Data-Driven Attack Detection Trends in IoT

Subjects: Computer Science, Information Systems

Contributor: Safwana Haque , Fadi El-Moussa , Nikos Komninos , Rajarajan Muttukrishnan

The Internet of Things is perhaps a concept that the world cannot be imagined without today, having become intertwined in everyday lives in the domestic, corporate and industrial spheres. However, irrespective of the convenience, ease and connectivity provided by the Internet of Things, the security issues and attacks faced by this technological framework are equally alarming and undeniable. In order to address these various security issues, researchers race against evolving technology, trends and attacker expertise.

IoT datasets machine learning cyberattack

1. Introduction

Technology is a rapidly evolving paradigm that is especially difficult to keep up with in the field of computing. This can be mainly accredited to the advancements made in semiconductor chips, which are continuously improved and exploited for research purposes. Some of the most recent buzz terms that can be commonly heard and are of relevance are machine learning (ML), federated learning (FL), blockchain and Internet of Things (IoT). These technologies can be further combined with one another to improve their individual outputs or efficiency and to generate an alternate byproduct or result. For example, FL can be used to ensure or enhance data privacy in the IoT and ML can be used to make automated predictions in IoT devices. On the other hand, blockchain can be used to improve trust and transparency in data transactions in IoT networks.

IoT, is a term coined by Kevin Ashton in 1999 [1] but only gained traction in 2013. Since 2017, IoT has grown tremendously and will continue to do so at an even greater rate according to market and industry surveys [2][3][4][5][6]. IoT has penetrated every sector of life, encompassing transportation, health, communication, agriculture, homes, etc., with even traditional devices having become ‘smart’, e.g., smart locks, smart cars, smart fridges, smart lights, smart speakers and smart watches. According to [7], as of 2020, there was an equal number of IoT and non-IoT devices in the world, and the amount of the former is estimated to triple by 2025. While making life easier, this explosive growth has introduced many related concerns, such as the need for more speed, storage, capabilities, efficiency, etc., which researchers are continually trying to address and improve.

One of the biggest growing concerns, however, is the security and privacy of users, data, devices and the IoT network, which are often overlooked by both manufacturers and consumers. Implementing failsafe systems can be a painstaking process, yet the failure to do so can lead to serious repercussions for both individual users and companies. Cybercrimes are very common and already impact existing home IoT networks. A recent incident reported by the British Broadcasting Corporation (BBC), for instance, revealed how a family became suspects to a

cybercrime that involved child abuse, to the detriment of their domestic life, income and mental health, the crime most likely having occurred via the hacking of their Wireless Fidelity (Wi-Fi) router, whose default password settings had not been changed [8]. Most cyberattacks commonly result from exploiting security vulnerabilities, such as weak/default password usage, poor update management, insecure interfaces, lack of user and data privacy, poor user awareness, lack of vendor standardization and many more.

Numerous steps must be continually taken to ensure that cybersecurity is maintained. These include the raising of user awareness/cyber education, security policy implementations, security software and tools (such as antivirus, firewalls, etc.) and, more recently, automated measures using machine and deep learning (DL) techniques. Exhaustive research has been carried out for conventional network and data security, but such work is severely lacking in emerging fields such as IoT. For example, numerous datasets have been generated and created by various studies and researchers on general-purpose networks, the earliest of which—known as the DARPA (Defense Advanced Research Projects Agency) dataset—dates back to 1998 [9]. Other datasets, found in [10][11][12], have been used to design intrusion detection and prevention systems (IDSs and IPSs, respectively). With respect to those widely used to train ML algorithms for IoT networks, older datasets, such as Knowledge Discovery in Databases (KDD) and Network Security Laboratory Knowledge Discovery in Databases (NSL-KDD), are believed to have shortcomings, e.g., there are a large number of duplicate records that could skew the machine training and learning process in the KDD dataset [13], and NSL-KDD, though an improvement over KDD, does not include more recent attack classes and IoT network properties. UNSW-NB15 [14] (by the University of New South Wales) and CIC-IDS2017 and CIC-IDS2018 [15] (by the Canadian Institute for Cybersecurity) are the more recent datasets used for IoT ML training, but as these datasets are not primarily concerned with IoT networks attack detection becomes limited.

2. What Are the Datasets Created Specifically for the Study of IoT Networks and Their Security?

The survey addresses this question by finding datasets that have been created using IoT devices in either a simulated environment or a physical network. In most cases, the IoT networks created are exposed to attacks and the network behavior is studied and analyzed under various attack conditions. Benign and attack data are collected and used to train ML and DL algorithms to create intrusion detection systems (IDSs). Ten datasets were found that are being studied and experimented on as part of this survey. Brief descriptions of these datasets are given below, while details of their attack capabilities can be found in **Table 1**.

- **Bot-IoT** [16] is a simulated dataset created to study and analyze network forensics using ML and DL techniques. It is based on five IoT scenarios consisting of a weather station, a smart fridge, motion-activated lights, a remotely activated garage door and a smart thermostat. These simulated environments were exposed to three categories of attacks: information gathering (port scans, operating system (OS) fingerprinting); denial of service (Transmission Control Protocol (TCP), User Datagram Protocol (UDP), Hypertext Transfer Protocol (HTTP) for both denial of service (DoS) and distributed denial of service (DDoS)), and information theft (keylogging and data theft), which are commonly exploited by botnets (bots). This dataset consists of more than 72 million

packet capture (PCAP) records. The distribution of attack records is not uniform, however, with the information theft attacks having the least number of records.

- **IoT Network Intrusion Dataset** [17] **(IoTNID)** was created using two real devices: a camera and a speaker. The dataset consists of reconnaissance, man-in-the-middle (MiTM), DoS and Mirai attacks. All the attack packets except those of Mirai were captured using the Nmap tool, while the Mirai attack packets were generated using a laptop.

- **IoT-23** [18] is a dataset created using three physical IoT devices: a Philips HUE smart Light Emitting Diode (LED) light, an Amazon Echo device and a Somfy smart door lock. These devices were set up to model 20 different malware scenarios and 3 benign scenarios (one for each device). Each malware scenario was exposed to a botnet (bot) attack, such as Mirai, Gafgyt, Torii, etc. This dataset was manually analyzed to provide benign and attack traffic features.

- **MedBIoT** [19] is a dataset that tries to emulate a medium-sized network consisting of 80 simulated devices and 3 real devices. The devices used were a switch, a light bulb, a lock and a fan. The setup was exposed to three types of botnets: Mirai, BASHLITE and Torii. This dataset aims to provide data for intrusion detection of botnets.

- **MQTT-IoT** [20] is a dataset based on a publish/subscribe message protocol called Message Queue Telemetry Transport (MQTT) used in the application/middleware layer. It is based on a simulated setup comprising 12 IoT sensors in four different attack scenarios (**Table 1**) and one benign scenario. This dataset was intended to be used for intrusion detection using ML techniques.

- **MQTTset** [21] is another dataset based on the MQTT communication protocol, in this case aimed at aiding the application of ML techniques in MQTT networks. The setup was simulated using eight different sensors of the following types: temperature, light, humidity, carbon monoxide (CO) gas, motion, smoke, door and fan to exploit five MQTT network attacks. This dataset removes features such as source and destination IP (Internet Protocol) addresses, port addresses and communication times among others that can be found in other datasets and focuses mainly on MQTT-based features.

- **N-BaIoT** [22]**:** The Network-based Detection of IoT (N-BaIoT) dataset was created using nine IoT devices, namely, two doorbells, one thermostat, one baby monitor, four security cameras and one webcam. These devices were of different makes and models. The network setup was exposed to two types of botnet attacks: Mirai and BASHLITE. Each of these botnets has other attacks, as specified in **Table 1**. This dataset comprises both benign and attack traffic intended for the study and detection of botnet attacks.

- **ToN_IoT** [23] is a dataset that aims at addressing the properties of both IoT and IIoT by collecting data from telemetric sources, operating systems and network data, hence the name ToN_IoT. Nine types of attacks were studied on the seven types of sensors specified in **Table 1**. This dataset explores the interaction of network elements across the edge, fog and cloud layers and tries to provide data for intrusion detection in large-scale IoT network scenarios.

References

1. Ashton, K. That 'Internet of Things' Thing. RFID JOURNAL, 22 June 2009. Available online: https://www.rfidjournal.com/that-internet-of-things-thing (accessed on 20 June 2021).
2. CISCO. Cisco Annual Internet Report (2018–2023) White Paper; CISCO: San Jose, CA, USA, 2020.

- **Edge-IIoTset** [24]: This is another dataset that was created to study IoT and IIoT devices and networks. Its design architecture consists of seven layers and 12 IoT (e.g., sound detection sensor, ultrasonic sensor, etc.) and IIoT devices (servo motor, stepper motor, etc.) The testbed was tested with 15 attacks which were categorized into 5 broad attack categories.
- **CICIoT2023** [25] is an IoT-based dataset that is the largest (as of 2023) in terms of the number of devices used to set up the network topology and the number of attacks studied. A total of 105 devices were used to design the testbed, and 33 attacks were carried out in the network for data collection, which were broadly classified into 7 attack categories. These attacks were carried out on the IoT devices using other IoT devices. This dataset also included Zigbee and Z-wave devices along with other IoT devices.

Table 1. IoT datasets summary.

	Year	Testbed Setup	Device Used	Attacks	Normal Traffic Gen Tool	Attack Traffic Gen Tool	Network Sim Tool	Packet Capture Tool
Bot-IoT [26]	2018	Virtual	5 devices simulated: smart refrigerator, smart garage door, weather monitoring system, smart lights, smart thermostat	Information gathering (service and OS scanning), denial of service (TCP, UDP, HTTP DoS and TCP, UDP, HTTP DDoS), information theft (keylogging, data theft)	Ostinato software [27]	Hping3 [28], Nmap [29], xprobe2 [30], golden-eye [31], Metasploit [32]	Node-red [33]	Tshark [34]; features extracted with Argus [35]
N-BaIoT [36]	2018	Real	9 real devices of types: doorbell, thermostat, baby monitor, security camera, webcam	BASHLITE (scan, junk, UDP flooding, TCP flooding, COMBO attack) and Mirai (scan, ack flooding, syn flooding, UDP flooding, UDP plain flooding)	N/A	Binaries and source code of BASHLITE and Mirai, respectively	N/A	Wireshark [37]
IoTNID [17]	2019	Real	2 real devices: Wi-Fi camera, speaker	Scanning (host, port, OS), man-in-the-middle,	N/A	Nmap	N/A	Monitor mode of wireless

3. Heuertz, B.; Velosa, A.; Thielemann, K.; Schulte, W.R.; Litan, A.; Pace, B. Predicts 2020: As IoT Use Proliferates, So Do Signs of Its Increasing Maturity and Growing Pains; Gartner: Hong Kong, China, 2019.

4. Hewlett Packard Enterprise. The Internet of Things: Today and Tomorrow; Hewlett Packard Enterprise: Hong Kong, China, 2019.

5. Ericsson. Connected Industries A Guide to Enterprise Digital Transformation Success A Report on Digital Transformation; Ericsson: Stockholm, Sweden, 2020.

6. The Economist Intelligence Unit. The IoT Business Index 2020: A Step Change in Adoption; The Economist Intelligence Unit: London, UK, 2020.

7. IoT Analytics. State of the IoT 2020: 12 Billion IoT Connections. 2020. Available online: https://iot-

and Information Systems Conference, MilCIS 2015, Canberra, ACT, Australia, 10–12 November 2015; Institute of Electrical and Electronics Engineers: Piscataway, NJ, USA, 2015.

15. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In Proceedings of the International Conference on

	Year	Testbed Setup	Device Used	Attacks	Normal Traffic Gen Tool	Attack Traffic Gen Tool	Network Sim Tool	Packet Capture Tool
				DoS attacks, Mirai (UDP, ACK, HTTP flooding, brute force)				network adapter
IoT-23 [38]	2020	Real	3 physical: speaker, light bulb, door lock	Mirai, Torii, Hide and Seek, Muhstik, Hakai, Internet Relay Chat Botnet (IRCBot), Hajime, Trojan, Kenjiro, Okiru, Gagfyt	N/A	Malware sample in a Raspberry Pi	N/A	Zeek [39]; features extracted with Zeek
MedBIoT [40]	2020	Mixed	80 virtual, 3 physical: switch, light bulb, lock, fan	Botnet malware: Mirai, BASHLITE and Torii	Scripts to trigger actions	Mirai and BashLite source codes, Torii sample	Docker [41]	tcpdump [42]; features extracted with Splunk [43]
MQTT-IoT [44]	2020	Virtual	12 MQTT sensors simulated	Aggressive scan, UDP scan, Sparta Secure Shell (SSH) brute force, MQTT brute-force attack	"Publish" MQTT command	Nmap, MQTT-PWN [45]	Virtual machines, VLC [46]	tcpdump
MQTT set [47]	2020	Virtual	10 simulated devices: temperature, light intensity, humidity, CO gas, motion, smoke, door opening/closure and fan status	Flooding denial of service, MQTT Publish flood, Slow DoS against Internet of Things Environments (SlowITe), malformed data, brute-	IoT-Flock [48]	MQTT-malaria [49], IoT-Flock, Message Queuing Telemetry Transport Security Assistant (MQTTSA) [50]	IoT-Flock	Eclipse Mosquitto [51]

Gener. Comput. Syst. 2019, 100, 779–796.

27. Ostinato Traffic Generator for Network Engineers. Available online: https://ostinato.org/ (accessed on 10 June 2023).
28. Kali Linux Tools. Hping3. Available online: https://www.kali.org/tools/hping3/ (accessed on 11 May 2023).

	Year	Testbed Setup	Device Used	Attacks	Normal Traffic Gen Tool	Attack Traffic Gen Tool	Network Sim Tool	Packet Capture Tool
				force authentication				
ToN_ IoT [52]	2020	Mixed	7 simulated sensors: fridge, garage door, GPS tracker, modbus, motion light, thermostat, weather sensor	Scanning, DoS, DDoS, ransomware, backdoor, injection, cross-site scripting, password and man-in-the-middle attacks	JavaScript in Node-RED	Nmap, Nessus [53], Python script, Metasploitable3, bash scripts on DVWA [54] and Security Shepherd [55], CeWL (Custom Word List generator) [56], Hydra [57], Ettercap tool [58]	NSX-VMware [59], Node-RED	Data logger on Node-RED server, Zeek
Edge-IIoT [60]	2022	Real	12 physical IoT and IIoT devices	DoS/DDoS (TCP SYN, UDP, HTTP, ICMP), information gathering (port scan, OS fingerprinting, vulnerability scan), MiTM (DNS and ARP spoofing), injection attack (XSS, SQL injection, uploading attack), malware (backdoor, password cracking, ransomware)	N/A	Hping3, Slowhttptest [61], Nmap, Netcat [62], Xprobe2, Nikto [63], Ettercap, XSSer [64], SQLmap [65], CeWL, OpenSSL cryptography toolkit [66]	N/A	Wireshark, Zeek and Tshark for feature extraction
CICIoT 23 [67]	2023	Real	67 IoT devices, 38 Zigbee and Z-wave devices	33 attacks in 7 categories (DDoS, DoS, Recon, web-based, brute force,	N/A	Hping3, udp-flood, slowloris, golang-httpflood, nmap, fping [68], DVWA, remot3d [69], BeEF [70],	N/A	Wireshark, tcpdump and dpkt package for feature extraction

42. TCPDUMP & LIBPCAP. Available online: https://www.tcpdump.org/ (accessed on 11 May 2023).
43. Splunk. The Key to Enterprise Resilience. Available online: https://www.splunk.com/ (accessed on 11 May 2023).
44. Hindy, H.; Bayne, E.; Bures, M.; Atkinson, R.; Tachtatzis, C.; Bellekens, X. Machine Learning Based IoT Intrusion Detection System: An MQTT Case Study (MQTT-IoT-IDS2020 Dataset). June

	Year	Testbed Setup	Device Used	Attacks	Normal Traffic Gen Tool	Attack Traffic Gen Tool	Network Sim Tool	Packet Capture Tool
4				spoofing, Mirai)		hydra, Ettercap, Mirai code		
4								

3. Are There Any Similarities or Differences among These Datasets?

To address this question, the IoT-related datasets found in the literature were compared. It was observed that all the datasets surveyed vary in respect to the number and types of devices used in the setup, the type of setup, whether simulation, real or mixed; the attacks the devices were exposed to, etc., as shown in **Table 1**. However, there are similarities among them which are discussed below:

- **Features:** Bot-IoT is the earliest IoT dataset considered and has been utilized by a number of researchers to carry out ML techniques for intrusion detection training. Even though this dataset employs the MQTT protocol, similar to the MQTT-IoT and MQTTset datasets, its feature set has no MQTT-based features such as those found in the latter two which are the only datasets that contain MQTT-related features. From **Table 2**, which shows the features common among the datasets studied, it can be seen that N-BaIoT and MedBIoT have 100 similar features to each other but have no common features with other datasets. Similarly, MQTT-IoT and MQTTset have MQTT-related features that are not found in other datasets. Over 15 features common to the IoT-IoT and IoT-23 datasets were also seen.

  The most common features found amongst the datasets were the five-tuple network flow features (source/destination IP address, source/destination port and protocol) and timestamps. A difference in opinion and research carried out regarding these features has been observed. While some studies, such as [47], removed common features like the source/destination IP and port addresses, as well as communication times, from their MQTTset to allow the identification of features independent of a particular connection/communication, others, such as [71], used these features in the IoT Network Intrusion Dataset to carry out ML training and testing for attack detection. These features, while important in identifying a network flow, carrying out network configurations and troubleshooting, could skew the ML training processes, leading to overfitting and the generation of high prediction rates. Other features, such as sequence or identification numbers, found in IoT-23, Bot-IoT, Edge-IIoT and IoTNID, could have similar effects.

  Most datasets have one or more of the three features (attack, category and subcategory labels) that are used to tag a flow as benign, attack or type of attack. The attack label is used to tag a traffic flow as either benign or attack traffic, which are sometimes denoted as 0 and 1, respectively. On the other hand, the category and subcategory labels are used in datasets where there are a number of different attack types and classes, e.g., the category is used to indicate that a flow belongs to a DoS attack while the subcategory indicates if it was a UDP, TCP, HTTP or ICMP (Internet Control Message Protocol) DoS attack. These features are not used in the

47. Vaccari, I.; Chiola, G.; Aiello, M.; Mongelli, M.; Cambiaso, E. MQTTset, a New Dataset for Machine Learning Techniques on MQTT. Sensors 2020, 20, 6578.
48. GitHub. ThingzDefense/IoT-Flock. Available online: https://github.com/ThingzDefense/IoT-Flock (accessed on 11 May 2023).
49. GitHub. etactica/mqtt-malaria. Available online: https://github.com/etactica/mqtt-malaria (accessed on 11 May 2023).
50. MQTTSA. Available online: https://sites.google.com/fbk.eu/mqttsa (accessed on 11 May 2023).
51. Eclipse Mosquitto. Available online: https://mosquitto.org/ (accessed on 11 May 2023).
52. 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.
53. Nessus. Available online: https://www.cs.cmu.edu/~dwendlan/personal/nessus.html (accessed on 11 May 2023).
54. Kali Linux Tools. Dvwa. Available online: https://www.kali.org/tools/dvwa/ (accessed on 11 May 2023).
55. OWASP Foundation. OWASP Security Shepherd. Available online: https://owasp.org/www-project-security-shepherd/ (accessed on 11 May 2023).
56. Kali Linux Tools. Cewl. Available online: https://www.kali.org/tools/cewl/ (accessed on 11 May 2023).
57. Kali Linux Tools. Hydra. Available online: https://www.kali.org/tools/hydra/ (accessed on 11 May 2023).
58. Ettercap. Available online: https://www.ettercap-project.org/index.html# (accessed on 11 May 2023).
59. VMware NSX. Networking and Security Virtualization. Available online: https://www.vmware.com/uk/products/nsx.html (accessed on 11 May 2023).
60. 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.

61. Kali Linux Tools. Slowhttptest. Available online: https://www.kali.org/tools/slowhttptest/ (accessed on 11 May 2023).
62. Netcat—SecTools Top Network Security Tools. Available online: https://sectools.org/tool/netcat/ (accessed on 11 May 2023).
63. Kali Linux Tools. Nikto. Available online: https://www.kali.org/tools/nikto/ (accessed on 11 May 2023).
64. XSSer. Cross Site "Scripter". Available online: https://xsser.03c8.net/ (accessed on 11 May 2023).
65. Sqlmap. Available online: https://sqlmap.org/ (accessed on 11 May 2023).
66. GitHub. Openssl/openssl. Available online: https://github.com/openssl/openssl (accessed on 11 May 2023).
67. Neto, E.C.P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; Ghorbani, A.A. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. Sensors 2023, 23, 5941.
68. Hping. Available online: https://hping.org/ (accessed on 11 May 2023).
69. Remot3d. Available online: https://kalilinuxtutorials.com/remot-3d-tool-large-pentesters/ (accessed on 11 May 2023).
70. BeEF. Available online: https://beefproject.com/ (accessed on 10 June 2023).
71. Ullah, I.; Mahmoud, Q.H. A Scheme for Generating a Dataset for Anomalous Activity Detection in

and usually supervised learning where the model is trained for the detection of the related attack class, while the label is useful for both supervised and unsupervised learning. In datasets where the labels are not explicitly given, such as in N-BaIoT, MQTTset, etc., the PCAP or comma-separated values (CSV) files are collected and organized separately for each type of attack or normal class for easy identification.

- **Attacks:** This is another important characteristic of an IoT dataset, as this would determine the type of attack an IDS would be able to detect when trained with the particular dataset. **Table 3** shows the types of attacks carried out in the test environment to create the datasets. The attacks have been categorized to show the layer of architecture they belong to. As IoT networks do not have a standardized architecture yet, such as the Open Systems Interconnection (OSI) model used in a conventional network, the attacks have been mapped to the OSI model depending on the layer the attack exploits.

  For example, an application-layer attack targets the highest layer of the OSI model, exploiting the application-level protocols and services. Some of the attacks seen in this category were cross-site scripting (XSS), SQL injection and HTTP DoS attacks. The most common form of transport-layer attacks seen in these datasets were the TCP and UDP DDoS/DoS attacks, which exploit the weaknesses of transport-layer protocols to overwhelm the network resources. Other layered attacks, such as ICMP flood/DoS attacks in the network layer, were observed, while only ARP (Address Resolution Protocol) spoofing was seen in the datalink layer. No physical layers have been studied in these datasets. Other malware or botnet attacks are more difficult to classify as they can span multiple layers!

Table 3. Attack distribution in IoT datasets.

Some datasets, such as N-BaIoT, IoT-23 and MedBIoT, contained traffic related to botnet attacks only. The

Dataset	Attack	A	N	T	D	M
Bot-IoT	Information gathering (service and OS scanning)		✓			
	TCP, UDP DoS/DDoS			✓		
	HTTP DoS/DDoS, information theft (keylogging, data theft)	✓				
N-BaIoT	BASHLITE/Mirai scan		✓			
	Mirai (ack flooding, syn flooding, UDP flooding, UDP plain flooding), BASHLITE (junk, UDP flooding, TCP flooding, COMBO attack)			✓		
	BASHLITE COMBO attack					✓
IoTNID	Scanning (host, port, OS)		✓			
	Man-in-the-middle	✓	✓			
	DoS attacks, Mirai (UDP, ACK)			✓		
	Mirai (HTTP flooding, brute force)	✓				

wrapper-based feature selection mechanisms. Comput. Secur. 2020, 94, 101863.
77. Das, A.; Ajila, S.A.; Lung, C.H. A Comprehensive Analysis of Accuracies of Machine Learning Algorithms for Network Intrusion Detection. In Machine Learning for Networking; Lecture Notes in

Dataset	Attack	A	N	T	D	M	Notes
IoT-23	Mirai, Torii, Hide and Seek, Muhstik, Hakai, Internet Relay Chat Botnet (IRCBot), Hajime, Trojan, Kenjiro, Okiru, Gagfyt					✓	
MedBIoT	Botnet malware: Mirai, BASHLITE and Torii					✓	
MQTT-IoT	Aggressive scan		✓	✓			
	UDP scan			✓			
	Sparta Secure Shell (SSH) brute force, MQTT brute-force attack	✓					
MQTTset	Flooding denial of service,		✓	✓			
	MQTT Publish flood, Slow DoS against Internet of Things Environments (SlowITe), malformed data, brute-force authentication	✓					
	scanning,		✓				
ToN_IoT	DoS, DDoS, and man-in-the-middle attacks		✓	✓			
	Ransomware, backdoor, injection, cross-site scripting, password	✓					
Edge-IIoT	DoS/DDoS (ICMP), MiTM (DNS spoofing)		✓				
	MiTM (ARP spoofing),				✓		
	DoS/DDoS (TCP SYN, UDP)			✓			
	Information gathering (port scan, OS fingerprinting, vulnerability scan),		✓	✓			
	HTTP DoS/DDoS, injection attack (XSS, SQL injection, uploading attack), malware (backdoor, password cracking, ransomware)	✓					
CICIoT2023	ACK fragmentation, UDP flood, UDP plain flood, RSTFIN flood, PSHACK flood, TCP flood, SYN flood, synonymous IP flood			✓			
	ICMP flood, ICMP fragmentation, DNS spoofing, ping sweep, OS scan, vulnerability scan, port scan, host discovery, GREIP flood, Greeth flood		✓				
	SlowLoris, HTTP flood, SQL injection, command injection, backdoor malware, uploading attack, XSS, browser hijacking, dictionary brute-force	✓					
	ARP spoofing				✓		
Will retain MQTT flag		✓	✓				
Will flag MQTT flag		✓	✓				

4. What ML and DL Techniques Have Been Applied to These Datasets for Attack Detection?

Common Features	Bot-IoT	N-BaIoT	IoT NID	IoT-23	Med BIoT	MQTT-IoT	MQTTset	ToN_IoT	Edge-IIoT	CICIoT 2023
Clean MQTT flag						✓	✓			
Reserved MQTT flag						✓	✓			
All 100 of MedBIoT features		✓			✓					
Label/attack	✓			✓		✓		✓	✓	
Subcategory	✓									
Category	✓							✓	✓	✓

r different Any IDS eristics of d, will be age in the ch an IDS

rithms, in underlying

[72] and test the model than their counter-ML algorithms. For these reasons, researchers have adopted a similar approach to DL as they have with ML, which is selecting the minimum and best features of a dataset to train an algorithm. It can be seen in [73], among other studies, that the runtime is reduced with a smaller feature set without (significantly) affecting the efficiency of the algorithm.

Some scientists, on the other hand, have tried to combine algorithms or create different ones similar to ensemble techniques [71][74]. Overall, it was seen from [47][73] and [75], for example, that tree-based algorithms, such as random trees (RTs), random forests (RFs), etc., performed better on average compared to others. Algorithms like Naïve Bayes (NB), though faster, had poorer performance comparatively [52][76][77]. It was also observed that the most commonly used ML algorithms were tree-based, while neural networks (NNs) are the most common for DL algorithms.

Despite various efforts, it was seen that some classes in the datasets did not yield promising results. For example, [75] found the prediction of benign traffic in IoT-23 to be poor, while [76] reported low precision rates for data theft and keylogging attack classes. Understanding the reasons behind these outcomes is important so that the datasets can be improved and newer ones without the same shortcomings can be generated in order to yield better detection results.

5. Any Other Methods Applied to These Datasets for Attack Detection?

It was observed that a different approach from the more traditional ML or DL is on the rise now. Known as federated learning, FL allows participating devices (in this case IoT devices or sensors) to retain their individual data (instead of sharing it with a server or datacenter) and to collaboratively train a shared prediction model. This method promotes privacy as node data are not exposed. Another advantage of this method is that data from devices can be non-IID (independent and identically distributed), meaning the devices could train the model at different times with different data sizes or parameters. This is a huge advantage, as IoT sensors differ in terms of their characteristics and the amount of information they gather.

An increasing number of studies using FL have been seen in the last two years. Seven of the discussed datasets in this study have been explored by researchers using FL. It is more common to see the use of DL or neural networks (NNs) in FL than traditional ML algorithms. This can be accredited to the fact that DL and NN models are better at learning and computing complex patterns in data with the use of multiple layers and deep architectures. This also reduces the need for manual feature engineering, as DL and NN algorithms can automatically deduce important features in the data used. A key difference between FL and ML is the use and transfer of models instead of data between devices and the training/testing server that allows privacy preservation of data. This is made possible with the use of transfer learning, where DL models can be pre-trained and deployed on the IoT devices, thereby reducing the need to train models from scratch. However, despite these benefits, DL algorithms are more resource-consuming compared to ML algorithms, e.g., in terms of training time, memory consumption, computational time, etc., which would add to the overheads of IoT devices, as they are usually limited in resources.