Article

CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment

Euclides Carlos Pinto Neto, Sajjad Dadkhah*, Raphael Ferreira, Alireza Zohourian, Rongxing Lu, Ali A. Ghorbani

[1] University of New Brunswick (UnB); e.neto@unb.ca, sdadkhah@unb.ca, raphael.ferreira@unb.ca, alireza.zohourian@unb.ca, rlu1@unb.ca, ghorbani@unb.ca

* Correspondence: Sajjad Dadkhah <sdadkhah@unb.ca>

Abstract: Nowadays, the Internet of Things (IoT) concept plays a pivotal role in society and brings new capabilities to different industries. The number IoT solutions in areas such as transportation and healthcare is increasing and new services are under development. In the last decade, society has experienced a drastic increase in IoT connections. In fact, IoT connections will increase in the next few years across different areas. Conversely, despite these benefits, several challenges still need to be faced to enable efficient and secure operations (e.g., interoperability, security, standards, and server technologies). Furthermore, although efforts have been made to produce datasets composed of attacks against IoT devices, several possible attacks are not considered. Most existing efforts do not consider an extensive network topology with real IoT devices. The main goal of this research is to propose a novel and extensive IoT attack dataset to foster the development of security analytics applications in real IoT operations. To accomplish this, 33 attacks are executed in an IoT topology composed of 105 devices. These attacks are classified into seven categories, namely DDoS, DoS, Recon, Web-based, Brute Force, Spoofing, and Mirai. Finally, all attacks are executed by malicious IoT devices targeting other IoT devices.

Keywords: Internet of Things (IoT); Dataset; Security; Machine Learning; Deep Learning; DoS; DDoS; Reconnaissance; Web Attacks; Brute Force; Spoofing; Mirai;

1. Introduction

Nowadays, the Internet of Things (IoT) concept plays a pivotal role in society and brings new capabilities to different industries [1] [2] [3]. IoT projects in areas such as transportation and healthcare are becoming more popular and new applications are under development [4]. This new paradigm relies on an extensively connected sensor network with multiple devices producing network traffic [5] [6] [7]. The research community and industrial labs have been evolving this concept for years and these devices are becoming more present in our daily lives [8] [9] [10].

Several areas have been transformed by this technology. For example, in healthcare applications, patients can be regularly monitored using IoT technology [11] [12] [13]. In transportation, IoT devices have been used to detect and prevent accidents [14] [15] [16]. Industrial IoT (IIoT) has also brought different solutions, such as high reliability and low latency automated monitoring and collaborative control [17]. IoT applications have also been developed for areas such as education [18], aviation [19], and forestry [20]. In the last decade, society has experienced a drastic increase in IoT connections [21]. In fact, IoT connections will increase in the next few years across different areas [22]. This motivates the creation and development of business ideas and new concepts that rely on a highly distributed infrastructure. Besides, various strategies have been proposed to solve potential problems in IoT operations, i.e., the deployment of new services is leveraged by the scientific findings achieved in the past few years.

Conversely, despite these benefits, several challenges still need to be faced to enable efficient and secure operations (e.g., interoperability, security, standards, and server technologies) [23] [24] [25]. The development of new applications may also bring new requirements to the systems. For example, the Internet of Vehicles (IoV) may require more restrictive response times than usual IoT applications [26]. Also, detecting and mitigating attacks performed against IoT devices is challenging due to several factors (e.g., distributed connections and light devices without security mechanisms) [27] [28] [29].

Furthermore, although efforts have been made to produce datasets composed of attacks against IoT devices, several possible attacks are not considered. Besides, most efforts do not consider an extensive network topology with real IoT devices. Finally, the attacks performed against IoT devices are executed by computer systems, highlighting the need for a dataset composed of attacks performed by malicious IoT devices. To enable the development of security analytics solutions for real-world scenarios, the data produced needs to (i) include a variety of attacks that can harm IoT operations, (ii) be collected from an extensive topology with real IoT devices of different types and brands, and (iii) include attacks performed by malicious IoT devices.

The main goal of this research is to propose a novel and extensive IoT attack dataset to foster the development of security analytics applications in real IoT operations. To accomplish this, 33 attacks are executed in an IoT topology composed of 105 devices. These attacks are classified into seven categories, namely DDoS, DoS, Recon, Web-based, Brute Force, Spoofing, and Mirai. Besides, all attacks are executed by malicious IoT devices targeting other IoT devices. This dataset includes multiple attacks not available in other IoT datasets and enables IoT professionals to develop new security analytics solutions. Furthermore, the data is available in different formats, allowing researchers to use features extracted in our evaluation or engineer new features.

The main contributions of this research are:

- We design a new realistic IoT attack dataset, using an extensive topology compose of several real IoT devices and adopting IoT devices as attackers and victims;
- We perform, document, and collect data from 33 attacks divided into 7 classes against IoT devices and demonstrated how they can be reproduced;
- We evaluate the performance of machine and deep learning algorithms using the CICIoT2023 dataset to classify and detect IoT network traffic as malicious or benign.

This paper is organized as follows: Section 2 presents an extensive comparison of the contributions of this research with other works present in the literature. Secondly, Section 3 introduces the CICIoT2023 dataset and presents the steps involved in the data collection. After that, Section 4 presents the feature extraction process and describes the data. Section 5 presents the Machine Learning (ML) evaluation in the classification of different attacks using the CICIoT2023 dataset. Finally, Section 6 presents the conclusion of this research.

2. Related Works

In the past few years, different contributions have been published regarding IoT security datasets. In fact, data has been produced with different goals and using different methods and resources. To better understand the characteristics of existing datasets, we review several initiatives present in the literature and compare them with the proposed CICIoT2023. The authors in [30] propose a novel network-based dataset for detecting botnet attacks in the IoT environment called N-BaioT (2018). Mirai and BASHLITE botnets were used to attack nine commercial IoT devices. Multiple features were extracted from the network traffic and used by a deep-learning autoencoder for attack detection. In [31], the authors introduce a host-based IoT dataset composed of data from real IoT devices. This dataset, called IoTHIDS (2018), is produced based on experiments considering a topology of three devices infected by Mirai, Hajime, Adira, BASHLITE, Doflo, Tsunami, and Wroba malware botnets.

IoT-SH (2019) [32] is a dataset composed of captures of twelve attacks (categorized into four classes) against eight different smart home devices. A three-layer Intrusion Detection System (IDS) is used considering various combinations of rule-based and machine learning approaches to classify the attacks. BoT-Iot (2019) is introduced in [33] as a realistic traffic dataset, produced considering heterogeneous network profiles. Multiple attacks are performed (e.g., DDoS, DoS, data theft, and scan) against five devices. In the evaluation process, a set of new features are selected and used based on Correlation Coefficient and Joint Entropy techniques. Various machine and deep learning models are trained to evaluate the attack detection accuracy.

The authors in [34] introduce the Kitsune (2019) dataset, which is composed of four different categories of attacks executed against nine IoT devices. In the experiments conducted, a security camera was infected by a real Mirai botnet sample. This dataset is intended to support the development of plug-and-play Network Intrusion Detection Systems (NIDS) to detect normal and malicious traffic. Similarly, IoTNIDS (2019) [35] represents an initiative focused on collecting data from a real-world IoT networking environment based on the interaction between two IoT devices (speaker and camera). Multiple attacks are analyzed in this effort, e.g., Mirai, MITM, DoS, and scanning. MedBIoT (2020) [36] is an IoT network architecture dataset based on using real and emulated devices. The authors evaluated multiple machine learning techniques using 100 statistical features extracted from the IoT network traffic. In [37], the authors propose the IoT-23 (2020) dataset. This contribution refers to a botnet dataset captured composed of real network environment captures of benign and malicious traffic.

IoTIDs (2020) [38] is proposed as a dataset composed of IoT-related flow-based features, selected and ranked by the correlation coefficients technique and the Shapira-Wilk algorithm, respectively. In the experiments, the authors performed four different attacks against two IoT devices (speaker and camera) and recorded the data. Multiple machine learning methods were used in the evaluation process (e.g., SVM, G-NB, LDA, and LR) focusing on attack detection and classification. The authors in [39] present the MQTT (2020) dataset with the primary goal of providing realistic data that includes a protocol dedicated to IoT network scenarios. Furthermore, eight IoT devices were connected to the MQTT broker and a set of 33 different features were extracted and provided to various machine learning algorithms. Similarly, MQTT-IoT-IDS (2020) [40] is another contribution focused on producing a dataset using a lightweight protocol, i.e., MQTT, which is used in IoT networks. The authors focus on replicating a realistic IoT network by using a camera feed, twelve MQTT sensors, and a broker. Five scenarios are considered based on the variation of the attacks performed. Several packet-based, uni- and bi-flow features are used alongside six different machine learning algorithms in the evaluation phase.

In [41], the authors proposed a new telemetry-based data-driven IoT/IIoT dataset called TON-IoT (2020). This heterogeneous dataset comprises both normal and attack samples captured in different scenarios. Targeting the development of a realistic dataset, the authors include attack sub-categories, data recorded from operating system logs, and network traffic. Several machine learning and deep learning algorithms are used in the evaluation phase and the achieved results are reported in detail. Finally, the Edge-IIoTSet (2022) dataset is introduced as a realistic cybersecurity resource for IoT and IIoT applications to enable the development of Intrusion Detection Systems (IDS) in centralized and distributed applications [42]. Throughout the paper, an in-depth description of the testbed used is presented. Besides, the authors also describe the dataset generation framework. Regarding the machine learning evaluation process, considerations of centralized and federated learning are presented. This dataset is focused on including attacks not present in previous datasets.

Table 1 compares all datasets reviewed with the proposed CICIoT2023 dataset. This analysis is performed considering the attacks executed in this research, i.e., these datasets may include attacks other than those showed in these tables.

Table 1. Comparison CICIoT2023 with existing IoT security datasets.

	Attack	IoTHIDS	N-BaIoT	Kitsune	IoTNIDS	IoT-SH	BoT-IoT	MedBIoT	IoT-23 (2020)	IoTIDS	MQTT	MQTT-IoT-IDS	X-IIoTID	WUSTL-IIoT	Edge-IIoTSet	CICIoT2023
DDoS	ACK Fragmentation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	UDP Flood	-	✓	-	✓	-	✓	-	-	-	-	-	-	-	✓	✓
	SlowLoris	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	ICMP Flood	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓
	RSTFIN Flood	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	PSHACK Flood	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	HTTP Flood	-	✓	-	✓	-	✓	-	-	-	-	-	-	-	✓	✓
	UDP Fragmentation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	ICMP Fragmentation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	TCP Flood	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	✓
	SYN Flood	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	✓
	SynonymousIP Flood	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
DoS	TCP Flood	-	✓			✓	✓	-	-	-	-	-	-	-	-	✓
	HTTP Flood	-	✓	-	-	-	✓	-	-	✓	-	-	-	-	-	✓
	SYN Flood	-	✓	✓	✓	-	-	-	-	✓	-	-	-	-	-	✓
	UDP Flood	-	✓	-	-	✓	✓	-	-	✓	-	-	-	-	-	✓
Recon	Ping Sweep	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	OS Scan	-	-	✓	✓	✓	✓	-	-	✓	-	✓	✓	-	✓	✓
	Vulnerability Scan	-	✓	-	-	-	-	-	-	-	-	✓	✓	-	✓	✓
	Port Scan	-	✓	-	✓	✓	✓	-	-	✓	-	✓	✓	-	✓	✓
	Host Discovery	-	-	-	✓	-	-	-	-	-	-	-	✓	-	-	✓
Web-Based	Sql Injection	-	-	-	-	-	-	-	-	-	-		-	-	✓	✓
	Command Injection	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	Backdoor Malware	-	-	-	-	-	-	-	-	-	-	-	✓	-	✓	✓
	Uploading Attack	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓
	XSS	-	-	-	-	-	-	-	-	-	-		-	-	✓	✓
	Browser Hijacking	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
Brute Force	Dictionary Brute Force	-	-	-	✓	-	-	-	-	-	✓	✓	✓	-	✓	✓
Spoofing	Arp Spoofing	-	-	✓	✓	✓				✓	-	-	✓		✓	✓
	DNS Spoofing	-	-	-	-	✓	-	-	-	-	-	-	-	-	✓	✓
Mirai	GREIP Flood	✓	✓	✓	✓	-	-	✓	✓	✓	-	-			-	✓
	Greeth Flood	✓	✓	✓	✓	-	-	✓	✓	✓	-	-	-	-	-	✓
	UDPPlain	✓	✓	✓	✓	-	-	✓	✓	✓	-	-	-	-	-	✓

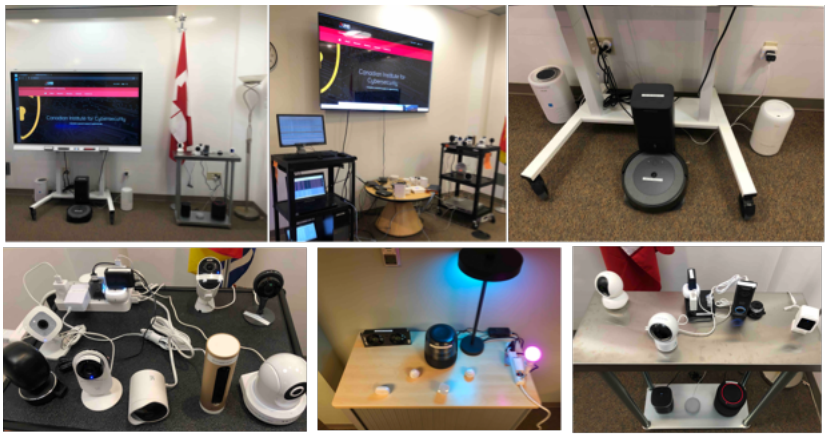

Figure 1. CIC IoT Lab.

3. The Proposed CICIoT2023

This Section introduces the CICIot2023 dataset. We aim to present an in-depth description of all steps and resources involved in producing this dataset. First, we describe the CIC IoT Lab. Then, we focus on the IoT topology, listing all IoT and network devices used and how they are connected. Then, we present a discussion on all attacks that have been executed. Finally, we provide insights into how the data was collected for benign and malicious scenarios.

3.1. IoT Lab

The production of IoT security data that can be used to support real applications is challenging for several reasons. One of the main problems is having an extensive network composed of several real IoT devices, similar to topologies of real IoT applications. Many works adopt simulated or very few IoT devices due to costs, network equipment required (e.g., switches, routers, and network tap), and personnel dedicated to maintaining such an infrastructure.

Thereupon, the Canadian Institute for Cybersecurity (CIC) has a distinguished presence in the cybersecurity ecosystem and a history of high-impact contributions to industry and academia. Examples are datasets used to develop new cybersecurity applications and several partnerships with industry to improve the cybersecurity practice and develop new solutions. This success enabled CIC to establish an IoT lab with a dedicated network to enable the development of IoT security solutions. In fact, by sharing the data collected from this extensive topology, we intend to foster the advancement of IoT security research and support several initiatives in different IoT security aspects.

Figure 1 shows the IoT lab in CIC and its devices. Indeed, IoT devices are distributed across the lab, in which some of them are placed on the table, others on the floor, and some on the walls. We adopt a local network topology and several power plugs are available in the lab. Besides, there are racks and storage rooms in order to organize the IoT and network devices.

3.2. IoT Topology

The IoT topology deployed to produce the CICIoT2023 is illustrated in Figure 2, and comprises 105 IoT devices. 67 IoT devices were directly involved in the attacks and other 38 Zigbee and Z-Wave devices were connected to 5 hubs.

These devices include, including smart home devices, cameras, sensors, and microcontrollers, are connected and configured to enable the execution of several attacks and capturing the corresponding attack traffic. The lab is also equipped with various tools and softwares, which enables us to perform several attacks and capture both benign and malicious attack traffic.

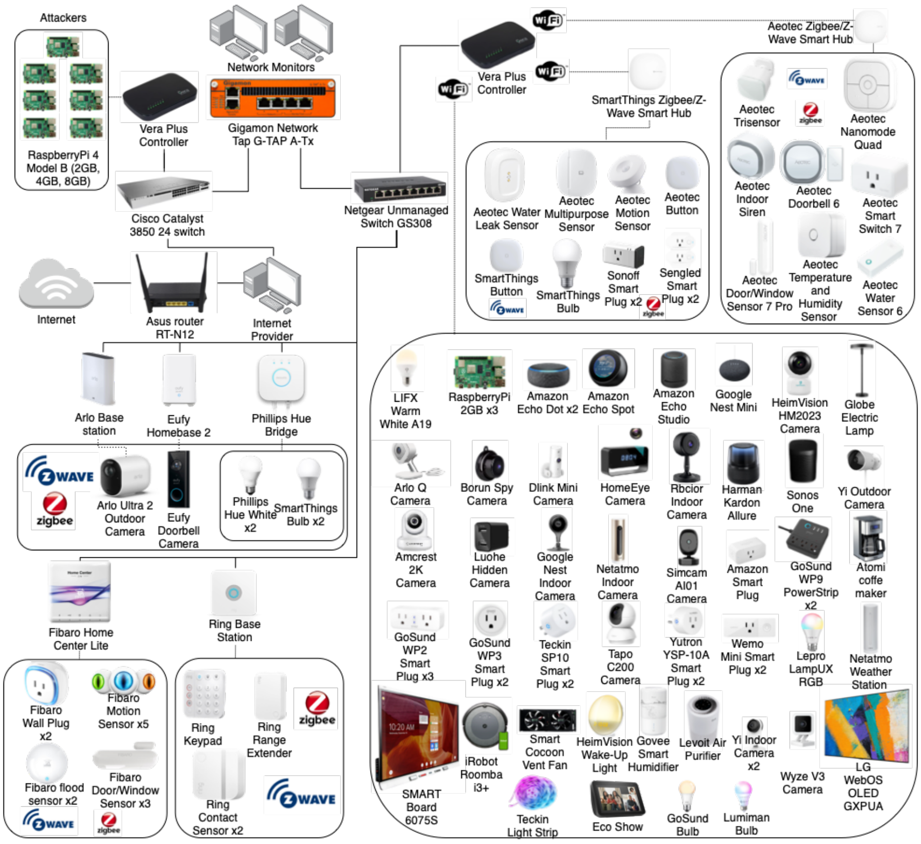

Figure 2. IoT network topology used in the experiments.

This topology is divided into two parts. In the first part, an ASUS router connects the network to the Internet and a Windows 10 Desktop computer shares this connectivity. Besides, a Cisco switch in placed between this computer and a VeraPlus access point connecting 7 RaspberryPi devices. These devices are responsible for executing the attacks and malicious activities in the experiments. Then, the cisco switch is connected to the second part through a Gigamon Network Tap. This network device collects all the IoT traffic and sends it to two network monitors, which are responsible for storing the traffic using wireshark. In fact, a network tap is a hardware device that allows for monitoring and analyzing network traffic by connecting to a network cable and providing a copy of the traffic to other monitoring and security tools. Network taps are connected in a way not to affect the normal operation and provide a full-duplex, non-intrusive, and passive way of accessing network traffic, without introducing any latency or affecting the performance of the network. This device has two network and two monitoring ports and is placed between the attackeing and legitimate devices, connecting one port to the attackers and the other to the victim networks. Using the monitor ports, we are able to capture the traffic to and from the IoT network.

In the second part, a Netgear Unmanneged Switch is connected to 5 gateways and base stations to enable the communication with IoT devices with protocols such as Zigbee and Z-Wave. Furthermore, another VeraPlus controller is connected to the switch. This controller is also connected to other 2 Zigbee/Z-Wave hubs and to several devices considered as victims in the attacks performed. The list of all IoT devices used in this dataset is presented in Table 2.

3.3. Data Collection of Benign and Malicious Scenarios

As described in Section 3.2, a network tap is dedicated to monitoring the network traffic. Every packet sent through the network is stored in separate computers. In fact, the

Table 2. List of IoT devices used to produce the dataset.

	Device Name	Category	MAC Address	Device Name	Category	MAC Address
Victms	Amazon Alexa Echo Dot 1	Audio	1C:FE:2B:98:16:DD	Lumiman bulb	Lighting	84:E3:42:42:ED:0B
	Amazon Alexa Echo Dot 2	Audio	A0:D0:DC:C4:08:FF	Philips Hue Bridge	Hub	00:17:88:60:D6:4F
	Amazon Alexa Echo Spot	Audio	1C:12:B0:9B:0C:EC	Smart Board	Home Automation	00:02:75:F6:E3:CB
	Amazon Alexa Echo Studio	Audio	08:7C:39:CE:6E:2A	Teckin Light Strip	Lighting	18:69:D8:EB:D4:3E
	Amazon Echo Show	Audio	2C:71:FF:05:F1:15	Teckin Plug 1	Power Outlet	D4:A6:51:76:06:64
	Google Nest Mini Speaker	Audio	CC:F4:11:9C:D0:00	Teckin Plug 2	Power Outlet	D4:A6:51:78:97:4E
	harman kardon (Ampak Technology)	Audio	B0:F1:EC:D3:E7:98	Wemo smart plug 1 (Wemo id: Wemo.Mini.AD3)	Power Outlet	30:23:03:F3:84:2B
	Sonos One Speaker	Audio	48:A6:B8:F9:1B:88	Wemo smart plug 2 (Wemo id: Wemo.Mini.4A3)	Power Outlet	30:23:03:F3:57:CB
	AMCREST WiFi Camera	Camera	9C:8E:CD:1D:AB:9F	Yutron Plug 1	Power Outlet	D4:A6:51:20:91:D1
	Arlo Base Station	Camera	3C:37:86:6F:B9:51	Yutron Plug 2	Power Outlet	D4:A6:51:21:6C:29
	Arlo Q Indoor Camera	Camera	40:5D:82:35:14:C8	LG Smart TV	Home Automation	AC:F1:08:4E:00:82
	Borun/Sichuan-AI Camera	Camera	C0:E7:BF:0A:79:D1	Netatmo Weather Station	Home Automation	70:EE:50:6B:A8:1A
	DCS8000LHA1 D-Link Mini Camera	Camera	B0:C5:54:59:2E:99	Raspberry Pi 4 - 2GB	NextGen	DC:A6:32:C9:E6:F4
	HeimVision Smart WiFi Camera	Camera	44:01:BB:EC:10:4A	Raspberry Pi 4 - 2GB	NextGen	DC:A6:32:C9: E4:C6
	Home Eye Camera	Camera	34:75:63:73:F3:36	Raspberry Pi 4 - 2GB	NextGen	DC:A6:32:C9:E5:02
	Luohe Cam Dog	Camera	7C:A7:B0:CD:18:32	Fibaro Door/Window Sensor 1	Sensor	N/A
	Nest Indoor Camera	Camera	44:BB:3B:00:39:07	Fibaro Door/Window Sensor 2	Sensor	N/A
	Netatmo Camera	Camera	70:EE:50:68:0E:32	Fibaro Door/Window Sensor 3	Sensor	N/A
	Rbcior Camera	Camera	10:5A:17:97:A5:C6	Fibaro Flood Sensor 1	Sensor	N/A
	SIMCAM 1S (AMPAKTec)	Camera	10:2C:6B:1B:43:BE	Fibaro Flood Sensor 2	Sensor	N/A
	TP-Link Tapo Camera	Camera	6C:5A:B0:44:1D:90	Fibaro Motion Sensor 1	Sensor	N/A
	Wyze Camera	Camera	7C:78:B2:86:0D:81	Fibaro Motion Sensor 2	Sensor	N/A
	Yi Indoor Camera	Camera	84:7A:B6:64:62:58	Fibaro Motion Sensor 3	Sensor	N/A
	Yi Indoor 2 Camera	Camera	84:7A:B6:62:3A:6C	Fibaro Motion Sensor 4	Sensor	N/A
	Yi Outdoor Camera	Camera	2C:D2:6B:66:D2:87	Fibaro Motion Sensor 5	Sensor	N/A
	Eufy HomeBase 2	Hub	8C:85:80:6C:B6:47	Fibaro Wall Plug 1	Power Outlet	N/A
	Amazon Plug	Power Outlet	B8:5F:98:D0:76:E6	Fibaro Wall Plug 2	Power Outlet	N/A
	Atomi Coffee Maker	Home Automation	68:57:2D:56:AC:47	Ring Alarm Keypad	Home Automation	N/A
	Cocoon Smart HVAC Fan	Home Automation	08:3A:F2:1F:BC:68	Ring Range Extender	Home Automation	N/A
	Globe Lamp ESP_B1680C	Lighting	50:02:91:B1:68:0C	Ring Contact Sensor (1)	Sensor	N/A
	GoSund Bulb	Lighting	C4:DD:57:13:07:C6	Ring Contact Sensor (2)	Sensor	N/A
	Gosund Power strip (1)	Power Outlet	50:02:91:1A:CE:E1	AeoTec TriSensor	Sensor	N/A
	GoSund Power strip (2)	Power Outlet	B8:F0:09:03:9A:AF	AeoTec Doorbell 6	Home Automation	N/A
	GoSund Smart plug WP2 (1)	Power Outlet	B8:F0:09:03:29:79	AeoTec Indoor Siren	Home Automation	N/A
	GoSund Smart Plug WP2 (2)	Power Outlet	50:02:91:10:AC:D8	AeoTec Smart Switch 7	Home Automation	N/A
	GoSund Smart plug WP2 (3)	Power Outlet	50:02:91:10:09:8F	AeoTec Water Sensor 6	Sensor	N/A
	GoSund Smart Plug WP3 (1)	Power Outlet	C4:DD:57:0C:39:94	AeoTec NanoMote Quad	Home Automation	N/A
	Gosund Smart Plug WP3 (2)	Power Outlet	24:A1:60:14:7F:F9	AeoTec Door/Window Sensor 7 Pro	Sensor	N/A
	Govee Smart Humidifer	Home Automation	D4:AD:FC:29:C8:A2	AeoTec Temperature and Humidity Sensor	Sensor	N/A
	HeimVision SmartLife Radio/Lamp	Lighting	D4:A6:51:30:64:B7	Philips Hue White 1	Lighting	N/A
	iRobot Roomba	Home Automation	50:14:79:37:80:18	Philips Hue White 2	Lighting	N/A
	LampUX RGB	Lighting	F4:CF:A2:34:48:6B	SmartThings Smart Bulb 1	Lighting	N/A
	Levoit Air Purifier	Home Automation	1C:9D:C2:8C:9A:94	SmartThings Smart Bulb 2	Lighting	N/A
	LIFX Lightbulb	Lighting	D0:73:D5:35:FB:C8	Aeotec Button	Home Automation	N/A
	SmartThings Hub	Hub	28:6D:97:7A:2B:2D	AeoTec Motion Sensor	Sensor	N/A
	AeoTec Smart Home Hub	Hub	28:6D:97:9E:F4:D5	AeoTec Multipurpose Sensor	Sensor	N/A
	Sengled Smart Plug 2	Power Outlet	N/A	AeoTec Water Leak Sensor	Sensor	N/A
	SmartThings Button	Home Automation	N/A	Sengled Smart Plug 1	Power Outlet	N/A
	SmartThings Smart Bulb 3	Lighting	N/A	Sonoff Smart Plug 2	Power Outlet	N/A
	Sonoff Smart Plug 1	Power Outlet	N/A	Arlo Ultra 2 Outdoor Camera	Camera	N/A
Attackers	Raspberry Pi 4 - 4GB	NextGen	E4:5F:01:55:90:C4	Raspberry Pi 4 - 2GB	NextGen	DC:A6:32:C9:E4:D5
	Raspberry Pi 4 - 8GB	NextGen	DC:A6:32:DC:27:D5	Raspberry Pi 4 - 2GB	NextGen	DC:A6:32:C9:E5:EF
	Raspberry Pi 4 - 2GB	NextGen	DC:A6:32:C9:E4:AB	Raspberry Pi 4 - 2GB	NextGen	DC:A6:32:C9:E4:90
	Raspberry Pi 4 - 2GB	NextGen	DC:A6:32:C9:E5:A4	Ring Base Station	Hub	B0:09:DA:3E:82:6C
	Fibaro Home Center Lite	Hub	AC:17:02:05:34:27	Eufy Doorbell Camera	Camera	N/A

network has two different interfaces, which are associated with two other monitoring ports that send incoming packets to these computers. Hence, the network traffic is monitored using Wireshark [43] and store in .pcap format. Since two data streams are stored, mergecap [44] is used to unify .pcap files for each experiment.

For each attack, a different experiment is performed targeting all applicable devices. In all scenarios, the attacks are performed by malicious IoT devices targeting vulnerable IoT devices. For example, DDoS attacks are executed against all devices, while web-based attacks target devices that support web applications. Table 3 depicts the tools used to perform all attacks.

3.3.1. Benign Data Generation

This represents legitimate use of the IoT network. In this sense, the main goal of the data-capturing procedure relies on gathering IoT traffic in idle states or with human interactions (e.g., sensor data, echo dot requests, and accessing video feeds from smart cameras).

3.3.2. Executing DoS and DDoS attacks

These attacks refer to flooding threats to compromise the availability of IoT operations. In the case of Denial-of-Service (DoS) attacks, one Raspberry Pi is responsible for flooding IoT devices. Furthermore, multiple Raspberry Pi's are used to execute Distributed Denial-of-Service (DDoS) attacks through an ssh-based master-client configuration. The attacks executed are:

- **ACK Fragmentation:** a relatively small number of maximum-sized packets is used to compromise the network operation. In many cases, these fragmented packets are successfully sent and handled by routers, firewalls, and intrusion prevention systems, given that fragmented packets recompilation is not performed [45];
- **Slowloris:** relies on using partial HTTP requests via open connections to a targeted Web server focusing on the application layer [46];
- **ICMP/HTTP/UDP/TCP Flood:** based on overwhelming a targeted device with different packet types [47] [48] [49];
- **RST-FIN Flood:** degrades networking capabilities by forwarding continuously RST-FIN packets towards a specific target [50];
- **PSH-ACK Flood:** degrades server operation by flooding using PUSH and ACK requests [51];
- **UDP Fragmentation:** refers to a special UDP flood that consumes more bandwidth while reducing the number of packets [52];
- **ICMP Fragmentation:** relies on the use of identical fragmented IP packets containing a portion of a fragmented ICMP message [53];
- **SYN Flood:** A SYN flood, on the other hand, is a specific type of TCP flood that targets the initial handshake of the TCP connection. The SYN flood sends a large number of SYN (synchronize) packets to the targeted server, but it never completes the handshake by sending the final ACK (acknowledge) packet [54];
- **Synonymous IP Flood:** an extensive number of manipulated TCP-SYN packets with source and destination addresses as the targeted address, which leads the server to use its resources to process the incoming traffic [55].

3.3.3. Gathering Information from the IoT Topology

These attacks gather all possible information about the target. Besides, an attacker can use a reconnaissance attack as a preparation step for other attacks. There are multiple ways to perform these attacks, and some of the most popular and threatening variations are:

- **Ping Sweep:** A ping sweep attack, also known as a ping scan, is a type of reconnaissance attack used to identify active hosts on a network. It involves sending a series of ICMP (Internet Control Message Protocol) Echo Request (ping) packets to a range of IP addresses on a network, and then analyzing the ICMP Echo Reply (pong) packets that are returned to identify which hosts are active and responding [56];

- **OS Scan:** An OS (operating system) scan attack, also known as an operating system fingerprinting attack, is a type of reconnaissance attack that is used to identify the type and version of an operating system running on a targeted host. The attacker uses various techniques to gather information about the targeted host, such as analyzing the responses to network packets, or examining the behavior of open ports and services, in order to determine the type and version of the operating system [57];
- **Vulnerability Scan:** A vulnerability scan attack is a type of network security assessment that involves automated tools to identify potential vulnerabilities in a computer system or network. The goal of a vulnerability scan is to identify security weaknesses that could be exploited by an attacker to gain unauthorized access to a system or steal sensitive information [58];
- **Port Scan:** A port scan attack is a type of reconnaissance attack that is used to identify open and active ports on a targeted host. The attacker sends a series of packets to various ports on the targeted host, attempting to establish a connection. The responses to these packets are then analyzed to determine which ports are open, closed, or filtered. [59]
- **Host Discovery:** A host discovery attack, also known as a host identification or host enumeration attack, is a type of reconnaissance attack that is used to identify active hosts on a network. It involves using various techniques to identify the IP addresses of devices that are connected to a network, and it is the first step in many cyber-attacks. [60]

3.3.4. Exploiting Web-Based vulnerabilities

To execute these attacks, web services running on IoT devices were targeted. Web-based attacks are concerned with targeting web services in several ways. These attack types include injection, hijacking, poisoning, spoofing, and DoS [61]. The web-based attacks executed in this research are:

- **SQL Injection:** an attack that targets web applications by injecting malicious SQL code into the application's input fields. The goal of an SQL injection attack is to gain unauthorized access to a database, steal sensitive information, or execute arbitrary commands on the database server [62];
- **Command Injection:** an attack that targets web applications by injecting malicious commands into an input field with the ultimate goal of gaining unauthorized access to a system, stealing sensitive information, or executing arbitrary commands on the targeted system [63];
- **Backdoor Malware:** involves installing malware on a targeted system that allows the attacker to gain unauthorized access to the system at a later time. The malware, known as a "backdoor," creates a hidden entry point into the system that can be used to bypass security measures and gain access to sensitive information or perform malicious actions [64];
- **Uploading Attack:** targets a web application by exploiting vulnerabilities in the application's file upload functionality. The goal of an uploading attack is to upload malicious files, such as malware, to a targeted system and use them to gain unauthorized access or execute arbitrary code on the targeted system;
- **Cross-Site Scripting (XSS):** allows an attacker to inject malicious code (e.g., a script) into a web page. The injected script can then be executed by the web browser of any user with access to the page, allowing the attacker to steal sensitive information (e.g., cookies, session tokens, and personal data) or to perform other malicious activities (e.g., traffic redirection) [65];

- **Browser Hijacking:** a type of cyber attack in which an attacker modifies a web browser's settings, such as the home page, default search engine, or bookmarks, in order to redirect the user to a different website or display unwanted ads. The goal of a browser hijacking attack is to generate revenue through advertising or to steal personal information [66].

3.3.5. Spoofing communication

Spoofing attacks enable malicious actors to operate as a victim system and gain illegitimate access to the network traffic. The main focus of such a procedure includes gaining access to systems, stealing data, and spreading malware [67]. Two of the most popular spoofing attacks are:

- **Arp Spoofing:** relies on the transmission of manipulated ARP (Address Resolution Protocol) messages to associate a MAC address with the IP address of other devices in the network. This enables attackers to intercept, modify, or block network traffic [68];
- **DNS Spoofing:** relies on the alteration of DNS entries in a DNS server's cache, redirecting users to manipulated or malicious websites. This enables attackers to steal sensitive information, spread malware, and perform other malicious actions [69].

3.3.6. Brute force threats

Brute-force attacks consist of the submission of data (e.g., passwords or passphrases) to eventually gain access to systems [70]. Among the several procedures that can be executed, a dictionary brute force attack is a type of attack that attempts to guess a password or passphrase by repeatedly trying words from a pre-defined list of words obtained from various sources. The goal of the attack is to find the correct password by trying all the words in the dictionary [71].

3.3.7. Mirai as an IoT threat

The Mirai attack is a large-scale DDoS that can target IoT devices. In this paper, we are conducting different variations of Mirai attacks by using five different raspberries as illustrated in Figure 3 alongside the connections considered in the different IoT network layers. In order to connect to the Internet, a gateway uses a Windows 10 instance to provide and monitor Internet access. This access is possible through a netgear unmanaged switch that connects attackers and general IoT devices. Several tools are used to perform the attacks and a special Mirai configuration is also adopted. Finally, an online IoT supervisor coordinates the operation of the multiple IoT devices in the topology (e.g., sensors, cameras, and smart speakers).

This attack infected devices to form a botnet that can flood targeted victims. This threat can cause disruption in different contexts and some of its most popular variations are:

- **GREIP:** Within GRE packet, this attack floods the target system with encapsulated packets. The internal data comprises random IPs and ports, whereas the external layer contains actual IPs [72];
- **GREETH:** This attack presents a similar procedure to GREIP. However, the main focus is on the packet encapsulation approach, which is based on the ethernet header [72];
- **UDP Plain:** This threat focuses on flooding targeted victim systems with UDP packets considering a repeated packet segment. However, the payload sent is different for each packet [72].

4. Feature Extraction & Data Description

The CICIoT2023 dataset is available in two different formats: pcap, and csv. Pcap files comprise the original data generated and collected in the CIC IoT network in

Table 3. CICIoT2023: Tools and frameworks used to execute attacks.

	Attack	Size	Tool
DDoS	ACK Fragmentation	285104	hping3 [73]
	UDP Flood	5412287	udp-flood [74]
	SlowLoris	23426	slowloris [75]
	ICMP Flood	7200504	hping3 [73]
	RSTFIN Flood	4045285	hping3 [73]
	PSHACK Flood	4094755	hping3 [73]
	HTTP Flood	28790	golang-httpflood [76]
	UDP Fragmentation	286925	udp-flood [74]
	ICMP Fragmentation	452489	hping3 [73]
	TCP Flood	4497667	hping3 [73]
	SYN Flood	4059190	hping3 [73]
	SynonymousIP Flood	3598138	hping3 [73]
DoS	TCP Flood	2671445	hping3 [73]
	HTTP Flood	71864	golang-httpflood [76]
	SYN Flood	2028834	hping3 [73]
	UDP Flood	3318595	hping3 [73] & udp-flood [74]
Recon	Ping Sweep	2262	nmap [77]& fping [78]
	OS Scan	98259	nmap [77]
	Vulnerability Scan	37382	nmap [77] & vulscan [79]
	Port Scan	82284	nmap [77]
	Host Discovery	134378	nmap [77]
Web-Based	Sql Injection	5245	DVWA [80]
	Command Injection	5409	DVWA [80]
	Backdoor Malware	3218	DVWA [80] & Remot3d [81]
	Uploading Attack	1252	DVWA [80]
	XSS	3846	DVWA [80]
	Browser Hijacking	5859	Beef [82]
Brute Force	Dictionary Brute Force	13064	nmap [77] & hydra [83]
Spoofing	Arp Spoofing	307593	ettercap [84]
	DNS Spoofing	178911	ettercap [84]
Mirai	GREIP Flood	751682	Adapted Mirai Source Code [85]
	Greeth Flood	991866	Adapted Mirai Source Code [85]
	UDPPlain	890576	Adapted Mirai Source Code [85]

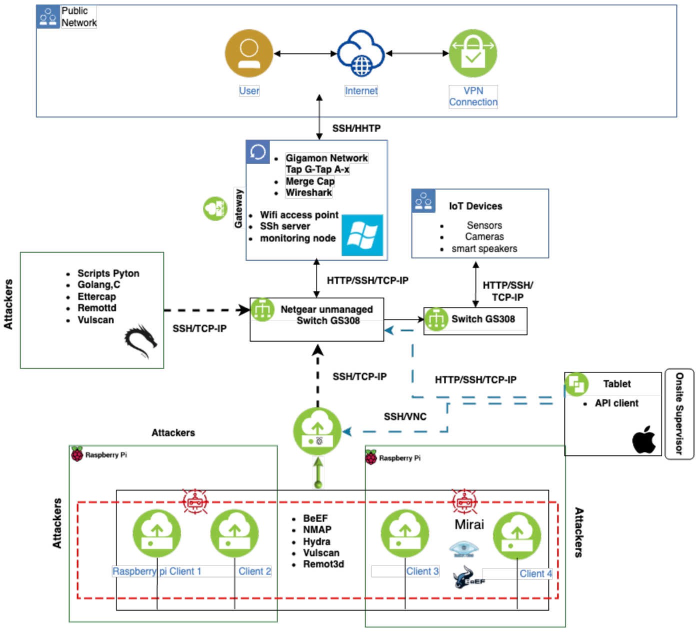

Figure 3. Basic attack framework for the dataset.

different scenarios. These files contain all packets sent and can be used to extract and engineer other features. Furthermore, csv files present a simpler way of loading and using the data. Those files are composed of features extracted from the original pcap files summarized by a fixed-size packet window. Figure 4 illustrates the process of converting pcap files into csv files.

Firstly, the network traffic data composed of captures of all attacks alongside benign traffic is used. As it represents about 548GB worth of traffic data, we split it into smaller chunks of 10MB to perform the conversion in parallel. This process is conducted using TCPDUMP [86]. After that, a parallel procedure is executed to extract several features using the DPKT package [87] and store them in separate csv files. These features are described in Table 4.

With the extracted features, we group the values captured in window sizes of 10 and 100 packets to mitigate data size discrepancy (e.g., DDoS and CommandInjection) and calculate their mean values using Pandas [88] and Numpy [89]. Finally, we combine all subfiles into a processed csv dataset using Pandas. Thereupon, the resulting csv datasets represent the combination of features of each data chunk.

Moreover, each attack conducted in this research presents different characteristics. For example, the network traffic generated by a DDoS attack tends to be larger than the network traffic generated by a Spoofing attack. Indeed, these differences can be also observed in other features of the dataset. Table 4 lists all features provided

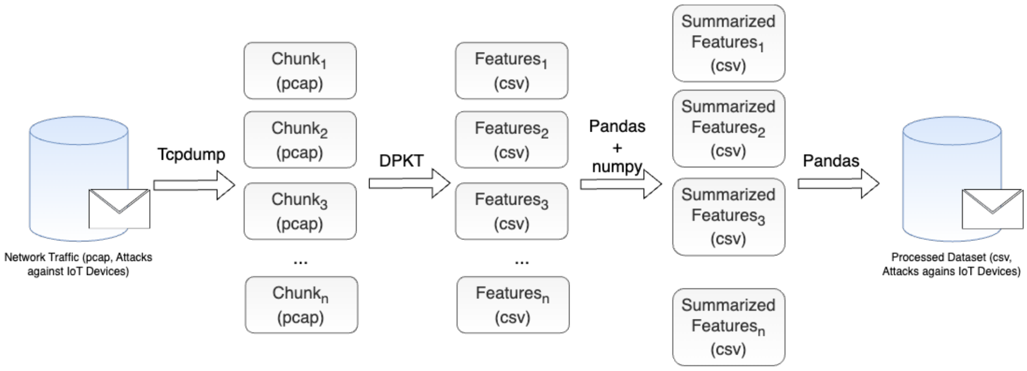
Network Traffic (pcap, Attacks against IoT Devices)
Tcpdump
Chunk$_1$ (pcap)
Chunk$_2$ (pcap)
Chunk$_3$ (pcap)
...
Chunk$_n$ (pcap)
DPKT
Features$_1$ (csv)
Features$_2$ (csv)
Features$_3$ (csv)
...
Features$_n$ (csv)
Pandas + numpy
Summarized Features$_1$ (csv)
Summarized Features$_2$ (csv)
Summarized Features$_3$ (csv)
Summarized Features$_n$ (csv)
Pandas
Processed Dataset (csv, Attacks agains IoT Devices)

Figure 4. Data processing: converting pcap files to csv.

in the dataset, which Table 5 presents the characteristics of these features. For each feature in the entire dataset, we present the mean, Standard Deviation (std), minimum (min), 25th percentile (25%), median (50%), 75th percentile (75%), and maximum (max) values.

5. Machine Learning (ML) Evaluation

In order to demonstrate how the CICIoT2023 dataset can be used to train Machine Learning (ML)-based attack detection and classification methods, Figure 5 illustrates the ML evaluation pipeline adopted in this research. Firstly, we combine all datasets produced following the procedure presented in Figure 4. In this sense, malicious and benign traffics are combined and shuffled into a single dataset (i.e., Blended Dataset) using PySpark [90]. Once the data is integrated, we evaluate ML performance from three different perspectives: (i) Multiclass classification - focussing on classifying 33 individual attacks-, (ii) Grouped classification - considering 7 attack groups (e.g., DDoS and DoS) -, and (iii) Binary classification (i.e., malicious and benign classification). In each case, the dataset is divided into the train (80%) and test (20%) sets, which are normalized using the Standard Scaler method [91] before the actual training process. Finally, the results obtained are summarized as integrated results.

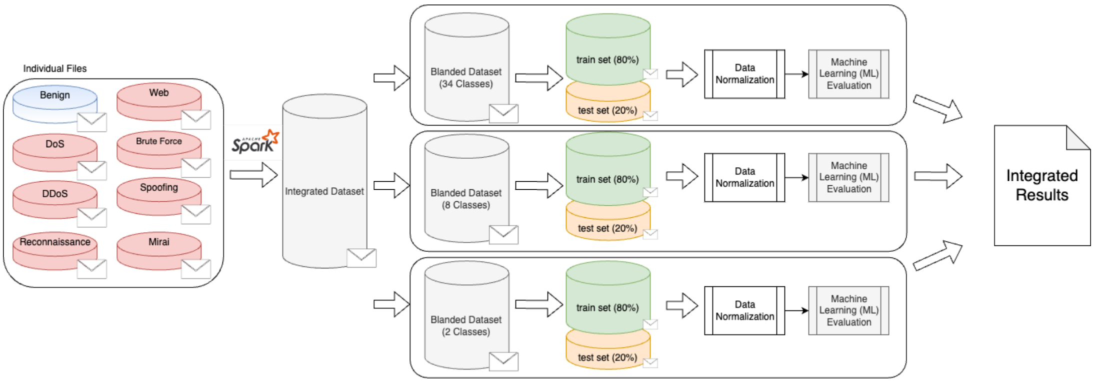

Figure 5. Machine Learning (ML) evaluation pipeline adopted in this research.

5.1. Metrics

The evaluation of different ML models and configurations is conducted based on evaluation metrics. Given that TP represents the True Positives, TN the True Negatives, FP the False Positive, and FN the False Negatives, the metrics used in this research are [92]:

- **Accuracy:** responsible for evaluating the classification models by depicting the proportion of correct predictions in a given dataset and is based on the following expression:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Recall:** the ratio of correctly identified labels to the total number of occurrences of that particular label:

$$Rec = \frac{TP}{TP + FN} \tag{2}$$

- **Precision:** the ratio of correctly identified labels to the total number of positive classifications:

$$Pre = \frac{TP}{TP + FP} \tag{3}$$

- **F1-Score:** geometric average of precision and recall:

$$F1 = 2 \times \frac{Pre \times Rec}{Pre + Rec} \tag{4}$$

Table 4. Features extracted from the network traffic.

#	Feature	Description
1	ts	Timestamp
2	flow duration	Duration of the packet's flow
3	Header Length	Header Length
4	Protocol Type	IP, UDP, TCP, IGMP, ICMP, Unknown (Integers)
5	Duration	Time-to-Live (ttl)
6	Rate	Rate of packet transmission in a flow
7	Srate	Rate of outbound packets transmission in a flow
8	Drate,	Rate of inbound packets transmission in a flow
9	fin flag number	Fin flag value
10	syn flag number	Syn flag value
11	rst flag number	Rst flag value
12	psh flag numbe	Psh flag value
13	ack flag number	Ack flag value
14	ece flag numbe	Ece flag value
15	cwr flag number	Cwr flag value
16	ack count	Number of packets with ack flag set in the same flow
17	syn count	Number of packets with syn flag set in the same flow
18	fin count	Number of packets with fin flag set in the same flow
19	urg coun	Number of packets with urg flag set in the same flow
20	rst count	Number of packets with rst flag set in the same flow
21	HTTP	Indicates if the application layer protocol is HTTP
22	HTTPS	Indicates if the application layer protocol is HTTPS
23	DNS	Indicates if the application layer protocol is DNS
24	Telnet	Indicates if the application layer protocol is Telnet
25	SMTP	Indicates if the application layer protocol is SMTP
26	SSH	Indicates if the application layer protocol is SSH
27	IRC	Indicates if the application layer protocol is IRC
28	TCP	Indicates if the transport layer protocol is TCP
29	UDP	Indicates if the transport layer protocol is UDP
30	DHCP	Indicates if the application layer protocol is DHCP
31	ARP	Indicates if the link layer protocol is ARP
32	ICMP	Indicates if the network layer protocol is ICMP
33	IPv	Indicates if the network layer protocol is IP
34	LLC	Indicates if the link layer protocol is LLC
35	Tot sum	Summation of packets lengths in flow
36	Min	Minimum packet length in the flow
37	Max	Maximumpacket length in the flow
38	AVG	Average packet length in the flow
39	Std	Standard deviation of packet length in the flow
40	Tot size	Packet's length
41	IAT	The time difference with the previous packet
42	Number	The number of packets in the flow
43	Magnitue	(Average of the lengths of incoming packets in the flow + Average of the lengths of outgoing packets in the flow) ** 0.5
44	Radius	(Variance of the lengths of incoming packets in the flow + Variance of the lengths of outgoing packets in the flow) ** 0.5
45	Covariance	Covariance of the lengths of incoming and outgoing packets
46	Variance	Variance of the lengths of incoming packets in the flow / The variance of the lengths of outgoing packets in the flow
47	Weight	Number of incoming packets * Number of outgoing packets

Table 5. Dataset Description

Feature	mean	std	min	25%	50%	75%	max
flow_duration	5.76544939	285.034171	0	0	0	0.10513809	394357.207
Header_Length	76705.9637	461331.747	0	54	54	280.555	9907147.75
Protocol Type	9.06568989	8.94553292	0	6	6	14.33	47
Duration	66.3507169	14.0191881	0	64	64	64	255
Rate	9064.05724	99562.4906	0	2.09185589	15.7542308	117.384754	8388608
Srate	9064.05724	99562.4906	0	2.09185589	15.7542308	117.384754	8388608
Drate	5.46E-06	0.00725077	0	0	0	0	29.7152249
fin_flag_number	0.08657207	0.28120696	0	0	0	0	1
syn_flag_number	0.20733528	0.40539779	0	0	0	0	1
rst_flag_number	0.09050473	0.28690351	0	0	0	0	1
psh_flag_number	0.08775006	0.28293106	0	0	0	0	1
ack_flag_number	0.12343168	0.32893207	0	0	0	0	1
ece_flag_number	1.48E-06	0.00121571	0	0	0	0	1
cwr_flag_number	7.28E-07	0.00085338	0	0	0	0	1
ack_count	0.09054283	0.28643144	0	0	0	0	7.7
syn_count	0.33035785	0.6635354	0	0	0	0.06	12.87
fin_count	0.09907672	0.32711642	0	0	0	0	248.32
urg_count	6.23982356	71.8524536	0	0	0	0	4401.7
rst_count	38.4681213	325.384658	0	0	0	0.01	9613
HTTP	0.04823423	0.21426079	0	0	0	0	1
HTTPS	0.05509922	0.22817383	0	0	0	0	1
DNS	0.00013068	0.01143079	0	0	0	0	1
Telnet	2.14E-08	0.00014635	0	0	0	0	1
SMTP	6.43E-08	0.00025349	0	0	0	0	1
SSH	4.09E-05	0.00639772	0	0	0	0	1
IRC	1.50E-07	0.00038722	0	0	0	0	1
TCP	0.57383427	0.49451846	0	0	1	1	1
UDP	0.21191758	0.40866676	0	0	0	0	1
DHCP	1.71E-06	0.00130903	0	0	0	0	1
ARP	6.62E-05	0.00813521	0	0	0	0	1
ICMP	0.16372157	0.37002273	0	0	0	0	1
IPv	0.99988731	0.01061485	0	1	1	1	1
LLC	0.99988731	0.01061485	0	1	1	1	1
Tot sum	1308.32257	2613.30273	42	525	567	567.54	127335.8
Min	91.6073456	139.695326	42	50	54	54	13583
Max	181.963418	524.030902	42	50	54	55.26	49014
AVG	124.668815	240.991485	42	50	54	54.0497296	13583
Std	33.3248065	160.335722	0	0	0	0.37190955	12385.2391
Tot size	124.691567	241.549341	42	50	54	54.06	13583
IAT	83182525.9	17047351.7	0	83071566	83124522.4	83343908	167639436
Number	9.49848933	0.81915318	1	9.5	9.5	9.5	15
Magnitue	13.12182	8.62857895	9.16515139	10	10.3923048	10.3967148	164.821115
Radius	47.0949848	226.769647	0	0	0	0.50592128	17551.2708
Covariance	30724.3565	323710.68	0	0	0	1.34421569	154902159
Variance	0.0964376	0.233001	0	0	0	0.08	1
Weight	141.51237	21.0683073	1	141.55	141.55	141.55	244.6

Table 6. Results obtained in the classification process conducted using different Machine Learning models (illustrated in Figure 6).

	Metric	Logistic Regression	Perceptron	Adaboost	Random Forest (RF)	Deep Neural Network (DNN)
34 classes	Accuracy	0.80231507	0.8195961	0.607888	0.99164365	0.986118011
	Recall	0.59520185	0.507506	0.607675	0.831586401	0.731868794
	Precision	0.486752461	0.454634	0.479621	0.704492066	0.665295126
	F1-score	0.49388408	0.4472933	0.473498	0.714021981	0.672346883
8 classes	Accuracy	0.831674188	0.8663152	0.351357	0.994368173	0.991147043
	Recall	0.696055597	0.6591315	0.487789	0.91001105	0.906642708
	Precision	0.512409686	0.5239188	0.464924	0.705407564	0.679434746
	F1-score	0.539424048	0.5551339	0.368663	0.71928904	0.69726491
2 classes	Accuracy	0.989023188	0.9817525	0.995899	0.99680798	0.994422814
	Recall	0.890400624	0.7970288	0.947303	0.965163906	0.933277496
	Precision	0.863157959	0.825432	0.965631	0.965395244	0.947579486
	F1-score	0.876258983	0.8105374	0.956273	0.965279544	0.940305998

5.2. *Evaluation*

In the evaluation process, we adopted five ML methods that have been successfully adopted in different applications including cybersecurity: Logistic Regression [93], Perceptron [94], Adaboost [95] [96] [97], Random Forest [98], and Deep Neural Network [99]. Figure 6 illustrates the performance of all methods when framing the classification problem as binary (i.e., malicious and benign), multiclass with 8 classes (i.e., benign and attack categories), and multiclass with 34 classes (i.e., benign and all individual attacks). These results are also depicted in Table 6.

For the binary classification, the results show that all methods present high performance. While accuracy is a metric that all methods reach over 98%, F1-score highlights the difference among these approaches. For example, Perceptron achieves 81%, showing that it suffers since the minority class (i.e., benign) is misclassified more often. In the classification of attack groups (i.e., 8 classes), the overall performance is degraded since the classification task becomes more challenging. The Logistic Regression, Perceptron, and Adaboost methods show a significant decrease in accuracy. This impact is even more perceptible for F1-score. However, both Random Forest and Deep Neural Network are able to maintain high accuracy and F-1 score. These methods also present a decrease in performance but are capable of achieving F1 scores of 70%.

Finally, the most challenging classification task is represented by a multiclass classification of individual attacks (i.e., 34 classes). In this scenario, both Random Forest and Deep Neural Network could maintain high accuracy with very similar results. The same applies to F1-score since a slight reduction was perceived (around 1%) compared to the 8-class challenge. Furthermore, this case study shows that the Logistic Regression, Perceptron, and Adaboost methods are not able to categorize attacks as efficiently, given that the average accuracy is below 80% and F1-score is less than 50% in all cases.

These results show how ML methods can be used to classify attacks against IoT operations. In fact, this is a starting point that can be considered in any ML-based cybersecurity solutions for IoT operations. This effort not only highlights that the use of other ML methods is possible (e.g., optimized methods), but also enables the adoption of similar strategies to solve IoT-specific problems. Finally, although we are focussing on 33 different attacks, future directions could also be tailored to address issues related to individual attacks or categories.

5.3. *Discussion*

To illustrate how these models are performing for each class, Tables 8 and 7 show the confusion matrix for Random Forest and Deep Neural Networks in the case of multiclass classification (8 classes).

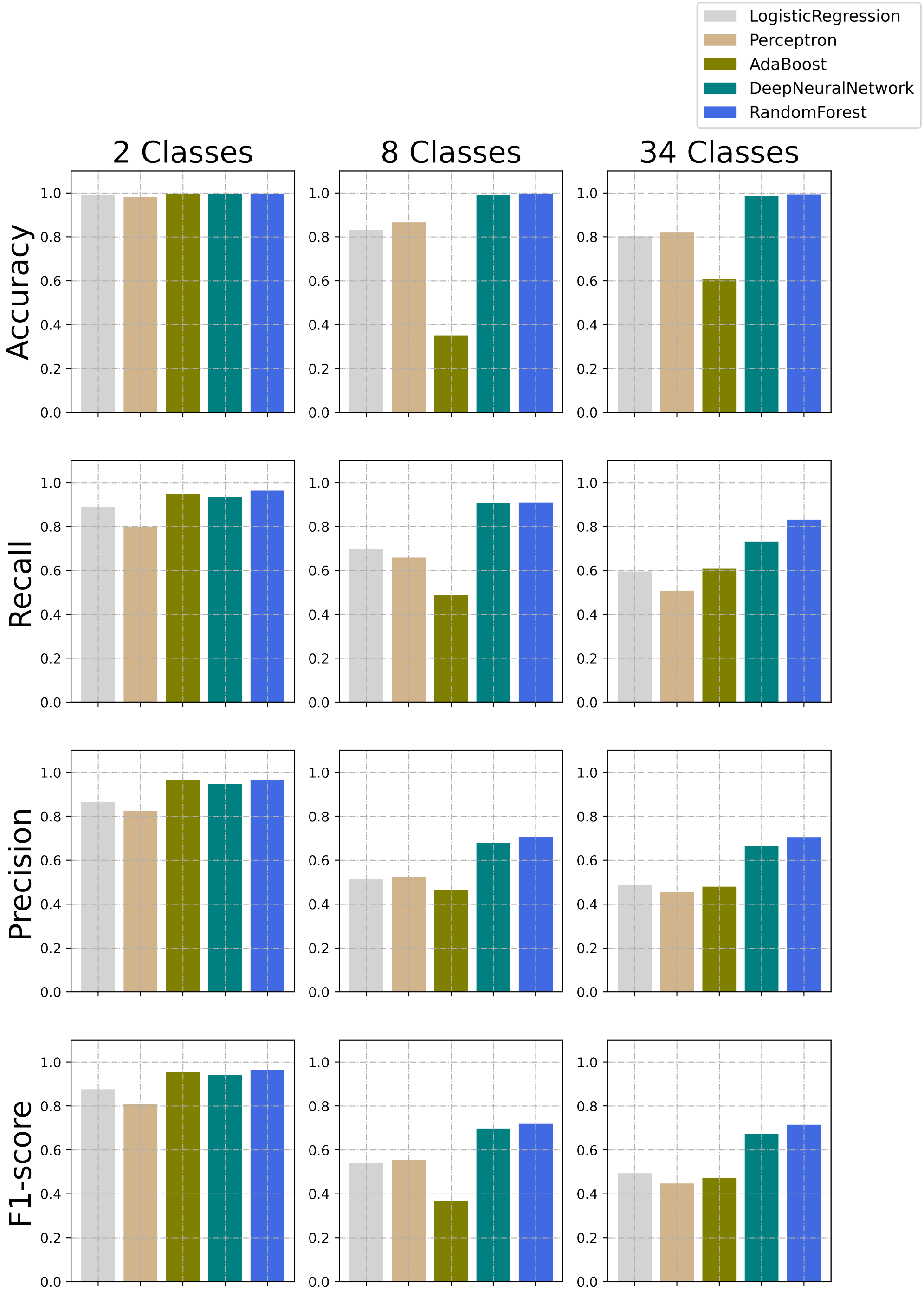

Figure 6. Results obtained in the classification process conducted using different Machine Learning models.

Table 7. Confusion matrix for Deep Neural Network in the case of multiclass classification (8 classes).

	Benign	BruteForce	DDoS	DoS	Mirai	Recon	Spoofing	Web
Benign	**230229**	1	7	2	0	9270	3812	1
BruteForce	1054	**438**	3	0	0	1216	271	1
DDoS	23	0	**7523853**	1012	545	653	65	0
DoS	15	0	4933	**1787065**	60	61	33	0
Mirai	10	0	258	41	**583283**	64	21	0
Recon	18517	2	968	30	1	**55656**	3455	1
Spoofing	30485	0	17	0	15	10021	**67257**	3
Web	1976	0	1	0	0	2028	1221	**207**

Table 8. Confusion matrix for Random Forest in the case of multiclass classification (8 classes).

	Benign	BruteForce	DDoS	DoS	Mirai	Recon	Spoofing	Web
Benign	**234929**	4	24	2	4	3192	5159	8
BruteForce	1342	**169**	1	0	0	844	626	1
DDoS	15	0	**7525049**	557	18	339	173	0
DoS	7	0	1088	**1790979**	34	12	47	0
Mirai	5	0	603	18	**582921**	100	30	0
Recon	11565	6	1418	11	16	**60006**	5591	17
Spoofing	14618	1	18	6	11	4743	**88371**	30
Web	1140	1	3	1	1	1265	2792	**230**

In both cases, it is possible to observe that some classes are very well classified, mainly those with a large number of occurrences in the dataset. For example, the misclassification rates for DDoS, DoS, and Mirai are very small, followed by Recon and Spoofing.

However, these models face challenges in classifying other attacks. For example, web-based attacks are usually classified as benign, Recon, or spoofing. The same occurs in the Brute Force classification. Finally, although the similarities in the data patterns lead the models to make these mistakes, the classification is successful in most cases, leading to the results depicted in Figure 6.

6. Conclusion

Nowadays, IoT is becoming increasingly important for society. In this context, the development of security solutions is pivotal to enabling efficient, secure, and dependable IoT operations. This research introduced a novel and extensive IoT attack dataset to foster the development of security analytics applications in real IoT operations. In this process, 33 attacks are executed in an IoT topology composed of 105 devices. These attacks are classified into seven categories (i.e., DDoS, DoS, Recon, Web-based, Brute Force, Spoofing, and Mirai) and all attacks are executed by malicious IoT devices targeting other IoT devices. Furthermore, this dataset includes multiple attacks not available in other IoT datasets and enables IoT professionals to develop new security analytics solutions using data in different formats. The dataset will be available through the CIC Dataset website (https://www.unb.ca/cic/datasets/index.html).

This work enables the development of several future works, e.g., the optimization of ML models, the analysis of features and how they influence different ML models, the interpretation of classifications, and the analysis of transferability based on the comparison to other datasets.

Acknowledgments: The authors graciously acknowledge the support from the Canadian Institute for Cybersecurity (CIC), the funding support from the NSERC Discovery Grant (no. RGPIN 231074), and Tier 1 Canada Research Chair to Dr. Ghorbani.

References

1. Da Xu, L.; He, W.; Li, S. Internet of things in industries: A survey. *IEEE Transactions on industrial informatics* **2014**, *10*, 2233–2243.
2. Nauman, A.; Qadri, Y.A.; Amjad, M.; Zikria, Y.B.; Afzal, M.K.; Kim, S.W. Multimedia Internet of Things: A comprehensive survey. *IEEE Access* **2020**, *8*, 8202–8250.

3. Habibzadeh, H.; Dinesh, K.; Shishvan, O.R.; Boggio-Dandry, A.; Sharma, G.; Soyata, T. A survey of healthcare Internet of Things (HIoT): A clinical perspective. *IEEE Internet of Things Journal* **2019**, *7*, 53–71.
4. Lee, S.K.; Bae, M.; Kim, H. Future of IoT networks: A survey. *Applied Sciences* **2017**, *7*, 1072.
5. Marjani, M.; Nasaruddin, F.; Gani, A.; Karim, A.; Hashem, I.A.T.; Siddiqa, A.; Yaqoob, I. Big IoT data analytics: architecture, opportunities, and open research challenges. *ieee access* **2017**, *5*, 5247–5261.
6. Hajjaji, Y.; Boulila, W.; Farah, I.R.; Romdhani, I.; Hussain, A. Big data and IoT-based applications in smart environments: A systematic review. *Computer Science Review* **2021**, *39*, 100318.
7. Dadkhah, S.; Mahdikhani, H.; Danso, P.K.; Zohourian, A.; Truong, K.A.; Ghorbani, A.A. Towards the development of a realistic multidimensional IoT profiling dataset. In Proceedings of the 2022 19th Annual International Conference on Privacy, Security & Trust (PST). IEEE, 2022, pp. 1–11.
8. Madakam, S.; Lake, V.; Lake, V.; Lake, V.; et al. Internet of Things (IoT): A literature review. *Journal of Computer and Communications* **2015**, *3*, 164.
9. Čolaković, A.; Hadžialić, M. Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues. *Computer networks* **2018**, *144*, 17–39.
10. Safi, M.; Kaur, B.; Dadkhah, S.; Shoeleh, F.; Lashkari, A.H.; Molyneaux, H.; Ghorbani, A.A. Behavioural Monitoring and Security Profiling in the Internet of Things (IoT). In Proceedings of the 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). IEEE, 2021, pp. 1203–1210.
11. Selvaraj, S.; Sundaravaradhan, S. Challenges and opportunities in IoT healthcare systems: a systematic review. *SN Applied Sciences* **2020**, *2*, 1–8.
12. Akkaş, M.A.; Sokullu, R.; Cetin, H.E. Healthcare and patient monitoring using IoT. *Internet of Things* **2020**, *11*, 100173.
13. Mohammed, J.; Lung, C.H.; Ocneanu, A.; Thakral, A.; Jones, C.; Adler, A. Internet of Things: Remote patient monitoring using web services and cloud computing. In Proceedings of the 2014 IEEE international conference on internet of things (iThings), and IEEE green computing and communications (GreenCom) and IEEE cyber, physical and social computing (CPSCom). IEEE, 2014, pp. 256–263.
14. Zantalis, F.; Koulouras, G.; Karabetsos, S.; Kandris, D. A review of machine learning and IoT in smart transportation. *Future Internet* **2019**, *11*, 94.
15. Uma, S.; Eswari, R. Accident prevention and safety assistance using IOT and machine learning. *Journal of Reliable Intelligent Environments* **2022**, *8*, 79–103.
16. Celesti, A.; Galletta, A.; Carnevale, L.; Fazio, M.; Ĺay-Ekuakille, A.; Villari, M. An IoT cloud system for traffic monitoring and vehicular accidents prevention based on mobile sensor data processing. *IEEE Sensors Journal* **2017**, *18*, 4795–4802.
17. Cheng, J.; Chen, W.; Tao, F.; Lin, C.L. Industrial IoT in 5G environment towards smart manufacturing. *Journal of Industrial Information Integration* **2018**, *10*, 10–19.
18. Al-Emran, M.; Malik, S.I.; Al-Kabi, M.N. A survey of Internet of Things (IoT) in education: Opportunities and challenges. *Toward social internet of things (SIoT): enabling technologies, architectures and applications* **2020**, pp. 197–209.
19. Pate, J.; Adegbija, T. AMELIA: An application of the Internet of Things for aviation safety. In Proceedings of the 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE, 2018, pp. 1–6.
20. Salam, A. Internet of things for sustainable forestry. In *Internet of Things for Sustainable Community Development*; Springer, 2020; pp. 147–181.
21. Cisco, U. Cisco annual internet report (2018–2023) white paper. *Cisco: San Jose, CA, USA* **2020**.
22. Vermesan, O.; Friess, P.; Guillemin, P.; Giaffreda, R.; Grindvoll, H.; Eisenhauer, M.; Serrano, M.; Moessner, K.; Spirito, M.; Blystad, L.C.; et al. Internet of things beyond the hype: Research, innovation and deployment. In *Building the Hyperconnected Society-Internet of Things Research and Innovation Value Chains, Ecosystems and Markets*; River Publishers, 2022; pp. 15–118.
23. Shafique, K.; Khawaja, B.A.; Sabir, F.; Qazi, S.; Mustaqim, M. Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. *Ieee Access* **2020**, *8*, 23022–23040.
24. Neto, E.C.P.; Dadkhah, S.; Ghorbani, A.A. Collaborative DDoS Detection in Distributed Multi-Tenant IoT using Federated Learning. In Proceedings of the 2022 19th Annual International Conference on Privacy, Security & Trust (PST). IEEE, 2022, pp. 1–10.
25. Kaur, B.; Dadkhah, S.; Xiong, P.; Iqbal, S.; Ray, S.; Ghorbani, A.A. Verification based scheme to restrict iot attacks. In Proceedings of the 2021 IEEE/ACM 8th International Conference on Big Data Computing, Applications and Technologies (BDCAT'21), 2021, pp. 63–68.
26. Sharma, S.; Kaushik, B. A survey on internet of vehicles: Applications, security issues & solutions. *Vehicular Communications* **2019**, *20*, 100182.
27. Safi, M.; Dadkhah, S.; Shoeleh, F.; Mahdikhani, H.; Molyneaux, H.; Ghorbani, A.A. A Survey on IoT Profiling, Fingerprinting, and Identification. *ACM Transactions on Internet of Things* **2022**, *3*, 1–39.
28. Abrishami, M.; Dadkhah, S.; Neto, E.C.P.; Xiong, P.; Iqbal, S.; Ray, S.; Ghorbani, A.A. Label Noise Detection in IoT Security based on Decision Tree and Active Learning. In Proceedings of the 2022 IEEE 19th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET). IEEE, 2022, pp. 046–053.

29. Erfani, M.; Shoeleh, F.; Dadkhah, S.; Kaur, B.; Xiong, P.; Iqbal, S.; Ray, S.; Ghorbani, A.A. A feature exploration approach for IoT attack type classification. In Proceedings of the 2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). IEEE, 2021, pp. 582–588.
30. Meidan, Y.; Bohadana, M.; Mathov, Y.; Mirsky, Y.; Shabtai, A.; Breitenbacher, D.; Elovici, Y. N-baiot—network-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Computing* **2018**, *17*, 12–22.
31. Bezerra, V.H.; da Costa, V.G.T.; Martins, R.A.; Junior, S.B.; Miani, R.S.; Zarpelao, B.B. Providing IoT host-based datasets for intrusion detection research. In Proceedings of the Anais do XVIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais, 2018, pp. 15–28.
32. Anthi, E.; Williams, L.; Słowińska, M.; Theodorakopoulos, G.; Burnap, P. A supervised intrusion detection system for smart home IoT devices. *IEEE Internet of Things Journal* **2019**, *6*, 9042–9053.
33. Koroniotis, N.; Moustafa, N.; Sitnikova, E.; Turnbull, B. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. *Future Generation Computer Systems* **2019**, *100*, 779–796.
34. Mirsky, Y.; Doitshman, T.; Elovici, Y.; Shabtai, A. Kitsune: an ensemble of autoencoders for online network intrusion detection. *arXiv preprint arXiv:1802.09089* **2018**.
35. Kang, H.; Ahn, D.H.; Lee, G.M.; Yoo, J.; Park, K.H.; Kim, H.K. IoT network intrusion dataset. *IEEE Dataport* **2019**.
36. Guerra-Manzanares, A.; Medina-Galindo, J.; Bahsi, H.; Nõmm, S. MedBIoT: Generation of an IoT Botnet Dataset in a Medium-sized IoT Network. In Proceedings of the ICISSP, 2020, pp. 207–218.
37. Parmisano, A.; Garcia, S.; Erquiaga, M. A Labeled Dataset with Malicious and Benign IoT Network Traffic. *Stratosphere Laboratory: Praha, Czech Republic* **2020**.
38. Ullah, I.; Mahmoud, Q.H. A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks. In Proceedings of the Canadian Conference on Artificial Intelligence, 2020, pp. 508–520.
39. Vaccari, I.; Chiola, G.; Aiello, M.; Mongelli, M.; Cambiaso, E. MQTTset, a New Dataset for Machine Learning Techniques on MQTT. *Sensors* **2020**, *20*, 6578.
40. Hindy, H.; Bayne, E.; Bures, M.; Atkinson, R.; Tachtatzis, C.; Bellekens, X. Machine Learning Based IoT Intrusion Detection System: An MQTT Case Study. *arXiv preprint arXiv:2006.15340* **2020**.
41. Alsaedi, A.; Moustafa, N.; Tari, Z.; Mahmood, A.; Anwar, A. TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for data-driven Intrusion Detection Systems. *IEEE Access* **2020**, *8*, 165130–165150.
42. Ferrag, M.A.; Friha, O.; Hamouda, D.; Maglaras, L.; Janicke, H. Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. *IEEE Access* **2022**, *10*, 40281–40306.
43. Lamping, U.; Warnicke, E. Wireshark user's guide. *Interface* **2004**, *4*, 1.
44. Baxter, J.H. *Wireshark essentials*; Packt Publishing Ltd, 2014.
45. Kumari, P.; Jain, A.K. A Comprehensive Study of DDoS Attacks over IoT Network and Their Countermeasures. *Computers & Security* **2023**, p. 103096.
46. Duravkin, I.; Loktionova, A.; Carlsson, A. Method of slow-attack detection. In Proceedings of the 2014 First International Scientific-Practical Conference Problems of Infocommunications Science and Technology. IEEE, 2014, pp. 171–172.
47. Harshita, H. Detection and prevention of ICMP flood DDOS attack. *International Journal of New Technology and Research* **2017**, *3*, 263333.
48. Sreeram, I.; Vuppala, V.P.K. HTTP flood attack detection in application layer using machine learning metrics and bio inspired bat algorithm. *Applied computing and informatics* **2019**, *15*, 59–66.
49. Acharya, A.A.; Arpitha, K.; Kumar, B. An intrusion detection system against UDP flood attack and ping of death attack (DDOS) in MANET. *International Journal of Engineering and Technology (IJET)* **2016**, *8*.
50. Cebeloglu, F.S.; Karakose, M. A cyber security analysis used for unmanned aerial vehicles in the smart city. In Proceedings of the 2019 1st International Informatics and Software Engineering Conference (UBMYK). IEEE, 2019, pp. 1–6.
51. Chen, E.Y. Detecting TCP-based DDoS attacks by linear regression analysis. In Proceedings of the Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005. IEEE, 2005, pp. 381–386.
52. Kaufman, C.; Perlman, R.; Sommerfeld, B. DoS protection for UDP-based protocols. In Proceedings of the Proceedings of the 10th ACM conference on Computer and communications security, 2003, pp. 2–7.
53. Gilad, Y.; Herzberg, A. Fragmentation considered vulnerable. *ACM Transactions on Information and System Security (TISSEC)* **2013**, *15*, 1–31.
54. Bogdanoski, M.; Suminoski, T.; Risteski, A. Analysis of the SYN flood DoS attack. *International Journal of Computer Network and Information Security (IJCNIS)* **2013**, *5*, 1–11.
55. Raptis, G.E.; Katsini, C.; Alexakos, C. Towards Automated Matching of Cyber Threat Intelligence Reports based on Cluster Analysis in an Internet-of-Vehicles Environment. In Proceedings of the 2021 IEEE International Conference on Cyber Security and Resilience (CSR). IEEE, 2021, pp. 366–371.
56. Al-Jarrah, O.; Arafat, A. Network intrusion detection system using neural network classification of attack behavior. *Journal of Advances in Information Technology Vol* **2015**, *6*.
57. Orebaugh, A.; Pinkard, B. *Nmap in the enterprise: your guide to network scanning*; Elsevier, 2011.

58. deRito, C.; Bhatia, S. Comparative Analysis of Open-Source Vulnerability Scanners for IoT Devices. In *Intelligent Data Communication Technologies and Internet of Things*; Springer, 2022; pp. 785–800.
59. Bhuyan, M.H.; Bhattacharyya, D.K.; Kalita, J.K. Surveying port scans and their detection methodologies. *The Computer Journal* **2011**, *54*, 1565–1581.
60. Wolfgang, M. Host Discovery with nmap. *Exploring nmap's default behavior* **2002**, *1*, 16.
61. Jensen, M.; Gruschka, N.; Herkenhöner, R. A survey of attacks on web services. *Computer Science-Research and Development* **2009**, *24*, 185–197.
62. Halfond, W.G.; Viegas, J.; Orso, A.; et al. A classification of SQL-injection attacks and countermeasures. In Proceedings of the Proceedings of the IEEE international symposium on secure software engineering. IEEE, 2006, Vol. 1, pp. 13–15.
63. Su, Z.; Wassermann, G. The essence of command injection attacks in web applications. *Acm Sigplan Notices* **2006**, *41*, 372–382.
64. Loi, H.; Olmsted, A. Low-cost detection of backdoor malware. In Proceedings of the 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, 2017, pp. 197–198.
65. Wassermann, G.; Su, Z. Static detection of cross-site scripting vulnerabilities. In Proceedings of the 2008 ACM/IEEE 30th International Conference on Software Engineering. IEEE, 2008, pp. 171–180.
66. Kumar, M.M.S.; Indrani, B. A Study on Web Hijacking Techniques and Browser Attacks. *International Journal of Applied Engineering Research* **2018**, *13*, 2614–2618.
67. van der Merwe, J.R.; Zubizarreta, X.; Lukčin, I.; Rügamer, A.; Felber, W. Classification of spoofing attack types. In Proceedings of the 2018 European Navigation Conference (ENC). IEEE, 2018, pp. 91–99.
68. Whalen, S. An introduction to arp spoofing. *Node99 [Online Document]* **2001**.
69. Maksutov, A.A.; Cherepanov, I.A.; Alekseev, M.S. Detection and prevention of DNS spoofing attacks. In Proceedings of the 2017 Siberian Symposium on Data Science and Engineering (SSDSE). IEEE, 2017, pp. 84–87.
70. Stiawan, D.; Idris, M.; Malik, R.F.; Nurmaini, S.; Alsharif, N.; Budiarto, R.; et al. Investigating brute force attack patterns in IoT network. *Journal of Electrical and Computer Engineering* **2019**, *2019*.
71. Vykopal, J.; Plesnik, T.; Minarik, P. Network-based dictionary attack detection. In Proceedings of the 2009 international conference on future networks. IEEE, 2009, pp. 23–27.
72. Abbas, S.G.; Hashmat, F.; Shah, G.A.; Zafar, K. Generic signature development for IoT Botnet families. *Forensic Science International: Digital Investigation* **2021**, *38*, 301224.
73. Tools, K. hping3 package description, 2019.
74. EPC-MSU. UDP Flood, 2023. https://github.com/EPC-MSU/udp-flood.
75. Yaltirakli, G. Slowloris. *github.com* **2015**.
76. Golang-HTTPFlood, 2020. https://github.com/Leeon123/golang-httpflood.
77. Lyon, G. Nmap security scanner. *línea] URL: http://nmap. org/[Consulta: 8 de junio de 2012]* **2014**.
78. Tools, K. Fping, 2023.
79. SCIP. Vulscan, 2023. https://github.com/scipag/vulscan.
80. DVWA. DAMN VULNERABLE WEB APPLICATION, 2017. https://github.com/digininja/DVWA.
81. KeepWannabe. Remot3d - An Easy Way To Exploiting, 2020. https://github.com/KeepWannabe/Remot3d.
82. BEEF. The Browser Exploitation Framework, 2023. https://beefproject.com.
83. Maciejak, D. Hydra, 2023. https://github.com/vanhauser-thc/thc-hydra.
84. Ornaghi, A.; Valleri, M. Ettercap, 2005.
85. Gamblin, J. Mirai BotNet, 2017. https://github.com/jgamblin/Mirai-Source-Code.
86. TCPDUMP. Tcpdump(1) man page. *https://www.tcpdump.org/manpages/tcpdump.1.html* **2022**.
87. DPKT. Dpkt documentation. *https://dpkt.readthedocs.io/en/latest/* **2022**.
88. PANDAS. pandas-dev/pandas: Pandas **2020**. https://doi.org/10.5281/zenodo.3509134.
89. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2.
90. Drabas, T.; Lee, D. *Learning PySpark*; Packt Publishing Ltd, 2017.
91. Scikit-learn. StandardScaler, 2023. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.
92. Danso, P.K.; Neto, E.C.P.; Dadkhah, S.; Zohourian, A.; Molyneaux, H.; Ghorbani, A.A. Ensemble-based Intrusion Detection for Internet of Things Devices. In Proceedings of the 2022 IEEE 19th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET). IEEE, 2022, pp. 034–039.
93. Bapat, R.; Mandya, A.; Liu, X.; Abraham, B.; Brown, D.E.; Kang, H.; Veeraraghavan, M. Identifying malicious botnet traffic using logistic regression. In Proceedings of the 2018 systems and information engineering design symposium (SIEDS). IEEE, 2018, pp. 266–271.
94. Shynk, J.J. Performance surfaces of a single-layer perceptron. *IEEE Transactions on Neural Networks* **1990**, *1*, 268–274.
95. AlShahrani, B.M.M.; et al. Classification of cyber-attack using Adaboost regression classifier and securing the network. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* **2021**, *12*, 1215–1223.
96. Rehman Javed, A.; Jalil, Z.; Atif Moqurrab, S.; Abbas, S.; Liu, X. Ensemble adaboost classifier for accurate and fast detection of botnet attacks in connected vehicles. *Transactions on Emerging Telecommunications Technologies* **2022**, *33*, e4088.

97. Khan, F.; Ahamed, J.; Kadry, S.; Ramasamy, L.K. Detecting malicious URLs using binary classification through ada boost algorithm. *International Journal of Electrical & Computer Engineering (2088-8708)* **2020**, *10*.
98. Choubisa, M.; Doshi, R.; Khatri, N.; Hiran, K.K. A simple and robust approach of random forest for intrusion detection system in cyber security. In Proceedings of the 2022 International Conference on IoT and Blockchain Technology (ICIBT). IEEE, 2022, pp. 1–5.
99. Xin, Y.; Kong, L.; Liu, Z.; Chen, Y.; Li, Y.; Zhu, H.; Gao, M.; Hou, H.; Wang, C. Machine learning and deep learning methods for cybersecurity. *Ieee access* **2018**, *6*, 35365–35381.