168.1.36 ICMP 98 Echo (ping) request id=8x0001, seq=41166/52, id=8x001, seq=41166/52, id=8x001, seq=41166/52, id=8x001, seq=41166/52, id=8x001, seq=41167/52, id=8x001, seq=41177/54, id=8x01, seq=41177/54, id=8x001, seq=41177/54, id=8x001, seq=41177/54, id=8x001, seq=41177/54, id=8x001, seq=41177/54, id=8x01, id=8x01, seq=41177/54, id=8x01, seq=41177/54, id=8x01, id=8x01, seq=4x01, id=8x001, seq=4x01, id=8x01,	57169	2022-07-22	14:59:33.	169620 1	92.168.1.36	192.168.1.22	ICMP	98	Echo	(ping)	reply	id=0x0001,	seq=411	63/52.	
57171 2022-07-22 14:59:13.41:93491 192.168.1.36 192.168.1.22 ICMP 98 Echo (ping) reply id=0x0001, seq=41166/52. 57172 2022-07-22 14:59:13.21:725 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) requyst id=0x0001, seq=41166/52. 57173 2022-07-22 14:59:13.21:725 192.168.1.36 192.168.1.32 ICMP 98 Echo (ping) requyst id=0x0001, seq=41166/53. 57173 2022-07-22 14:59:13.21:725 192.168.1.36 192.168.1.32 ICMP 98 Echo (ping) requyst id=0x0001, seq=41169/53. 57173 2022-07-22 14:59:13.20:660:192.168.1.22 192.168.1.32 ICMP 98 Echo (ping) requyst id=0x0001, seq=41179/54. 57175 2022-07-22 14:59:13.20:660:192.168.1.32 192.168.1.36 ICMP 98 Echo (ping) requyst id=0x0001, seq=41175/55. 57175 2022-07-22 14:59:13.20:660:192.168.1.32 192.168.1.36 ICMP 98 Echo (ping) requyst id=0x0001, seq=41176/55. 57178 2022-07-22 15:99:13.20:67719 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) requyst id=0x0001, seq=41178/55. 57178 2022-07-22 15:99:18.20:80:12:12 192.168.1.22 IDM id=0x0001, seq=41178/55. id=0x0001, seq=41178/55. 57178 2022-07-22 15:90:16.170:70:19:16.1.22 239.255.255.250 SDD 217 M-SEARCH + HTTP/1.1 id=0x0001, seq=411	57170	2022-07-22	14:59:34	193463 1	92.168.1.22	192.168.1.36	ICMP	98	Echo	(ping)	request	id=0x0001,	seq=411	66/52.	-
57172 2022-07-22 14:59:35.217228 192.168.1.36 ICMP 98 Echo (ping) request id=x00001, seq=41169/53. 57173 2022-07-22 14:59:35.21725 192.168.1.36 ICMP 98 Echo (ping) request id=x0001, seq=41169/53. 57175 2022-07-22 14:59:36.24207 192.168.1.36 ICMP 98 Echo (ping) reply id=x0001, seq=41169/53. 57175 2022-07-22 14:59:36.24207 192.168.1.36 I22.168.1.36 ICMP 98 Echo (ping) reply id=x00001, seq=41175/55. 57175 2022-07-22 14:59:38.26064 192.168.1.32 I22.168.1.36 ICMP 98 Echo (ping) reply id=x0001, seq=41175/55. 57177 2022-07-22 14:59:38.260677 102.168.1.32 I22.168.1.36 ICMP 98 Echo (ping) reply id=x0001, seq=41176/55. 57179 2022-07-22 15:39:38.260677 102.168.1.32 I20.168.1.32 IS0.168.1.22	57171	2022-07-22	14:59:34	193491 1	92.168.1.36	192.168.1.22	ICMP	98	Echo	(ping)	reply	id=0x0001,	seq=411	66/52.	-
57173 2022-07-22 14:59:35.217255 192.168.1.36 192.168.1.22 ICMP 98 Echo (ping) reply id=0x0001, seq=41109/53. 57174 2022-07-22 14:59:36.242071 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) reply id=0x0001, seq=41172/54. 57175 2022-07-22 14:59:36.242071 192.168.1.36 192.168.1.32 ICMP 98 Echo (ping) reply id=0x0001, seq=41172/54. 57175 2022-07-22 14:59:37.250601 192.168.1.36 192.168.1.32 ICMP 98 Echo (ping) reply id=0x0001, seq=41172/54. 57175 2022-07-22 14:59:37.25061 192.168.1.36 192.168.1.32 ICMP 98 Echo (ping) reply id=0x0001, seq=41172/54. 57175 2022-07-22 14:59:37.25071 192.168.1.36 192.168.1.36 ICMP 98 Echo (ping) reply id=0x0001, seq=41178/55. 57175 2022-07-22 15:90:18.37971 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) reply id=0x0001, seq=41178/55. 60386 2022-07-22 15:30:16.37071 192.168.1.22 192.168.1.22 192.255.255.250 SSDP 217 M-SEARCH + HTTP/1.1 60386 2022-07-22 15:30:16.37071 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH + HTTP/1.1 10630601, seq=4118/56. 60386 2022-07-22 15:30:16.37071 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH + HTTP/1.1 106302 202-07-21 15:30:16.37071 192.	57172	2022-07-22	14:59:35.	217228 1	192.168.1.22	192.168.1.36	ICMP	98	Echo	(ping)	request	id=0x0001,	seq=411	69/53.	-
57174 2022-07-22 14:59:36.242034 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request id=xx0001, seq=41172/54. 57175 2022-07-22 14:59:36.242034 192.168.1.36 ICMP 98 Echo (ping) request id=xx0001, seq=41172/54. 57175 2022-07-22 14:59:37.265832 192.168.1.36 ICMP 98 Echo (ping) request id=xx0001, seq=41172/54. 57175 2022-07-22 14:59:37.265832 192.168.1.36 ICMP 98 Echo (ping) request id=xx0001, seq=41175/55. 57178 2022-07-22 14:59:37.265812 192.168.1.32 192.168.1.36 ICMP 98 Echo (ping) request id=xx0001, seq=41175/55. 57178 2022-07-22 14:59:37.265812 192.168.1.32 192.168.1.36 ICMP 98 Echo (ping) request id=xx0001, seq=41176/55. 57179 2022-07-22 15:39:15.26757 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request id=xx0001, seq=41176/55. 57179 202-07-22 15:39:15.26757 192.168.1.22 192.255.255.55 5500 217 M-5EARCH * MTTP/1.1 id=50x001, seq=4118/55.	57173	2022-07-22	14:59:35.	.217255 1	192.168.1.36	192.168.1.22	ICMP	98	Echo	(ping)	reply	id=0x0001,	seq=411	69/53.	-
57175 2022-07-22 14:59:36.242075 192.168.1.36 192.168.1.22 ICMP 98 Echo (ping) request id=80001, seq=41172/54. 57176 2022-07-22 14:59:38.24027 192.168.1.36 ICMP 98 Echo (ping) request id=80001, seq=41172/54. 57176 2022-07-22 14:59:38.24026 192.168.1.36 ICMP 98 Echo (ping) request id=80001, seq=41172/54. 57178 2022-07-22 14:59:38.24064 192.168.1.36 ICMP 98 Echo (ping) request id=80001, seq=41178/55. 57179 2022-07-22 15:99:38.24077 192.168.1.32 192.168.1.36 ICMP 98 Echo (ping) request id=80001, seq=41178/55. 57179 2022-07-22 15:99:38.24076 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request id=80001, seq=41178/55. 57180 2022-07-22 15:30:16.16.3707 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request id=80001, seq=41181/56. 60384 2022-07-22 15:30:16.16.3707 192.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 60386 2022-07-22 15:30:16.37071 192.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 60386 2022-07-2	57174	2022-07-22	14:59:36	242034 1	192.168.1.22	192.168.1.36	ICMP	98	Echo	(ping)	request	id=0x0001,	seq=411	72/54.	-
57176 2022-07-22 14159137.265832 192.168.1.36 ICMP 98 Echo (ping) request id=0x0001, seq=41175/55. 57177 2022-07-22 14159137.265802 192.168.1.36 ICMP 98 Echo (ping) request id=0x0001, seq=41175/55. 57178 2022-07-22 14159138.280464 192.168.1.36 ICMP 98 Echo (ping) request id=0x0001, seq=41178/55. 57178 2022-07-22 14159138.280464 192.168.1.36 ICMP 98 Echo (ping) request id=0x0001, seq=41178/55. 57170 2022-07-22 14159138.280464 192.168.1.36 ICMP 98 Echo (ping) request id=0x0001, seq=41178/55. 57108 2022-07-22 1519018.0316484 192.168.1.32 IDM 98 Echo (ping) request id=0x0001, seq=4118/55. 60385 2022-07-22 1519017.016.01.32 209.255.255.55 S050 217 M-SEARCH HTTP/1.1 60385 2022-07-22 1519017.017.015 103.168.1.22 209.255.255.250 SD0 217 M-SEARCH HTTP/1.1 60387 2022-07-22 1519017.077195 103.168.1.22 209.255.255.250	57175	2022-07-22	14:59:36	242079 1	192.168.1.36	192.168.1.22	ICMP	98	Echo	(ping)	reply	id=0x0001,	seq=411	72/54.	
57177 2022-07-22 14:59:137, 265866 192.168.1.36 192.168.1.32 ICMP 98 Echo (ping) request 1d=0x0001, seq=41175/55. 57178 2022-07-22 14:59:137, 265866 192.168.1.36 ICMP 98 Echo (ping) request 1d=0x0001, seq=41175/55. 57180 2022-07-22 14:59:137, 265866 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request 1d=0x0001, seq=41175/55. 57180 2022-07-22 14:59:139.314484 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request id=0x0001, seq=41181/56. 66384 2022-07-22 15:30:16.37071 102.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 60386 2022-07-22 15:30:16.37071 192.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 60386 2022-07-22 15:30:17.377164 102.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 60386 2022-07-22 15:30:17.377064 102.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 60392 2022-07-22 15:30:13.339643 102.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 <td>57176</td> <td>2022-07-22</td> <td>14:59:37</td> <td>265832 1</td> <td>192.168.1.22</td> <td>192.168.1.36</td> <td>ICMP</td> <td>98</td> <td>Echo</td> <td>(ping)</td> <td>request</td> <td>id=0x0001,</td> <td>seq=411</td> <td>75/55.</td> <td>22⁽¹⁾</td>	57176	2022-07-22	14:59:37	265832 1	192.168.1.22	192.168.1.36	ICMP	98	Echo	(ping)	request	id=0x0001,	seq=411	75/55.	22 ⁽¹⁾
57178 2022-07-22 14:59:38.289646 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request 1d=0x0001, seq=41178/55. 57179 2022-07-22 14:59:38.289646 192.168.1.36 ICMP 98 Echo (ping) request 1d=0x0001, seq=41178/55. 57180 2022-07-22 14:59:38.289646 192.168.1.36 ICMP 98 Echo (ping) request 1d=0x0001, seq=41178/55. 57180 2022-07-22 15:39:16.176376 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request 1d=0x0001, seq=41178/55. 60384 2022-07-22 15:39:16.176376 192.168.1.22 239.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60385 2022-07-22 15:39:16.176376 192.168.1.22 239.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60387 2022-07-22 15:39:16.176376 192.168.1.22 239.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60387 2022-07-22 15:39:16.176376 192.168.1.22 239.255.255 SSDP 217 M-SEARCH * HTTP/1.1 60392 2022-07-22 15:39:16.3.177976 192.168.1.22	57177	2022-07-22	14:59:37	265860 1	192.168.1.36	192.168.1.22	ICMP	98	Echo	(ping)	reply	id=0x0001,	seq=411	75/55.	
57179 2022-07-22 14:59:38.289677 192.168.1.32 102.168.1.22 ICMP 98 Echo (ping) reply id=0x0001, seq=41181/56_ 60384 2022-07-22 15:30:16.176376 192.168.1.22 129.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60385 2022-07-22 15:30:16.37027 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60386 2022-07-22 15:30:17.330648 192.168.1.22 239.255.255.550 SDP 217 M-SEARCH * HTTP/1.1 60387 2022-07-22 15:30:10.17305 102.168.1.22 239.255.255.550 SDP 217 M-SEARCH * HTTP/1.1 60387 2022-07-22 15:30:10.1730548 129.168.1.22 239.255.255.550 SDP 217 M-SEARCH * HTTP/1.1 60392 2022-07-22 15:30:10.30643 129.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 60392 2022-07-22 15:30:10.30643 129.168.1.22 239.255.255.250 SDP 217 M-SEARCH * HTTP/1.1 60392 2022-07-22 15:30:10.307661 129.168.1.22 239.255.255.250 SDP	57178	2022-07-22	14:59:38	289646 1	192.168.1.22	192.168.1.36	ICMP	98	Echo	(ping)	request	id=0x0001,	seq=411	78/55.	
57180 2022-07-22 14:59:39.314484 192.168.1.22 192.168.1.36 ICMP 98 Echo (ping) request id=0x0001, seq=41181/56_ 60384 2022-07-22 15:30:16.176376 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH + HTTP/1.1 60385 2022-07-22 15:30:16.170376 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH + HTTP/1.1 60386 2022-07-22 15:30:16.177105 101.166.1.22 239.255.255.250 SSDP 217 M-SEARCH + HTTP/1.1 60386 2022-07-22 15:30:16.177105 101.166.1.22 239.255.255.50 SSDP 217 M-SEARCH + HTTP/1.1 60391 202-07-22 15:30:16.330643 102.168.1.22 239.255.255.50 SSDP 217 M-SEARCH + HTTP/1.1 60392 2022-07-22 15:30:10.330643 102.168.1.22 239.255.255.50 SSDP 217 M-SEARCH + HTTP/1.1 60392 2022-07-22 15:30:10.370760 192.168.1.22 239.255.255.50 SDP 217 M-SEARCH + HTTP/1.1	57179	2022-07-22	14:59:38	289677 1	192.168.1.36	192.168.1.22	ICMP	98	Echo	(ping)	reply	id=0x0001,	seq=411	78/55.	-
60384 2022-07-22 15:30:16.176376 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60385 2022-07-22 15:30:17.177195 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60387 2022-07-22 15:30:17.13705 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60387 2022-07-22 15:30:17.13705 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60387 2022-07-22 15:30:18.137057 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60392 2022-07-22 15:30:18.1330643 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60392 2022-07-22 15:30:18.1370649 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1 60392 2022-07-22 15:30:18.1370649 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1	57186	2022-07-22	14:59:39	.314484 1	192.168.1.22	192.168.1.36	ICMP	98	Echo	(ping)	request	id=0x0001,	seq=411	81/56.	-
66385 2022-07-22 15:80:16.337027 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH HTTP/1.1 60386 2022-07-22 15:80:17.177195 102.168.1.22 239.255.255.250 SSDP 217 M-SEARCH HTTP/1.1 60387 2022-07-22 15:80:17.339648 102.168.1.22 239.255.255.250 SSDP 217 M-SEARCH HTTP/1.1 60387 2022-07-22 15:80:18.17827 102.168.1.22 239.255.255.50 SDP 217 M-SEARCH HTTP/1.1 60392 2022-07-22 15:80:18.339643 102.168.1.22 239.255.255.50 SDP 217 M-SEARCH HTTP/1.1 60392 2022-07-22 15:80:18.370761 102.168.1.22 239.255.255.250 SDP 217 M-SEARCH HTTP/1.1 60392 2022-07-22 15:30:10.370761 102.168.1.22 239.255.255.250 SDP 217 M-SEARCH HTTP/1.1	603	A 2022 07 2	15.20.1	6 176376	103 168 1 33	220 255 255 250	SEDD	21-	7 M CC	ADCH *	HTTD/1	1			
60383 2022-07-22 13:50:10.3370J 19:108.122 23:25:25:25 SSDP 217 M-SEARCH HTTP/1.1 60386 0022-07-22 15:30:17.13705 192:168.1.22 23:9.25:25:25:25 SSDP 217 M-SEARCH HTTP/1.1 60387 2022-07-22 15:30:18.137827 192:168.1.22 23:9.255:25:25 SSDP 217 M-SEARCH HTTP/1.1 60392 2022-07-22 15:30:18.139643 192:168.1.22 23:9.255:25:25 SSDP 217 M-SEARCH HTTP/1.1 60392 2022-07-22 15:30:18.139643 192:168.1.22 239:255:25:25 SSDP 217 M-SEARCH HTTP/1.1 60392 2022-07-22 15:30:18.139643 192:168.1.22 239:255:25:55 SSDP 217 M-SEARCH HTTP/1.1	6034	SE 2022-07-2	2 15.30.1	6 227017	102 168 1 22	239.235.255.256	SEDD	21	7 M CE	ARCH +	HTTP/1.	÷			
60387 2022-07-22 15:10:117.330648 102.168.1.22 239.255.255.350 250 217 M-SEARCH HTTP/1.1 60391 2022-07-22 15:10:116.17827 102.168.1.22 239.255.255.250 S5DP 217 M-SEARCH HTTP/1.1 60392 2022-07-22 15:10:116.130643 102.168.1.22 239.255.255.250 S5DP 217 M-SEARCH HTTP/1.1 60392 2022-07-22 15:10:10.137064 192.168.1.22 239.255.255.250 S5DP 217 M-SEARCH HTTP/1.1 60393 2022-07-22 15:10:10.137064 192.168.1.22 239.255.255.250 S5DP 217 M-SEARCH HTTP/1.1	603	6 2022-07-2	2 15.30.1	7 177105	192.100.1.22	239.233.233.230	SSOP	21	7 M CE	ARCH +	HTTP/1.	1			
60307 2022-07-22 13:36:17.33945 192.108.1.22 239.255.255.259 230 217 M:SEARCH M:IT/1.1 60302 2022-07-22 15:30:18.139453 192.168.1.22 239.255.255.259 SSDP 217 M:SEARCH M:IT/1.1 60302 2022-07-22 15:30:18.339643 192.168.1.22 239.255.255.259 SSDP 217 M:SEARCH H:ITP/1.1 60303 2022-07-22 15:30:18.339643 189.2168.1.22 239.255.255.259 SSDP 217 M:SEARCH H:ITP/1.1	603	30 2022-07-2	2 15.30.1	7.330648	192.100.1.22	239.255.255.250	SSUP	21	A M CE	ARCH *	HTTP/1.	1			
60392 2022-07-22 15:30:18.339643 192.168.1.22 239.255.255 SSDP 217 M-SEARCH * HTTP/1.1 60393 2022-07-22 15:30:19.179769 192.168.1.22 239.255.255.250 SSDP 217 M-SEARCH * HTTP/1.1	603	2022-07-2	2 15:30:1	0 170527	192.100.1.22	239.235.255.250	SSOP	211	7 M.CE	ARCH *	HTTP/1	1			
60393 2022-07-22 15:30:19:17969 192:108:1.122 239:255.255.250 SSDP 217 M-SEARCH * HTTP/1.1	603	2022-07-2	2 15:30:1	9 220642	102 168 1 22	230 255 255 250	SEDD	21	7 M-CE	ARCH *	HTTP/1	1			
COSSS FORE OF RE ESTIMATION EXERCICITERE ESTERSTESSTESSTESSTESSTESSTESSTESSTESSTE	603	3 2022-07-2	2 15.30.1	9 179769	192.168.1.22	239 255 255 250	SSOP	213	7 M-SE	ARCH *	HTTP/1	1			
60304 2022-07-22 15-30-10 330021 102 168 1 22 230 255 258 SCDD 217 M-SEADCH * HTTD/1 1	603	4 2022-07-2	2 15:30.1	0 330021	192 168 1 22	230 255 255 250	SSOP	21	7 M.SE	ARCH +	HTTP/1	1			
60396 2022-07-22 15:30:21 374906 192-168.1.22 224.0.0.252 IGMPv2 60 Membership Report group 224.0.0.252	603	6 2022-07-2	2 15:30:2	1.374906	192,168,1,22	224.0.0.252	TGMPy2	66	a Memb	ership	Report of	group 224.0.	0.252		
50404 2022-07-22 15:30:24 875005 192 158 1 22 224 0 0 251 TGMPv2 50 Membership Report group 224 0 0 251	604	4 2022-07-2	2 15:30:2	4.875006	192,168,1,22	224 0 0 251	TGMPv2	66	a Memb	ershin	Report	group 224.0.	0.251		
58485 2822-87-22 15:38:25 875477 192 158 1 22 239 255 255 258 TGMPv2 58 Membership Report group 239 255 255 258	604	15 2022-07-2	2 15:30:2	5.875477	192, 168, 1, 22	239 255 255 258	TGMPv2	60	a Memb	ershin	Report	group 239.25	5.255.25	0	
60406 2022-07-22 15:30:28.875619 192.168.1.22 239.255.102.18 IGMPv2 60 Membership Report group 239.255.102.18	6044	6 2022-07-2	2 15:30:2	8.875619	192.168.1.22	239,255,102,18	IGMPv2	64	a Memb	ership	Report a	group 239.25	5.102.18		
60408 2022-07-22 15:30:42.875414 192.168.1.22 239.255.102.18 IGMPv2 60 Membership Report group 239.255.102.18	604	8 2022-07-2	2 15:30:4	2.875414	192.168.1.22	239.255.102.18	IGMPv2	64	a Memb	ership	Report a	group 239.25	5.102.18		
60409 2022-07-22 15:30:43.874526 192.168.1.22 224.0.0.251 IGMPv2 60 Membership Report group 224.0.0.251	604	9 2022-07-2	2 15:30:4	3.874526	192.168.1.22	224.0.0.251	IGMPv2	64	0 Memb	ership	Report .	group 224.0.	0.251		
2247 2022-07-22 11:09:48.070830 40.99.202.102 192.168.1.36 SMTP 177 S: 220 L04P123CA0664.outlook.office365.com M	224	7 2022-07-2	2 11:09:4	8.070830	3 40.99.202.10	192.168.1.36	SMTP		177 S	: 220	L04P123C	A0664.outlo	ok.offic	e365.	com M
2249 2022-07-22 11:09:48.070975 192.168.1.36 40.99.202.102 SMTP 90 C: EHLO ucs-4579.localnet	224	9 2022-07-2	2 11:09:4	8.07097	5 192.168.1.36	40.99.202.102	SMTP		90 C	: EHLO	ucs-457	9.localnet			
2250 2022-07-22 11:09:48.091328 40.99.202.102 192.168.1.36 SMTP 271 S: 250-L04P123CA0664.outlook.office365.com H	225	0 2022-07-2	2 11:09:4	8.091328	40.99.202.10	192.168.1.36	SMTP		271 S	: 250-	L04P123C	A0664.outlo	ok.offic	e365.	com H
2251 2022-07-22 11:09:48.091458 192.168.1.36 40.99.202.102 SMTP 76 C: STARTTLS	225	1 2022-07-2	2 11:09:4	8.091458	8 192.168.1.36	40.99.202.102	SMTP		76 C	: STAR	TTLS				
2252 2022-07-22 11:09:48.118371 40.99.202.102 192.168.1.36 SMTP 95 S: 220 2.0.0 SMTP server ready	225	2 2022-07-2	2 11:09:4	8.11837:	1 40.99.202.10	192.168.1.36	SMTP		95 S	: 220	2.0.0 SM	TP server r	eady		
2353 2022-07-22 11:11:00.476780 52.97.208.6 192.168.1.36 SMTP 165 S: 220 L04P123CA0018.outlook.office365.com M	235	3 2022-07-2	2 11:11:0	0.47678	3 52.97.208.6	192.168.1.36	SMTP		165 S	: 220	L04P123C	A0018.outlo	ok.offic	e365.	com M
2355 2022-07-22 11:11:00.477010 192.168.1.36 52.97.208.6 SMTP 78 C: EHLO ucs-4579.localnet	235	5 2022-07-2	2 11:11:0	0.477010	3 192.168.1.36	52.97.208.6	SMTP		78 C	: EHLO	ucs-457	9.localnet			
2356 2022-07-22 11:11:00.497241 52.97.208.6 192.168.1.36 SMTP 259 S: 250-L04P123CA0018.outlook.office365.com H	235	6 2022-07-2	2 11:11:0	0.49724:	1 52.97.208.6	192.168.1.36	SMTP		259 S	: 250-	L04P123C	A0018.outlo	ok.offic	e365.	com H
2357 2022-07-22 11:11:00.497390 192.168.1.36 52.97.208.6 SMTP 64 C: STARTTLS	235	7 2022-07-2	2 11:11:0	0.497390	9 192.168.1.36	52.97.208.6	SMTP		64 C	: STAR	TTLS				
2358 2022-07-22 11:11:00.516801 52.97.208.6 192.168.1.36 SMTP 83 S: 220 2.0.0 SMTP server ready	235	8 2022-07-2	2 11:11:0	0.51680	1 52.97.208.6	192.168.1.36	SMTP		83 S	: 220	2.0.0 SM	TP server r	eady		
2696 2022-07-22 11:13:44.102980 40.99.201.246 192.168.1.36 SMTP 177 S: 220 L04P265CA0173.outlook.office365.com M	269	6 2022-07-2	2 11:13:4	4.102980	40.99.201.24	5 192.168.1.36	SMTP		177 S	: 220	L04P265C	A0173.outlo	ok.offic	e365.	com M
2698 2022-07-22 11:13:44.103127 192.168.1.36 40.99.201.246 SMTP 90 C: EHLO ucs-4579.localnet	269	8 2022-07-2	2 11:13:4	4.10312	7 192.168.1.36	40.99.201.246	SMTP		90 C	: EHLO	ucs-457	9.localnet			
2699 2022-07-22 11:13:44.123489 40.99.201.246 192.168.1.36 SMTP 271 5: 250-L04P265CA0173.outlook.office365.com H	269	9 2022-07-2	2 11:13:4	4.123489	40.99.201.24	5 192.168.1.36	SMTP		271 S	: 250-	LO4P265C	A0173.outlo	ok.offic	e365.	com H
2700 2022-07-22 11:13:44.123628 192.168.1.36 40.99.201.246 SMTP 76 C: STARTTLS	276	0 2022-07-2	2 11:13:4	4.123628	8 192.168.1.36	40.99.201.246	SMTP		76 C	: STAR	TTLS				
2701 2022_07_22 11:13:44 146400 40 00 201 246 102 168 1 36 CMTD 05 5: 220 2 0 0 CMTD conven readu	270	1 2022-07-2	2 11.13.4	146400	40 99 201 24	107 168 1 36	SMTP		95 5	. 220	2 A A SM	TD server r	vhea		

Figure 11: Sample Benign pcap file.

The next section continues with the Machine learning flow process.

3.3 Machine Learning Flow Process Design

This process was used in the second phase of the implantation process. Each flow content has been briefly explained in how they were used in this section. The components used to perform the ML process are seen in Table 7.

S/N	Components	Description
1	Hardware	Intel Core i7 8th Gen PC
2	Operating System	Windows 10
3	Memory	34 GB RAM
4	Libraries	Pandas, Sklearn, Matplotlib
5	Storage	MS Excel
	Feature extraction tool	CICFlowMeter 4.0
6	Programming language	Python 3.9
7	IDE	(Anaconda Jupyter Notebook)

Table 7: Components Utilised

Figure 12 illustrates the ML flow process design using supervised learning models.

3.3.1 Raw Data Collection

The raw data collected as pcap files stored on the host PC is required to be transformed into a format that is compatible with model building in machine learning. A tool named CICFlowMeter was used to make this transformation of the raw data into a possible real dataset CSV format as shown in Figure 13, this was stored in the host PC as an excel file. CICFlowMeter can be accessed using this link (https://github.com/ahlashkari/CICFlowMeter).



Figure 12: Machine Learning Flow process design.

3.3.2 Feature Extraction

CICFlowMeter 4.0 is a network traffic flow generator and analyser. over 84 statistical network traffic features, such as Duration, Number of packets, Number of bytes, Length of packets, etc., can be extracted separately in the forward and backward directions when using it to generate bidirectional flows, where the first packet determines the forward (source to destination) and backward (destination to source) directions. Picking features from a list of already-existing features, adding new features, and adjusting the flow timeout period are extra functionalities in this tool (Lashkari et al. 2017; Patil et al. 2022). Figures 14 and 15 (with CICFlowMeter 3.0) show the type of features extracted for this experiment.

1	Sec Port Dst IP	Dst Port	Protocol	Timestamp	Flow Dura	a Tot Fwd P	Tot Bwd P	TotLen Fw T	lotLen Bw Fr	vd Pkt Li Fi	nd Pkt Li P	Fwd Pkt L	Fwd Pkt L	Burd Pkt L	Bood Pikt L	Bwd Pkt I	L Bwd Pkt L	Flow Byte	s, Flow Pkts	Flow IAT I	Flow IAT 9	Flow IAT N	Flow IAT F	vd IAT T Fwd	I IAT N Fwd I/	AT S Fw	d IAT N Fwd II	T N Bwd IAT 1	Bed IAT N	Bed IAT S	Bwd IAT N	Bwd IAT N F
2	80 192.168.1	49702	6	5 11/07/2022 20:3	2 6305158	0	- 4	0	0	0	0	0	0	0	0	0	0 0	(0.634401	2101719	1850449	4000328	303519	0	0	0	0	0 6305158	2101719	1850449	4000328	303519
3	49702 192.168.1	80	ι e	5 11/07/2022 20:3	2 1200971	1	1	0	0	0	0	0	0	0	D	0	0 0	0	1.665319	1200971	0	1200971	1200971	0	0	0	0	0 0	0	0	0	0
4	80 192.168.1	55462	6	5 11/07/2022 20:4	5 14304632	0	5	0	0	0	0	0	0	0	0	0	0 0	(0.349537	3576158	3314477	8000974	302562	0	0	0	0	0 1.43E+07	3576158	3314477	8000974	302562
5	55462 192,168,1	80	6	5 11/07/2022 20:4	5 1200834	1	1	0	0	0	0	0	0	0	0		0 0	(1.665509	1200834	0	1200834	1200834	0	0	0	0	0 0	0	0	0	0
6	55462 192 168 1	80	6	5 11/07/2022 20:4	5 4799955	1	1	0	0	0	0	0	0	0	0		0 0		0.416671	4799955	0	4799955	4799955	0	0	0	0	0 0	0	0	0	0
7	80 192 168 1	55926		5 11/07/2022 20:4	6 14303206	0	5	0	0	0	0	0	0	0	0		0 0		0.349572	3575802	3314312	8000014	301980	0	0	0	0	0 1.43E+07	3575802	3314312	8000014	301980
8	80 192 168 1	57633		5 11/07/2022 20-5	0 41141255	1	1	0	0	0	0	0	0	0	0		0		0.048613	4 11F+07	0	4 11F+07	4 11F+07	0	0	0	0	0 0	0	0	0	0
9	80 192 168 1	57760		5 11/07/2022 20:5	0 42356120	2	1	507	189	507		253.5	358 5031	189	189	189		16,4321	0.070828	2 125+07	3.00F+07	4.24F+07	169	169	169	0	169	169 0		0		
n	80 197 168 1	58255		5 11/07/2022 20:1	4 14002211		- 4	0	0				0	0	0				0.285669	4557404	3054632	8000381	2001302	0	0		0	0 1405+07	4557404	3054632	8000381	2001302
ñ	80 192 168 1	\$7930		11/07/2022 20-1	8 14002419		4	0						0					0.285665	4667473	1055807	8001634	1000855	0	0			0 1 405+07	4667473	1055927	8001634	1000855
12	80 102 168 1	67016		5 11/07/2022 20:2	8 14000720		-	0	0		0			0	0				0.285600	4666021	1055578	8001148	2000410	0	0		0	0 1405-07	4444023	1055578	8001148	2000410
12	80 103 168 1	57017		11/07/2022 20:3	2 1/000500			0						0	0	-			0.105662	4667501	205/707	8000250	2000620	0	0			0 1 405-07	4667501	205/207	0000250	2000520
1	80 101 108 1	57527		11/07/2022 20.3	4 140000000		- 1	0			0			0		-			0.285603	4667000	3034737	8001154	2000039	0	0			0 1.405-07	4007301	3034131	8001350	2000603
2	80 102 168 1	57948		11/07/2022 20.3	4 6001100			0			0			0	0	-			0.483834	1007980	3034977	40001336	2003022	0	0		0	0 1.400+07	400/980	3034977	4000434	2003522
1	80 192.208.1	5/949		11/07/2022 20:3	* 6001300		3	0						0		-			0.00000	3000530	3434030	4000434	2000000	0	0	-		0 60001100	4000530	2424030	4000434	2000000
20	80 192.168.1	58060		5 11/07/2022 20:3	e ee111e22		4	0						0		-			0.06801	2.210+07	2.840+07	4.210+07	1999042	0	0	-		0 1999042	1333245		1333245	1333245
10	80 192.168.1	580/8		5 11/07/2022 20:3	4 14001300			0	0		0	0	0	0	0	-			0.285692	4667035	3054806	8000117	2000615	0	0		0	0 1.402+07	4667033	3034806	8000117	2000615
10	80 192.208.1	50004		11/07/2022 20.3	* 14001132			0			0			0		-			0.285691	4007044	3033040	8000385	12222294	0	0	-		0 1.400407	4007044	3033040	8000385	13333394
12	80 192.168.1	58092		5 11/07/2022 20:1	4 14000819		4	0	0		0	0		0	0	-	0		0.285698	4666940	3054822	8000256	2000898	0	0	-	0	0 1.406+07	4666940	3054822	8000256	2000898
0.	80 192.168.1	58083		5 11/07/2022 20:1	4 14002072	0	-	0	0		0	0	0	0	0	-			0.285672	460/33/	3054556	1444421	2000/22	0	0		0	0 1.402+07	400/33/	3034536	1444421	2000/22
2	80 192.168.1	58091		5 11/07/2022 20:3	4 14001043	0		0	0		0	0	0	0	0		0		0.285693	4667014	3054803	8000272	2000955	0	0		0	0 1.405+07	466/014	5054803	8000272	2000955
2	80 192.168.1	58577		5 11/07/2022 20:1	5 14004445	0	4	0	0	0	0	0	0	0	0		0		0.285624	4668148	3054047	7999258	2000104	0	0	0	0	0 1.408+07	4668148	3054047	7999258	2000104
2.	80 192.168.1	58221		5 11/07/2022 20:1	4 66196448	1	2	0	0	0	0	0	0	0	0		0 0		0.04532	5.51E+07	4.40E+07	6.42E+07	2005425	0	0	0	0	0 2001425	2001425	0	2001425	2001425
24	80 192.168.1	58222	6	5 11/07/2022 20:1	4 48086678	1	2	0	0	0	0	0	•	0	0		0 0		0.062387	2.40E+07	3.12E+07	4.61E+07	2000496	0	0	0	0	0 2000495	2000496	0	2000496	2000496
2	80 192.168.1	58223		5 11/07/2022 20:3	4 14002887	0	4	0	0	0	0	0	0	0	0		0 0	(0.285655	4667629	3055055	8000822	2000669	0	0	0	0	0 1.40E+07	4667629	3055055	8000822	2000669
28	80 192.168.1	58239	6	5 11/07/2022 20:1	4 45041762	1	1	0	0	0	0	0	0	0	0		0 0		0.044403	4.50E+07	0	4.50E+07	4.50E+07	0	0	0	0	0 0	0	0	0	0
23	80 192.168.1	58240	6	5 11/07/2022 20:1	4 63151777	1	1	0	0	0	0	0	0	0	0		0 0	0	0.03167	6.32E+07	0	6.32E+07	6.32E+07	0	0	0	0	0 0	0	0	0	0
28	80 192.168.1	58244	6	5 11/07/2022 20:3	4 14000963	0	- 4	0	0	0	0	0	0	0	0		0 0	0	0.285695	4666988	3055049	8000259	2000201	0	0	0	0	0 1.40E+07	4666988	3055049	8000259	2000201
25	80 192.168.1	58245	6	5 11/07/2022 20:1	4 14001227	0	4	0	0	0	0	0	0	0	0		0 0		0.285689	4667076	3055438	8000803	2000013	0	0	0	0	0 1.40E+07	4667076	3055438	8000803	2000013
3	80 192.168.1	58246		5 11/07/2022 20:3	4 14001753	0	- 4	0	0	0	0	0	0	0	0		0 0		0.285679	4667251	3054774	8000235	2000730	0	0	0	0	0 1.40E+07	4667251	3054774	8000235	2000730
31	80 192.168.1	58247	6	5 11/07/2022 20:1	4 14002510	0	- 4	0	0	0	0	0	0	0	0		0 0	(0.285663	4667503	3054587	8000210	2001000	0	0	0	0	0 1.40E+07	4667503	3054587	8000210	2001000
37	80 192.168.1	58249	6	5 11/07/2022 20:1	4 13997005	0	4	0	0	0	0	0	0	0	0		0 0		0.285775	4665668	3056599	8000038	1996344	0	0	0	0	0 1.40E+07	4665668	3056599	8000038	1996344
33	80 192.168.1	58252		5 11/07/2022 20:3	4 14013025	0	- 4	0	0	0	0	0	0	0	0		0 0		0.285449	4671008	3054339	8001136	2000123	0	0	0	0	0 1.40E+07	4571008	3054339	8001136	2000123
34	80 192.168.1	58254	6	5 11/07/2022 20:1	4 37378081	1	1	0	0	0	0	0	0	0	0		0 0	0	0.053507	3.74E+07	0	3.74E+07	3.74E+07	0	0	0	0	0 0	0	0	0	0
35	80 192.168.1	58363	6	5 11/07/2022 20:3	4 14000867	0	- 4	0	0	0	0	0	0	0	0		0 0		0.285697	4666956	3054618	8000154	2001340	0	0	0	0	0 1.40E+07	4666956	3054618	8000154	2001340
36	80 192.168.1	58384	6	5 11/07/2022 20:3	4 43246454	1	2	0	0	0	0	0	0	0	0		0 0		0.06937	2.16E+07	2.78E+07	4.12E+07	2000631	0	0	0	0	0 2000631	2000631	0	2000631	2000631
37	80 192.168.1	58385	6	5 11/07/2022 20:1	4 14000519	0	- 4	0	0	0	0	0	0	0	0		0 0	0	0.285704	4666840	3055017	8000175	2000280	0	0	0	0	0 1.40E+07	4666840	3055017	8000175	2000280
35	80 192.168.1	58389	6	5 11/07/2022 20:3	4 63057242	1	1	0	0	0	0	0	0	0	0	0	0 0	(0.031717	6.31E+07	0	6.31E+07	6.31E+07	0	0	0	0	0 0	0	0	0	0
35	80 192.168.1	58410	6	5 11/07/2022 20:1	5 14002187	0	4	0	0	0	0	0	0	0	0	0	0 0	(0.28567	4667396	3054814	8000366	2000725	0	0	0	0	0 1.40E+07	4667396	3054814	8000366	2000725
40	80 192.168.1	58411		5 11/07/2022 20:1	5 14003256	0	4	0	0	0	0	0	0	0	0	0	0 0	(0.285648	4667752	3055168	8001123	2000802	0	0	0	0	0 1.40E+07	4667752	3055168	8001123	2000802
41	80 192.168.1	58412	6	5 11/07/2022 20:1	5 14001455	0	- 4	0	0	0	0	0	0	0	0	0	0 0	(0.285685	4667152	3055684	8001134	1999848	0	0	0	0	0 1.40E+07	4667152	3055684	8001134	1999848
4	80 192,168,1	58413	6	5 11/07/2022 20:1	5 14000682	0	4	0	0	0	0	0	0	0	0	0	0 0	(0.2857	4666894	3055255	8000308	1999763	0	0	0	0	0 1.40E+07	4666894	3055255	8000308	1999763
43	80 192,168,1	58414	6	5 11/07/2022 20:3	5 14000700	0	4	0	0	0	0	0	0	0	0	0	0 0	0	0.2857	4666900	3055623	8001027	2000071	0	0	0	0	0 1.40E+07	4666900	3055623	8001027	2000071
44	80 192.168.1	58415	6	5 11/07/2022 20:1	5 32082297	1	2	0	0	0	0	0	0	0	0		0 0	0	0.09351	1.60E+07	1.99E+07	3.01E+07	2000718	0	0	0	0	0 2000718	2000718	0	2000718	2000718
4	80 192,168,1	58416	6	5 11/07/2022 20:1	5 46125116	1	2	0	0	0	0	0	0	0	0		0 0	0	0.06504	2.31E+07	2.98E+07	4.41E+07	1999463	0	0	0	0	0 1999463	1999463	0	1999463	1999463
4	80 192 168 1	58749		5 11/07/2022 20:5	2 14303469	0	5	0	0	0	0	0	0	0	0		0 0		0.349566	3575867	3314101	7999970	302148	0	0	0	0	0 1.43E+07	3575867	3314101	7999970	302148
4	80 192 168 1	58541	6	5 11/07/2022 20:1	5 21003939	0	8	0	1323	0	0	0	0	189	0	165.379	66.82159	62,98819	0.380881	3000563	3231148	9600210	600591	0	0	0	0	0 2.10E+07	3000563	3231148	9600210	600591
4	80 192 168 1	58543		5 11/07/2022 20:1	5 21003027	0	8	0	1323	0	0			189	0	165.379	66.82159	62,99090	0 380897	3000432	3231084	9600294	600603	0	0	0	0	0 2 10E+07	3000432	3231084	9600294	600603
100	80 103 168 1	EBEAA		11/07/2022 20:3	5 66736336		2	0						0					0.0/6101	2 215.07	4.405.07	6.416-07	1003036	0	0		-	0 1002006	1003036		1002034	1003036

Figure 13: Sample of raw data transformed to CSV format.

1	Flow ID	22	Flow Pkts/s	43	Fwd Pkts/s	64	Fwd Blk Rate Avg	
2	Src IP	23	Flow IAT Mean	44	Bwd Pkts/s	65	Bwd Byts/b Avg	
3	Src Port	24	Flow IAT Std	45	Pkt Len Min	66	Bwd Pkts/b Avg	
4	Dst IP	25	Flow IAT Max	46	Pkt Len Max	67	Bwd Blk Rate Avg	
5	Dst Port	26	Flow IAT Min	47	Pkt Len Mean	68	Subflow Fwd Pkts	
6	Protocol	27	Fwd IAT Tot	48	Pkt Len Std	69	Subflow Fwd Byts	
7	Timestamp	28	Fwd IAT Mean	49	Pkt Len Var	70	Subflow Bwd Pkts	
8	Flow Duration	29	Fwd IAT Std	50	FIN Flag Cnt	71	Subflow Bwd Byts	
9	Tot Fwd Pkts	30	Fwd IAT Max	51	SYN Flag Cnt	72	Init Fwd Win Byts	
10	Tot Bwd Pkts	31	Fwd IAT Min	52	RST Flag Cnt	73	Init Bwd Win Byts	
11	TotLen Fwd Pkts	32	Bwd IAT Tot	53	PSH Flag Cnt	74	Fwd Act Data Pkts	
12	TotLen Bwd Pkts	33	Bwd IAT Mean	54	ACK Flag Cnt	75	Fwd Seg Size Min	
13	Fwd Pkt Len Max	34	Bwd IAT Std	55	URG Flag Cnt	76	Active Mean	
14	Fwd Pkt Len Min	35	Bwd IAT Max	56	CWE Flag Count	77	Active Std	
15	Fwd Pkt Len Mean	36	Bwd IAT Min	57	ECE Flag Cnt	78	Active Max	
16	Fwd Pkt Len Std	37	Fwd PSH Flags	58	Down/Up Ratio	79	Active Min	
17	Bwd Pkt Len Max	38	Bwd PSH Flags	59	Pkt Size Avg	80	Idle Mean	
18	Bwd Pkt Len Min	39	Fwd URG Flags	60	Fwd Seg Size Avg	81	Idle Std	
19	Bwd Pkt Len Mean	40	Bwd URG Flags	61	Bwd Seg Size Avg	82	Idle Max	
20	Bwd Pkt Len Std	41	Fwd Header Len	62	Fwd Byts/b Avg	83	Idle Min	
21	Flow Byts/s	42	Bwd Header Len	63	Fwd Pkts/b Avg	84	Label	

Figure 14: 84 features extracted for a new dataset using CICFlowMeter 4.0.

1Dst Port21Flow IAT Max41Pkt Len Min61Bwd Byts/b2Protocol22Flow IAT Min42Pkt Len Max62Bwd Pkts/b3Timestamp23Fwd IAT Tot43Pkt Len Mean63Bwd Blk Rate4Flow Duration24Fwd IAT Mean44Pkt Len Std64Subflow Fwd5Tot Fwd Pkts25Fwd IAT Mean44Pkt Len Std64Subflow Fwd6Tot Bwd Pkts26Fwd IAT Max46FIN Flag Cnt66Subflow Bwd7TotLen Fwd Pkts27Fwd IAT Min47SYN Flag Cnt67Subflow Bwd8TotLen Bwd Pkts28Bwd IAT Tot48RST Flag Cnt68Init Fwd Win9Fwd Pkt Len Max29Bwd IAT Mean49PSH Flag Cnt69Init Bwd Win10Fwd Pkt Len Min30Bwd IAT Std50ACK Flag Cnt70Fwd Act Data11Fwd Pkt Len Mean31Bwd IAT Max51URG Flag Cnt71Fwd Seg Siz12Fwd Pkt Len Mean33Fwd PSH Flags53ECE Flag Cnt73Active Max13Bwd Pkt Len Max33Fwd PSH Flags54Down/Up Ratio74Active Max15Bwd Pkt Len Min34Bwd PSH Flags55Pkt Size Avg75Active Max15Bwd Pkt Len Min36Bwd URG Flags55Pkt Size Avg75Active Max								
2Protocol22Flow IAT Min42Pkt Len Max62Bwd Pkts/b /b3Timestamp23Fwd IAT Tot43Pkt Len Mean63Bwd Blk Rate4Flow Duration24Fwd IAT Mean44Pkt Len Std64Subflow Fwd5Tot Fwd Pkts25Fwd IAT Std45Pkt Len Var65Subflow Fwd6Tot Bwd Pkts26Fwd IAT Max46FIN Flag Cnt66Subflow Bwd7TotLen Fwd Pkts27Fwd IAT Min47SYN Flag Cnt67Subflow Bwd8TotLen Bwd Pkts28Bwd IAT Tot48RST Flag Cnt68Init Fwd Win9Fwd Pkt Len Max29Bwd IAT Mean49PSH Flag Cnt69Init Bwd Win10Fwd Pkt Len Maa31Bwd IAT Std50ACK Flag Cnt70Fwd Act Data11Fwd Pkt Len Mean31Bwd IAT Max51URG Flag Cnt71Fwd Seg Siz12Fwd Pkt Len Maa33Fwd PSH Flags53ECE Flag Cnt73Active Mean13Bwd Pkt Len Maa33Fwd PSH Flags53ECE Flag Cnt73Active Max15Bwd Pkt Len Mean35Fwd URG Flags55Pkt Size Avg75Active Max15Bwd Pkt Len Std36Bwd URG Flags55Pkt Size Avg77Idle Mean16Bwd Pkt Len Std36Bwd URG Flags55Fwd Seg Size Avg77	1	Dst Port	21	Flow IAT Max	41	Pkt Len Min	61	Bwd Byts/b Avg
3 Timestamp23 Fwd IAT Tot43 Pkt Len Mean63 Bwd Blk Rate4 Flow Duration24 Fwd IAT Mean44 Pkt Len Std64 Subflow Fwd5 Tot Fwd Pkts25 Fwd IAT Std45 Pkt Len Var65 Subflow Fwd6 Tot Bwd Pkts26 Fwd IAT Max46 FIN Flag Cnt66 Subflow Bwd7 TotLen Fwd Pkts27 Fwd IAT Min47 SYN Flag Cnt67 Subflow Bwd8 TotLen Bwd Pkts28 Bwd IAT Tot48 RST Flag Cnt68 Init Fwd Win9 Fwd Pkt Len Max29 Bwd IAT Mean49 PSH Flag Cnt69 Init Bwd Wir10 Fwd Pkt Len Maa29 Bwd IAT Max51 URG Flag Cnt71 Fwd Seg Siz12 Fwd Pkt Len Maa31 Bwd IAT Max51 URG Flag Cnt72 Active Mean13 Bwd Pkt Len Maa32 Bwd IAT Min52 CWE Flag Count72 Active Mean13 Bwd Pkt Len Maa33 Fwd PSH Flags53 ECE Flag Cnt73 Active Std14 Bwd Pkt Len Min34 Bwd PSH Flags55 Pkt Size Avg75 Active Min15 Bwd Pkt Len Std36 Bwd URG Flags55 Pkt Size Avg76 Idle Mean17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byts/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Rate Avg80 Label	2	Protocol	22	Flow IAT Min	42	Pkt Len Max	62	Bwd Pkts/b Avg
4Flow Duration24Fwd IAT Mean44Pkt Len Std64Subflow Fwd5Tot Fwd Pkts25Fwd IAT Std45Pkt Len Var65Subflow Fwd6Tot Bwd Pkts26Fwd IAT Max46FIN Flag Cnt66Subflow Bwd7TotLen Fwd Pkts27Fwd IAT Min47SYN Flag Cnt67Subflow Bwd8TotLen Bwd Pkts28Bwd IAT Tot48RST Flag Cnt68Init Fwd Win9Fwd Pkt Len Max29Bwd IAT Mean49PSH Flag Cnt69Init Bwd Win10Fwd Pkt Len Min30Bwd IAT Std50ACK Flag Cnt70Fwd Act Data11Fwd Pkt Len Mean31Bwd IAT Max51URG Flag Cnt71Fwd Seg Siz12Fwd Pkt Len Max33Fwd PSH Flags53ECE Flag Cnt73Active Mean13Bwd Pkt Len Min34Bwd PSH Flags53ECE Flag Cnt73Active Max15Bwd Pkt Len Min34Bwd PSH Flags55Pkt Size Avg75Active Max15Bwd Pkt Len Std36Bwd URG Flags55Fwd Seg Size Avg77Idle Mean16Bwd Pkt Len Std36Bwd Header Len57Bwd Seg Size Avg77Idle Max17Flow Byts/s37Fwd Header Len58Fwd Byts/b Avg78Idle Max18Flow Pkts/s38Bwd Hets/s59Fwd Pkts/b Avg79<	3	Timestamp	23	Fwd IAT Tot	43	Pkt Len Mean	63	Bwd Blk Rate Avg
5 Tot Fwd Pkts25 Fwd IAT Std45 Pkt Len Var65 Subflow Fwd6 Tot Bwd Pkts26 Fwd IAT Max46 FIN Flag Cnt66 Subflow Bwd7 TotLen Fwd Pkts27 Fwd IAT Min47 SYN Flag Cnt67 Subflow Bwd8 TotLen Bwd Pkts28 Bwd IAT Tot48 RST Flag Cnt68 Init Fwd Win9 Fwd Pkt Len Max29 Bwd IAT Mean49 PSH Flag Cnt69 Init Bwd Wir10 Fwd Pkt Len Min30 Bwd IAT Std50 ACK Flag Cnt70 Fwd Act Data11 Fwd Pkt Len Mean31 Bwd IAT Max51 URG Flag Cnt71 Fwd Seg Siz12 Fwd Pkt Len Max32 Bwd IAT Min52 CWE Flag Count72 Active Mean13 Bwd Pkt Len Min34 Bwd PSH Flags53 ECE Flag Cnt73 Active Std14 Bwd Pkt Len Min35 Fwd URG Flags55 Pkt Size Avg75 Active Max15 Bwd Pkt Len Std36 Bwd URG Flags56 Fwd Seg Size Avg76 Idle Mean16 Bwd Pkt Len Std38 Bwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byt/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Rate Avg80 I abel	4	Flow Duration	24	Fwd IAT Mean	44	Pkt Len Std	64	Subflow Fwd Pkts
6 Tot Bwd Pkts26 Fwd IAT Max46 FIN Flag Cnt66 Subflow Bwd7 TotLen Fwd Pkts27 Fwd IAT Min47 SYN Flag Cnt67 Subflow Bwd8 TotLen Bwd Pkts28 Bwd IAT Tot48 RST Flag Cnt68 Init Fwd Win9 Fwd Pkt Len Max29 Bwd IAT Mean49 PSH Flag Cnt69 Init Bwd Wir10 Fwd Pkt Len Max29 Bwd IAT Max50 ACK Flag Cnt70 Fwd Act Data11 Fwd Pkt Len Mean31 Bwd IAT Max51 URG Flag Cnt71 Fwd Seg Siz12 Fwd Pkt Len Max32 Bwd IAT Min52 CWE Flag Count72 Active Mean13 Bwd Pkt Len Max33 Fwd PSH Flags53 ECE Flag Cnt73 Active Std14 Bwd Pkt Len Min34 Bwd PSH Flags55 Pkt Size Avg75 Active Max15 Bwd Pkt Len Std36 Bwd URG Flags56 Fwd Seg Size Avg76 Idle Mean16 Bwd Pkt Len Std36 Bwd URG Flags57 Bwd Seg Size Avg77 Idle Std17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Rate Avg80 Label	5	Tot Fwd Pkts	25	Fwd IAT Std	45	Pkt Len Var	65	Subflow Fwd Byts
7 TotLen Fwd Pkts27 Fwd IAT Min47 SYN Flag Cnt67 Subflow Bwd8 TotLen Bwd Pkts28 Bwd IAT Tot48 RST Flag Cnt68 Init Fwd Win9 Fwd Pkt Len Max29 Bwd IAT Mean49 PSH Flag Cnt69 Init Bwd Wir10 Fwd Pkt Len Min30 Bwd IAT Std50 ACK Flag Cnt70 Fwd Act Data11 Fwd Pkt Len Mean31 Bwd IAT Max51 URG Flag Cnt71 Fwd Seg Siz12 Fwd Pkt Len Std32 Bwd IAT Min52 CWE Flag Count72 Active Mean13 Bwd Pkt Len Max33 Fwd PSH Flags53 ECE Flag Cnt73 Active Std14 Bwd Pkt Len Min34 Bwd PSH Flags54 Down/Up Ratio74 Active Max15 Bwd Pkt Len Mean35 Fwd URG Flags55 Pkt Size Avg75 Active Min16 Bwd Pkt Len Std36 Bwd URG Flags57 Bwd Seg Size Avg77 Idle Std17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg78 Idle Mean19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Rate Avg80 Label	6	Tot Bwd Pkts	26	Fwd IAT Max	46	FIN Flag Cnt	66	Subflow Bwd Pkts
8 TotLen Bwd Pkts28 Bwd IAT Tot48 RST Flag Cnt68 Init Fwd Win9 Fwd Pkt Len Max29 Bwd IAT Mean49 PSH Flag Cnt69 Init Bwd Wir10 Fwd Pkt Len Min30 Bwd IAT Std50 ACK Flag Cnt70 Fwd Act Data11 Fwd Pkt Len Mean31 Bwd IAT Max51 URG Flag Cnt71 Fwd Seg Siz12 Fwd Pkt Len Std32 Bwd IAT Min52 CWE Flag Count72 Active Mean13 Bwd Pkt Len Max33 Fwd PSH Flags53 ECE Flag Cnt73 Active Std14 Bwd Pkt Len Min34 Bwd PSH Flags54 Down/Up Ratio74 Active Max15 Bwd Pkt Len Mean35 Fwd URG Flags55 Pkt Size Avg75 Active Min16 Bwd Pkt Len Std36 Bwd URG Flags56 Fwd Seg Size Avg77 Idle Std17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byts/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Rate Avg80 Label	7	TotLen Fwd Pkts	27	Fwd IAT Min	47	SYN Flag Cnt	67	Subflow Bwd Byts
9 Fwd Pkt Len Max29 Bwd IAT Mean49 PSH Flag Cnt69 Init Bwd Wir10 Fwd Pkt Len Min30 Bwd IAT Std50 ACK Flag Cnt70 Fwd Act Data11 Fwd Pkt Len Mean31 Bwd IAT Max51 URG Flag Cnt71 Fwd Seg Siz12 Fwd Pkt Len Std32 Bwd IAT Min52 CWE Flag Count72 Active Mean13 Bwd Pkt Len Max33 Fwd PSH Flags53 ECE Flag Cnt73 Active Std14 Bwd Pkt Len Min34 Bwd PSH Flags54 Down/Up Ratio74 Active Max15 Bwd Pkt Len Mean35 Fwd URG Flags55 Pkt Size Avg75 Active Min16 Bwd Pkt Len Std36 Bwd URG Flags56 Fwd Seg Size Avg77 Idle Std17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byts/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Rate Avg80 Label	8	TotLen Bwd Pkts	28	Bwd IAT Tot	48	RST Flag Cnt	68	Init Fwd Win Byts
10Fwd Pkt Len Min30Bwd IAT Std50ACK Flag Cnt70Fwd Act Data11Fwd Pkt Len Mean31Bwd IAT Max51URG Flag Cnt71Fwd Seg Siz12Fwd Pkt Len Std32Bwd IAT Min52CWE Flag Count72Active Mean13Bwd Pkt Len Max33Fwd PSH Flags53ECE Flag Cnt73Active Std14Bwd Pkt Len Min34Bwd PSH Flags54Down/Up Ratio74Active Max15Bwd Pkt Len Mean35Fwd URG Flags55Pkt Size Avg75Active Min16Bwd Pkt Len Std36Bwd URG Flags56Fwd Seg Size Avg76Idle Mean17Flow Byts/s37Fwd Header Len57Bwd Seg Size Avg77Idle Std18Flow Pkts/s38Bwd Header Len58Fwd Byts/b Avg78Idle Max19Flow IAT Mean39Fwd Pkts/s59Fwd Pkts/b Avg79Idle Min20Flow IAT Std40Bwd Pkts/s60Fwd Blk Rate Avg80Iabel	9	Fwd Pkt Len Max	29	Bwd IAT Mean	49	PSH Flag Cnt	69	Init Bwd Win Byts
11Fwd Pkt Len Mean31Bwd IAT Max51URG Flag Cnt71Fwd Seg Size12Fwd Pkt Len Std32Bwd IAT Min52CWE Flag Count72Active Mean13Bwd Pkt Len Max33Fwd PSH Flags53ECE Flag Cnt73Active Std14Bwd Pkt Len Min34Bwd PSH Flags54Down/Up Ratio74Active Max15Bwd Pkt Len Mean35Fwd URG Flags55Pkt Size Avg75Active Min16Bwd Pkt Len Std36Bwd URG Flags56Fwd Seg Size Avg76Idle Mean17Flow Byts/s37Fwd Header Len57Bwd Seg Size Avg77Idle Std18Flow Pkts/s38Bwd Header Len58Fwd Byts/b Avg78Idle Max19Flow IAT Mean39Fwd Pkts/s59Fwd Pkts/b Avg79Idle Min20Flow IAT Std40Bwd Pkts/s60Fwd Blk Rate Avg80Iabel	10	Fwd Pkt Len Min	30	Bwd IAT Std	50	ACK Flag Cnt	70	Fwd Act Data Pkts
12Fwd Pkt Len Std32Bwd IAT Min52CWE Flag Count72Active Mean13Bwd Pkt Len Max33Fwd PSH Flags53ECE Flag Cnt73Active Std14Bwd Pkt Len Min34Bwd PSH Flags54Down/Up Ratio74Active Max15Bwd Pkt Len Mean35Fwd URG Flags55Pkt Size Avg75Active Min16Bwd Pkt Len Std36Bwd URG Flags56Fwd Seg Size Avg76Idle Mean17Flow Byts/s37Fwd Header Len57Bwd Seg Size Avg77Idle Std18Flow Pkts/s38Bwd Header Len58Fwd Byts/b Avg78Idle Max19Flow IAT Mean39Fwd Pkts/s59Fwd Pkts/b Avg79Idle Min20Flow IAT Std40Bwd Pkts/s60Fwd Blk Rate Avg80Iabel	11	Fwd Pkt Len Mean	31	Bwd IAT Max	51	URG Flag Cnt	71	Fwd Seg Size Min
13 Bwd Pkt Len Max33 Fwd PSH Flags53 ECE Flag Cnt73 Active Std14 Bwd Pkt Len Min34 Bwd PSH Flags54 Down/Up Ratio74 Active Max15 Bwd Pkt Len Mean35 Fwd URG Flags55 Pkt Size Avg75 Active Min16 Bwd Pkt Len Std36 Bwd URG Flags56 Fwd Seg Size Avg76 Idle Mean17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byts/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Rate Avg80 I abel	12	Fwd Pkt Len Std	32	Bwd IAT Min	52	CWE Flag Count	72	Active Mean
14 Bwd Pkt Len Min34 Bwd PSH Flags54 Down/Up Ratio74 Active Max15 Bwd Pkt Len Mean35 Fwd URG Flags55 Pkt Size Avg75 Active Min16 Bwd Pkt Len Std36 Bwd URG Flags56 Fwd Seg Size Avg76 Idle Mean17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byts/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Rate Avg80 I abel	13	Bwd Pkt Len Max	33	Fwd PSH Flags	53	ECE Flag Cnt	73	Active Std
15 Bwd Pkt Len Mean35 Fwd URG Flags55 Pkt Size Avg75 Active Min16 Bwd Pkt Len Std36 Bwd URG Flags56 Fwd Seg Size Avg76 Idle Mean17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byts/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Bate Avg80 I abel	14	Bwd Pkt Len Min	34	Bwd PSH Flags	54	Down/Up Ratio	74	Active Max
16 Bwd Pkt Len Std36 Bwd URG Flags56 Fwd Seg Size Avg76 Idle Mean17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byts/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Blk Bate Avg80 Label	15	Bwd Pkt Len Mean	35	Fwd URG Flags	55	Pkt Size Avg	75	Active Min
17 Flow Byts/s37 Fwd Header Len57 Bwd Seg Size Avg77 Idle Std18 Flow Pkts/s38 Bwd Header Len58 Fwd Byts/b Avg78 Idle Max19 Flow IAT Mean39 Fwd Pkts/s59 Fwd Pkts/b Avg79 Idle Min20 Flow IAT Std40 Bwd Pkts/s60 Fwd Bik Bate Avg80 Label	16	Bwd Pkt Len Std	36	Bwd URG Flags	56	Fwd Seg Size Avg	76	Idle Mean
18 Flow Pkts/s 38 Bwd Header Len 58 Fwd Byts/b Avg 78 Idle Max 19 Flow IAT Mean 39 Fwd Pkts/s 59 Fwd Pkts/b Avg 79 Idle Min 20 Flow IAT Std 40 Bwd Pkts/s 60 Fwd Bik Bate Avg 80 Label	17	Flow Byts/s	37	Fwd Header Len	57	Bwd Seg Size Avg	77	Idle Std
19 Flow IAT Mean 39 Fwd Pkts/s 59 Fwd Pkts/b Avg 79 Idle Min 20 Flow IAT Std 40 Bwd Pkts/s 60 Fwd Blk Rate Avg 80 Label	18	Flow Pkts/s	38	Bwd Header Len	58	Fwd Byts/b Avg	78	Idle Max
20 Flow IAT Std 40 Bwd Pkts/s 60 Fwd Blk Rate Avg 80 Label	19	Flow IAT Mean	39	Fwd Pkts/s	59	Fwd Pkts/b Avg	79	Idle Min
Lo non mi ota do bita ritojo do riva bit ritate Avg do Edber	20	Flow IAT Std	40	Bwd Pkts/s	60	Fwd Blk Rate Avg	80	Label

Figure 15: Open-source dataset 80 extracted features with CICFlowMeter 3.0.

CICFlowMeter tool was used to extract the raw data (Pcap files) to a real dataset into a CSV format with the type of features shown in Figures 14 and 15 which is suitable for building algorithms in an ML system. The next section describes the extracted dataset used for this study.

3.3.2.1 Extracted Dataset Description

This research consists of a newly generated dataset and an open-source dataset CSE-CIC-IDS2018 ('IDS 2018' 2022) consisting of the benign and DDoS classes that make up the binary class label which forms up the class attribute. DDoS is an attack traffic label, and benign is a normal traffic class. The benchmark dataset was deployed in this study for the training of large data sizes since limited time is required to complete this research. Therefore, generating a larger sample data size could not be achieved. However, choosing an open-source dataset for this study was difficult due to the lack of DDoS-specific datasets, despite being one of the most severe security attacks in cloud computing.

1. Newly Generated Dataset

This data consists of 28972 label sample sizes of benign and DDoS traffic flow and 84 features. The major tool used to perform DDoS attacks is slowloris on different terminals seen as systems against the system in the Owncloud.

2. CSE-CIC-IDS2018 Dataset

HTTP denial of service: In this case, the major tools used are Slowloris and LOIC, which have been shown to render Web servers fully unreachable with only one attacking machine. The dataset includes numerous attack types, such as DoS, Infiltration, DDoS, and brute force. Only DDoS attacks are taken into consideration for this study.

The dataset provided both normal and attack traffic consisting of 155047 benign and DDoS labels and 80 features. The open source dataset can be accessed using this link (https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv and https://www.unb.ca/cic/datasets/ids-2018.html).

Two major attributes of the dataset are Class labels and class distribution

Class labels: Class labels is the target features in the features extracted from the dataset. Each dataset element represents a momentary sample of the network activity. These occurrences are classified based on the type of traffic, such as whether it is malicious or benign. The new and benchmark datasets are encoded to maintain homogeneity in the class labelling scheme. Binary classification designates Normal traffic as Benign and malicious traffic as a DDOS attack. Table 8 depicts the class labels of the datasets.

Table 8: Class labels for the datasets

Label	Scenario
Benign	Normal Traffic
DDoS attack	Malicious Traffic

Class distribution records for the new dataset: Each dataset used in this study is presented in the table with the number of records of each class label.

The main data generated for this research Figure 16 shows the number of records of each class label. The percentage records of the class labels were represented using a pie chart as shown in Figure 17.



Figure 16: Class distribution of the new dataset.



Figure 17: New dataset class labels percentage records.

Class Distribution for CSE-CIC-IDS2018 Dataset

For the open-source dataset that is used for larger sample size training, the records of each class label are shown in Figure 18. The percentage of class label records is displayed in Figure 19.



Figure 18: Class distribution for Open-source dataset. https://deepscienceresearch.com



Figure 19: Open-source dataset class labels percentage records.

3.3.3 Feature Selection

This involves choosing necessary features that have a greater impact on the output variable. It implies that we should only choose characteristics (independent variables) that have a strong relationship to the output variable. It is the step that matters the most when building a machine learning model (Swapnilbobe 2021). Since the dataset generated contains lots of features, training the model with those features will impact the accuracy of the model to go down. Therefore, to tackle this issue, only the features that have a greater impact on the dependent variable should be used (the output variable).

3.3.3.1 Correlation Coefficient

The linear relationship between two or more variables is measured using correlation. We can forecast one variable based on another using correlation. Because the desirable variables have a strong correlation with the target, correlation can be used to select features. Variables should also be uncorrelated among themselves while being correlated with the objective. One can estimate one variable from another if the two are correlated. As a result, if two features are correlated, the model only actually requires one of them https://deepscienceresearch.com 37

as the other does not provide any new information (Gupta 2020). The correlation coefficient method was used to detect the features with the highest positive correlation and highest negative correlation. These features were removed to achieve high accuracy.

3.3.3.2 Feature Selection for The New Dataset

Out of 84 features extracted from the new dataset using CICFlowMeter 4.0, 31 features were selected for Model evaluation using correlation coefficient and its heatmap. These features greatly impacted the results achieved in this experiment. Figure 20 illustrates the heatmap of features selected. Also, to remove the duplicate heat map, a masking technique was used as shown in Figure 21. Figure 22 shows the 31 features selected for ML model training, testing and validation.



Figure 20: New dataset feature selection heat map.



Figure 21: Application of Mask on the heat map.

Features = ['Src Port', 'Protocol', 'Tot Fwd Pkts', 'TotLen Fwd Pkts', 'Fwd Pkt Len Max', 'Fwd Pkt Len Mean', 'Bwd Pkt Len Max', 'Bwd Pkt Len Mean', 'Flow Byts/s', 'Flow IAT Mean', 'Flow IAT Max', 'Fwd IAT Tot', 'Fwd IAT Std', 'Fwd IAT Min', 'Bwd IAT Mean', 'Bwd IAT Max', 'Fwd Header Len', 'Fwd Pkts/s', 'Pkt Len Min', 'Pkt Len Mean', 'Pkt Len Var', 'SYN Flag Cnt', 'PSH Flag Cnt', 'Pkt Size Avg', 'Bwd Seg Size Avg', 'Subflow Fwd Byts', 'Subflow Bwd Byts', 'Active Mean', 'Active Max', 'Idle Mean', 'Idle Max']

Figure 22: 31 Features selected for new dataset ML building.

3.3.3 Feature Selection for the CSE-CICIDS2018 Dataset

32 features were selected for model training, testing and validation. Figure 23 and 24 (application of mask) shows the heatmap feature selection. Figure 25 depicts the main feature of the evaluation.







Figure 24: Application of mask on the heat map.

https://deepscienceresearch.com

Figure 25: 32 Main features used for ML model building.

3.3.4 Data Pre-processing

This involves transforming unusable raw data into something that can be used. Apply label encoding to the categorical class label to transform it into discrete form (0,1), where 0 represents a class that is benign, and 1 represents a DDoS attack.

3.3.4.1 Data Cleaning and Transformation

Before using the dataset to build ML models, cleaning the data was essential. For instance, checking strings and negative values can disrupt the smooth running of the process. In this section, cleaning, like checking missing data and removing undefined data, was performed to ensure that it was transformed into a smooth processing dataset.

Missing Data: In machine learning, handling missing data is essential since it can cause any model to make inaccurate predictions. Considering this, null values are removed by moving the most recent valid observation up the column axis. The fillna method from the pandas' library is used to do this (Pandas 2022), as seen in the example below.

```
data.fillna(method ='ffill', inplace = True)
```

Undefined data: Data that has null values removed may be undefinable. After propagation, a null field with no cells on its left becomes NaN since there are no cells to supply a value. These values are therefore decoded as 0. The Fillna approach is used for all of this.

```
data=data.fillna(0)
```

However, no missing values and strings were found in the chosen datasets.

https://deepscienceresearch.com

Transformation: Modeling may not be possible given the format of the data that was collected. In these situations, data and data types must be changed before being fed into the models according to the CRISP-DM technique. Since models do not perform well with strings or do not work at all, several data features were converted into numeric or float.

3.3.5 Dataset Splitting

The datasets were divided into two subsets: training and testing also used for validation. For ML model building, this split was done in an 80:20 ratio, respectively. Where 80% is for training and 20% for testing and validation. The split is performed using the train_test_split helper method from the scikit-learn library. This method divides training into two stages: the training set and the validation set. First, the model is trained using the training set. The model's performance on data that it has not been trained on is then estimated using the validation set. A stratified 5-fold technique was used for validation in the context of this study.

3.3.6 Modelling

The building of the learning model and the creation of the forecasted labels are the two components of the classification phase. Scikit-learn, a Python toolkit for data mining, data analysis, and machine learning, is used to carry out these tasks.

3.3.6.1 Selection of Models

Five models were selected namely Random Forest, Support Vector Machine, Naïve Bayes, Decision Tree and K-Nearest Neighbours for training and validation of the datasets (New and open datasets) used in this experiment. These models were selected based on their performance in intrusion detection.

3.3.7 Training

The chosen algorithms were supplied with training data during the training phase so they can be used to develop machine learning models. The training set is utilised for training. https://deepscienceresearch.com 42 The target attribute was present in the input data source at this stage of the procedure (class label). To map the input characteristics to the target attribute, patterns must be found in the training set. A model is created based on the patterns that were observed. New DDoS datasets and an open-source dataset were used in this study as the input data source, and the target attribute is the nature of network traffic—either an attack or normal traffic—DDoS attack or Benign. The open dataset (CSE-CICIDS2018 DDoS dataset) was deployed and trained since a large sample data size ranging from 100000 to 300000 could not be generated due to the limited time assigned in this research. The data size was extracted from the open-source dataset for training and the results were also compared using the new data generated.

On each of the datasets provided, five algorithms are trained. A variety of techniques from the scikit-learn library are used during training. Table 9 shows the method used for each model. Figure 26 shows the Sample code used for training the dataset.

Model	Scikit-learn Methods & Classifiers
Random Forest	sklearn.ensemble.RandomForestClassifier
SVM	sklearn.svm.LinearSVC
Naive Bayes	sklearn.naive_bayes.GaussianNB
K-NN	sklearn.neighbor.KNeighborsClassifier
Decision Tree	sklearn.tree.DecisionTreeClassifier

Table 9: Scikit-learn Python library classifiers used for building models.



Figure 26: Sample python code for ML model training (Wan n.d. 2019).

3.3.8 Validation

5-fold cross-validation was used to validate the model after training. To evaluate if the model performs were fit for the dataset, cross-validation was used. This approach lessened the overfitting mistakes that happen when a model is too closely fitted to a variety of data samples. Cross-validation is carried out in iterations, with each iteration requiring the division of the dataset into k subsets, or folds. As shown in Figure 27, the model is trained on k-1 folds while the other fold is saved for testing. Up until every fold has functioned as a test fold, this process was repeated. The process is finished by calculating the average value, which was then used to summarise the evaluation metric.



★ Test samples

Figure 27: 5-Fold cross-validation

A stratified k-fold technique is employed in this study with the validation dataset (20% of the whole set). A variant of k-fold cross-validation known as stratified k-fold makes sure that the distribution of classes is uniform across all folds. Utilizing the StratifiedKFold function with k=5 from the scikit-learn library was implemented. Figure 28 shows the sample code for ML model validation.

```
#Cross Validation
from sklearn.model_selection import cross_val_score,KFold
from sklearn.model_selection import cross_val_score,StratifiedKFold

import time
s=time.time()
RF=RandomForestClassifier()
stratifiedkf=StratifiedKFold(n_splits=5)
score=cross_val_score(RF,x,y,cv=stratifiedkf)
res2 = time.time()
print('RF took ',res2-res1,'seconds')
print("Cross Validation Scores are {}".format(score.mean()*100))
```

Figure 28: Sample python code for ML model validation (Wan n.d. 2019).

3.3.9 Testing

The models were tested with imaginary data at the end of the modelling phase. The test set those results from the data split is the unseen data used at this point (20%). Testing is done to evaluate a model's ability to represent data and how well it will function in the future. This study made sure that any model modifications were made before testing, ensuring that testing data would only be utilised once. Accuracy, precision, recall, and

F-measure are the main performance metrics that were created to analyse the performance of the DDoS datasets. Although computation time was included.

3.3.10 Evaluation

Creating performance metrics is an essential step in analysing the effectiveness of the algorithm. Several metrics are generated for this study including:

Accuracy

The number of cases that a classification model correctly and wrongly classifies is one approach to quantifying its performance. A confusion matrix is a popular representation of these values. A tabulated visual representation of the effectiveness of supervised learning systems is called a confusion matrix. Instance counts in actual classes are represented by rows, while instance counts in predictive classes are represented by columns where 0 is Benign traffic and 1 is attack traffic. The confusion matrix for a binary classification task is shown in Figure 29.

		Predicted Class							
		0	1						
Actual Class	0	1000	100						
Actual Class	1	200	50						

Figure 29: Illustration of confusion matric for a binary classifier.

The efficiency of a single classifier can be determined using a confusion matrix. However, combining the matrix's components into a single value is more useful and understandable. In this work, the accuracy measure is used to summarise the matrix and is calculated as follows:

Accuracy = properly *labelled instances/Total instances x100%*

Equation 1: Computation of accuracy

Precision

Accuracy alone usually is insufficient to judge the efficiency of a learning algorithm. Even while accuracy shows whether the model is being trained properly, it does not provide precise information about the application. As a result, other performance criteria, like precision, are used. The proportion of accurately categorised positives, or true positives, is referred to as precision. There are numerous circumstances in which false positives could have an impact. In the context of this study, a high false positive rate suggests that legitimate traffic may be mistakenly classified as harmful. Beyond academics, this may lead to a waste of time and resources.

Precision is calculated as follows:

Precision = True positives/True positives x False positives

Equation 2: Computation of precision

Recall

Recall is an additional performance measure. How many of the real positives were discovered or identified is measured by recall. It is also a very crucial statistic because undiscovered positives or false negatives may have negative impacts in specific circumstances. For example, an algorithm that does not account for all DDoS attack instances means that malicious network traffic will go undetected, increasing the possibility of disruption to the system and the users. Recall is calculated as:

Recall = True positives/True positives x False negatives

Equation 3: Computation of recall

F1- Score/Measure

The F-measure is a metric that combines precision and recall giving a total accuracy score of an algorithm. An algorithm that successfully detects threats while producing a

few false alarms has a high F-measure score because it has low false positives and false negatives. F1-score is calculated as follows:

F1-score = 2 x (Precision x Recall/ Precision + Recall)

Equation 4: Computation of F1-score

Computation Time

Processing time is the last performance metric used in this evaluation. This indicates the length of time a model needs to train rather than being directly tied to classification. This measure provides information about the efficiency of the algorithm being computed within a time frame. The reported processing time was based on a windows 10 operating system with 34GB RAM and an i7 processor.