This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

A Survey of Public IoT Datasets for Network Security Research

François De Keersmaeker, *Graduate Student Member, IEEE,* Yinan Cao, Gorby Kabasele Ndonda, and Ramin Sadre, *Member, IEEE*

***Abstract*—Publicly available datasets are an indispensable tool for researchers, as they allow testing new algorithms on a wide range of different scenarios and making scientific experiments verifiable and reproducible. Research in IoT security is no exception. In particular, the design of traffic classification and intrusion detection solutions for network security relies on network traces obtained from real networks or realistic testbeds.**

In this paper, we provide a detailed survey on the existing datasets containing IoT network traffic. We classify them according to several features that help researchers quickly find the datasets that fit their specific needs. In total, we survey 74 datasets that we found by analyzing more than 100 scientific articles. We also discuss the weaknesses of existing datasets, identify challenges, and point to future directions for creating new IoT datasets.

***Index Terms*—IoT, network, security, datasets, taxonomy, traffic generation.**

I. Introduction

With the growing popularity and fast development of the Internet of Things (IoT), the challenges faced by manufacturers and users become numerous and diverse. In particular, the security of IoT devices and applications is a major concern [1]–[4]. A myriad of private and high-value data is collected, carried and transmitted by the IoT, escalating the risk of data breaches. Furthermore, as demonstrated by attacks such as the Mirai botnet [5], devices are not only vulnerable to attacks, but can also be misused to attack other Internet participants. Therefore, the design of new security mechanisms has become one of the top priorities in IoT research [6].

A significant challenge when trying to improve IoT security is the limit in terms of energy and computational capabilities inherent to IoT devices. For example, running a full-fledged anti-malware software, as commonly done on general purpose computers, is not possible on most IoT devices. As a solution, researchers have proposed the usage of Network-based Intrusion Detection Systems (N-IDS) that run on a more powerful device located in a central place in the IoT network, such as a gateway or router [7]–[9]. Such N-IDSs monitor the network traffic arriving to or originating from IoT devices and classify it as benign and malicious, the latter triggering an alarm or a countermeasure. The state of the art is to use binary Machine Learning (ML) classifiers for this task [8], [10], [11].

To train and validate an ML-based classifier, datasets are needed that contain a relevant subset of the data the classifier will be later applied on. In many cases, the definition of whether a classifier is "good" or not is meaningless without the support of a large amount of empirical data. In the case of a N-IDS, the classifier must be trained and validated on a sufficiently large, representative, and correctly labelled dataset consisting of benign and malicious network traffic.

In general, the relevance of a dataset for this task can be measured along three axes, namely *Quantity*, *Quality* and *Availability* [12]. The quantity refers to the size of a dataset, i.e., the number of records it contains. The underlying expectation is that the larger a dataset is, the more likely it contains representative samples. By extension, quantity can also refer to the number of datasets themselves. Quality represents the likeliness of the dataset to produce useful results when used, i.e., to emulate real-world cases. For a dataset consisting of network traffic, it encompasses diverse characteristics, such as the number of different end hosts appearing in the dataset, the duration of the traffic collection, the constraints of the tools used for the collection process, etc. Datasets exhibiting low quality data may induce biased, or even wrong results. Finally, availability refers to the fact that researchers need access to the dataset. Obviously, releasing datasets for reproducibility is an important part of the scientific method [13].

Different approaches exist to obtain IoT-related network datasets. In *testbed-driven* generation, researchers generate datasets by instrumenting an environment containing IoT devices and recording the network traffic under normal conditions, as well as under attacks. This is demanding work, as it necessitates monetary (in terms of hardware) and time (the duration of data capture) resources. In contrast, *synthetic* dataset generation requires less resources as it is based on the simulation or emulation of the IoT devices, the communication network, and the applications running on top of it. This approach is very flexible, but its difficulty lies in achieving a realistic behavior of the different components. Finally, *empirical* datasets are obtained by recording the network traffic of real IoT devices used by real users. Such datasets are "realistic" by design, but collecting them comes with various technical and non-technical challenges, such as privacy concerns.

If researchers cannot collect their own dataset, they can employ a dataset that has been publicly released to the community. Finding such a dataset that fits one's needs for a specific use case can be a tedious task. *The purpose of this*

The first two authors contributed equally to this work.

F. De Keersmaeker is funded by the Research Fellow [ASP] grant 40006064 of the F.R.S.-FNRS, Belgium.

Y. Cao was funded by the ARC project RAINDROP 22/27-126 of the Fédération Wallonie-Bruxelles, Belgium.

G. Kabasele Ndonda was funded by the project SWITCH (#8488) of the Pôle Mecatech, Belgium.

F. De Keersmaeker, Y. Cao, G. Kabasele Ndonda and R. Sadre are affiliated with UCLouvain, Belgium.

article is to help the IoT network security research community by providing a comprehensive survey of publicly available datasets and classifying them along various characteristics. To this end, we analyzed over 100 scientific articles and more than 70 datasets. Not all datasets are suitable for network security research; therefore, we have limited the scope of the article as explained below.

A. Scope

Our paper focuses on datasets that can be used to design and validate network security solutions for IoT networks. Here, the term "IoT network" refers to the network allowing IoT devices to communicate with other devices and services, such as cloud servers or smartphones running companion apps. We are *not* interested in network communication between general-purpose devices, i.e., PCs, servers, smartphones or tablets, even if happening in the context of an IoT application, for example between a database server storing IoT sensor data and a web browser querying that data.

The found datasets mainly represent two IoT domains: Smart Homes (SH) and Wireless Sensor Networks (WSN). The former consist of household objects, e.g. light bulbs, fridge, television, to which networking and remote control capabilities have been given, making them so-called "smart". The latter consist of sensor nodes that measure environmental conditions and send the collected data to a server over a wireless network.

In the context of Industry 4.0, the IoT paradigm has also gained popularity for industrial control and monitoring, forming the Industrial IoT (IIoT) subset of IoT. We also consider datasets from this domain as in scope for this paper. However, we do not include datasets gathered at traditional Industrial Control Systems (ICS) and SCADA systems, such as [14]. While those systems also use connected sensors, they differ in several points from IoT networks, such as in the usage of operator-programmable field devices (PLCs), real-time capable hardware and software, and, in some cases, proprietary network protocols and equipment [15]. Similarly, we do not consider datasets containing the internal communication between Electronic Control Units (ECU) or similar embedded systems, e.g., from the Controller Area Network (CAN) in cars, such as [16].

As this paper focuses on network traffic, datasets containing only the *content* of IoT communications, e.g., telemetry data, are *not* considered here. Such datasets exist and contain, for example, information gathered by sensors in parking lots, records of user interactions with Smart Home devices, health data from smart health devices, or soil information collected by a WSN in a plantation. Similarly, host-centric digital forensics datasets, containing information such as disk images, computer malware binaries, or RAM dumps, are not covered.

It is worth noting that we have included several popular IP traffic datasets that do not contain any IoT traffic. We have done so because these datasets have been used by researchers working on IoT security, either because they did not have access to IoT-specific datasets or because IoT datasets with the desired content did not exist. The datasets in question may still be valuable to IoT researchers, and indeed one could argue that modern IoT devices using Internet protocols such as HTTP might behave similarly to traditional Internet hosts.

Finally, for some of the datasets mentioned, we did not find a way to obtain them directly. Those have been nevertheless included for two reasons. First, they could potentially be obtained by directly contacting the authors. Second, we deemed those datasets contain relevant data or show an interesting measurement methodology which could be reproduced by researchers for similar use cases.

B. Contribution

This article surveys the state-of-the-art in datasets for research in IoT network security. It gives a clear view over the available datasets and their properties. For the latter, we have identified and described in detail 11 characteristics that help researchers to find datasets suitable for a particular use case. In total, we present and classify 74 datasets, focusing on the ones that are related to IoT devices, include network traffic, preferably benign and malicious, and are publicly available.

Furthermore, we discuss issues with existing datasets and point to future directions in the creation of IoT-specific network datasets. Therefore, our survey can also serve as a guide to identify protocols and scenarios neglected so far by dataset creators.

C. Organization

This paper is organized as follows. In Section II, we provide an overview of IoT technologies, in particular the network protocols, since they are an important aspect of IoT network datasets. In Section III, we present related surveys and research. In Section IV, the methodology we followed to perform the survey and to select the relevant datasets is explained. Section V details the characteristics we defined for the classification of the datasets, whereas Section VI presents the result of the classification itself. Then, the datasets are presented in more detail in Section VII (for the IoT-related datasets) and in Section VIII (for the non-IoT datasets). Section IX discusses current challenges and future directions for the generation of IoT datasets. Section X concludes this paper.

II. Overview of IoT

The IoT has evolved rapidly since the concept was proposed 20 years ago, and now it is playing a vital role in many areas, such as e-health, agriculture, industrial manufacturing, Smart Homes, and Smart Cities. With the development and popularization of 5G, it is expected to continuously expand in the future. In this section, we provide an overview of the general concept of IoT and go into more detail about the technologies and protocols used to connect IoT devices.

A. Definition and general architecture

The term "Internet of Things" (IoT) stands for a set of technologies that allow to equip physical objects with sensors and processing ability and to let them exchange information

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

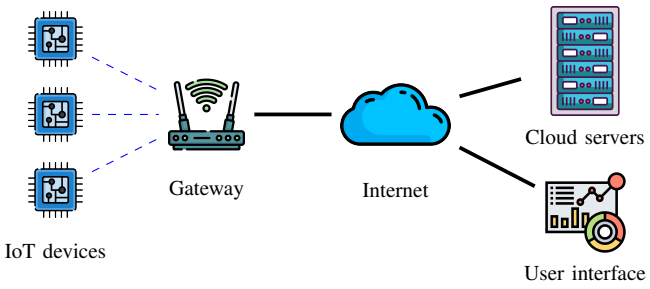
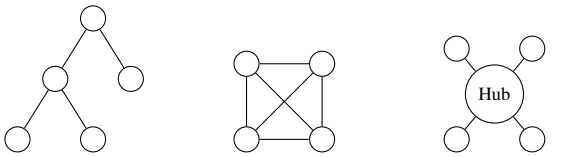


Fig. 1: Architecture of a typical IoT deployment (icons from flaticon.com)



(a) Tree topology (b) Mesh topology (c) Star topology

Fig. 2: End devices sub-network topologies

with other devices over communication networks. Figure 1 shows the architecture of a typical IoT deployment consisting of the following components:

1) *IoT devices* sense, and sometimes also act, in the physical world. They send data to and receive commands from servers and other devices.
2) In order to reduce energy consumption and save resources, IoT devices often use communication technologies and network protocols (see Section II-B) that are not compatible with IP networks. A *gateway* acts as a translator between IoT networks and the Internet.
3) *(Cloud) servers* located in the Internet receive and analyze the data sent by the IoT devices.
4) *Human-machine user interfaces*, such as web-based dashboards or mobile companion applications, allow the users to inspect the data and control the devices.

As explained in Section I, the focus of this paper is on datasets containing the network communication of IoT devices. The latter can take various forms, such as Smart Home devices (light bulbs, refrigerator, power plugs, camera, etc.), Smart City facilities, connected sensors in a WSN, or Smart Health devices. The topology of the sub-networks formed by IoT devices is protocol-dependent, the most popular ones being explained hereafter. If the end devices are able to communicate with each other, two possibilities exist. The network might take a hierarchical structure, forming a tree network (Figure 2a), with a root node and several intermediate layers of devices between the root and the devices at lowest level. If there is no hierarchy between the nodes, and if every node is connected to every other in its surroundings, they form a mesh network (Figure 2b). Both topologies result in a multi-hop network that can cover long distances even with short-range communication technologies. If the devices only communicate with a central communication hub, a star topology is obtained (Figure 2c).

B. Protocols for IoT networks

Devices that do not have resource or energy constraints, e.g., Smart TVs, can employ the same network technologies and protocols as general-purpose computers, that are Ethernet and Wi-Fi, and IP-based protocols, such as TCP and UDP on the transport layer, and HTTP(S) on the application layer.

However, for devices with limited computing capabilities and/or energy constraints, less complex and power-consuming protocols have been designed. In the following, we will present those frequently found in the surveyed datasets. In the well-known OSI model of communication layers [17], the covered protocols appear in the physical layer, the datalink or Medium Acess Control (MAC) layer, the network layer, and the application layer, and will be presented in this order.

1) Physical and datalink layers

Bluetooth: Bluetooth is a short-range radio frequency (RF) communication technology. Bluetooth link control hardware, integrated as either one chip or a radio module and a baseband module, implements the RF, baseband, and link manager portions of the Bluetooth specification. Bluetooth devices are generally organized into groups of two to eight devices called piconets, consisting of a single master device and one or more slave devices [18]. The Bluetooth Low Energy variant (BLE) has been designed to consume less power, and is thus more adapted and optimized for IoT networks [19].

IEEE 802.15.4: Covering the two first OSI communication layers, i.e., the physical and datalink layer, the IEEE802.15.4 standard [20] was developed in 2003 to build low-power and Low-Rate Wireless Personal Area Networks (LR-WPAN). Their range lies between 10 and 100 meters, so they can be easily used in Smart Homes or WSNs where the sensors are located close to each other, but require a mesh topology if long range communication is needed. As the standard only defines the first two layers of the OSI model, additional protocols have been design to implement the upper layers, as described hereafter.

LoRa & Sigfox: Low-Power Wide Area Network (LP-WAN) technologies [21], [22] allow to build IoT network with point-to-point communication ranges up to several kilometers [23], and thus are particularly suitable for Smart City or agriculture applications. The two most popular LP-WANs are LoRaWAN [24] and Sigfox [25]. The former uses a proprietary physical layer modulation, LoRa, above which the free LoRaWAN datalink layer is defined for Medium Access Control (MAC). In contrast, the Sigfox physical modulation is licensed to multiple manufacturers, but the MAC layer is proprietary. Other differences include the data rate, which is higher with LoRaWAN, and the range, which is greater with Sigfox [26]. As both technologies use non-IP stacks, they necessitate a gateway between the LP-WAN network and the Internet.

Z-Wave: Z-Wave is a proprietary protocol for the communication in the unlicensed sub-GHz band [27] and has been originally developed for home automation solutions. It has a range of up to 100 meters for point to point communication and also supports larger mesh networks. It is based on a non-IP protocol stack; IP hosts therefore must communicate to Z-

Wave devices via a Z-Wave-over-IP gateway that understands encapsuled Z-Wave messages.

2) Network layer

Zigbee: Zigbee is an IEEE 802.15.4-based specification by the Zigbee Alliance for the creation of wireless Personal Area Networks (PAN) [28]. It supports star, tree, and mesh networks. The nodes can configure dynamically and have the capability of ad-hoc routing. A designated coordinator node manages the network and is also responsible for the generation of the network encryption key. The original Zigbee design has limited interoperability with IP networks but a specification published in 2013 (Zigbee IP) uses 6LoWPAN and RPL (see below) to route standard IPv6 traffic.

6LoWPAN: The 6LoWPAN specification defines encapsulation and discovery mechanisms that allow to operate IPv6 over IEEE 802.15.4 and other networks [29]. To do so, it compresses the large IPv6 header by assuming default values for header fields unless explicitly stated otherwise, and using lightweight encoding for typical values. Since existing Internet transport and application layer protocols can directly run on 6LoWPAN, the gateway only needs to translate 6LoWPAN packets to full IPv6 packets and vice versa to provide interoperability between the IoT network and the Internet.

RPL: The Routing Protocol for Low-power and Lossy Networks (RPL) is a routing protocol designed for IEEE 802.15.4 based mesh sensor networks and other unreliable networks [30]. In such networks, link failures are frequent and communication is expensive. RPL organizes nodes in a tree-like directed acyclic graph (DODAG) optimized for the communication of IoT devices with external servers.

3) Application layer

MQTT: The Message Queuing Telemetry Transport (MQTT) protocol [31] is an application-layer messaging protocol following the publish/subscribe paradigm. Unlike the other protocols presented in this section, it is not specific to IoT, and usually runs over TCP, but has nevertheless found many usages in IoT applications. It allows for large-scale asynchronous communication between publishers, that send messages related to a topic (e.g., a temperature sensors broadcasting a temperature update), and subscribers, that receive all the messages related to the topic they are subscribed to (e.g., a heater that must switch on or off). Between the publishers and subscribers lies a broker, which stores the topic subscriptions, receives published messages and forwards them to the subscribers [32]. The MQTT protocol is lightweight and open, which makes it suitable for a wide range of applications, including in constrained environments, such as IoT and other Machine-to-Machine (M2M) communication.

CoAP: The Constrained Application Protocol (CoAP) is an application layer protocol specifically design for the communication with resource constrained devices over low-bandwidth networks. CoAP provides a REST-based request/response interaction model [33] that is similar to HTTP, but suitable for devices that do no have the necessary resources to process large HTTP messages. Since CoAP runs on top of UDP, it defines its own mechanisms for reliable communication and uses DTLS for (optional) encryption. Although it is a pull-based protocol by design, an extension allows devices to periodically push measurements to the client, thus avoiding the overhead of repeated requests.

Modbus: Modbus is an application-layer protocol created by Schneider Electric, initially developed for the communication between field devices (e.g., Programmable Logic Controllers) and control servers (e.g., Master Termninal Units) in Industrial Control Systems (ICS) [34]. As it is an open and easy to implement protocol, Modbus is a *de facto* standard for ICS communication, such as in Supervisory Control and Data Acquisition (SCADA) systems. Given its prevalence in ICS, and thanks to one of its variants running on top of TCP [35] which makes it possible to run on any device with IP capabilities, the Modbus protocol is a first choice for IIoT networks [36]–[38].

Modbus is a client/server protocol working in a request/response fashion where the field devices typically act as Modbus servers while the control server is the client [39]. The field devices are connected to sensors and actuators through I/O modules. The state of the I/O ports of a device as well as its internal state is accessible through *registers* that can be queried or modified by the client by sending Modbus requests.

4) Others

A large number of other protocols and network technologies have been used to implement IoT networks, as presented in [40]. In particular, the usage of mobile networks has a long history, be it with the now obsolete 2G, the current LTE-M and NB-LTE (both 4G) and the latest 5G. However, we have not found public IoT datasets from such networks and therefore have not included them here. If a surveyed dataset includes traffic related to a protocol not described here, it will be presented along the dataset in the remaining of the paper.

III. Related Work

To our knowledge, there has been no previous extensive survey of network trace datasets for IoT security. However, there have been publications on other types of devices and other aspects of (IoT) security. In the following, we will present them and discuss their relationship to our work. Table I gives a summary.

Jing *et al.* [41] proposed a categorization of security-related measurement data. For each category of data, the authors provided a specific classification and discussion of its advantages and disadvantages. Four high-level categories of data were presented: packet-level data, flow-level data, connection-level data and host-level data. Packet-level data contains packet header information, packet payload information, and packet activity information. Port mirroring is one of the commonly used packet collection methods. Flow-level data contains information about *flows*, i.e., streams of packets that share one or more attributes. Flow-level measurements reduce the measurement overhead and can be useful if the packet payload is not needed. Connection-based data consists of statistical information about the traffic between two IP addresses. Host-level data is collected on a host device and contains information describing its internal state and its

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

evolution during an attack. The authors do not present any dataset.

Grajeda *et al.* [12] surveyed the available datasets for digital forensics, along with their origin and types. The datasets contain information that is important for the understanding of the behavior and impact of malicious software on a host device, such as RAM dumps, email datasets, apk files, etc. More than 50 datasets were presented, of which only three contained network traffic. Moreover, the authors discussed weaknesses in the existing datasets and future directions, such as lack of variety, data de-identification, and strategies to share complex data.

Chaabouni *et al.* [42] focused on IoT Network Intrusion Detection Systems (N-IDS) based on Machine Learning (ML). The authors explained the specific aspects of learning techniques for IoT security. The survey also presented several free datasets and N-IDS implementation tools for further research. However, since the survey's focus was on detection techniques, only six datasets were discussed.

Buczak *et al.* [43] presented machine learning and data mining (DM) methods for cyber analytics. They discussed in detail the different types of data used by ML and DM approaches (PCAP captures, Netflow, etc.) and mention four public datasets (DARPA 1998/1999, KDD'99, NSL-KDD) containing network traffic from general-purpose computers. Most of these datasets are outdated.

Ring *et al.* [44] provided a survey of more than 30 datasets for IDS research and described the packet-based and flow-based data in detail. The authors proposed 15 different properties that cover a wide range of criteria to assess the existing datasets. Most of them are similar or identical to our own characteristics, e.g., year, public availability, presence of attack traffic, duration, etc. The ones that we do not include were deemed less important (e.g., anonymity, as payload encryption becomes more widespread), equal for nearly all datasets (e.g., presence of normal traffic), or linked to another characteristic (e.g., size of data, which is linked to the recording duration). The survey is comprehensive, but the datasets mostly consist of traffic from general-purpose computer networks and hosts.

Sánchez *et al.* [45] presented a review of different types of devices and state-of-the-art assessing techniques for device identification and misbehavior detection. The authors also discussed 37 public datasets of which 30 contain network data, of which, in turn, only 10 contain IoT traffic (one contains radio spectrum measurements from IoT devices). The other datasets contain network traffic from general-purpose computers and industrial control systems, as well as non-network data (telemetry data, host performance metrics, etc.).

Conti *et al.* [46] proposed a comprehensive overview of Industrial Control Systems (ICS) along with their security challenges, and presented different ICS testbeds and datasets. They reviewed around 50 testbeds, classified along nine features, and over 20 datasets categorized by 12 characteristics. As we did with Ring *et al.* [44], we share some of those characteristics, namely the type of data, recording duration, and presence of attacks. Their characteristics that we did not share include metadata (e.g., dataset name, reference and resource), dataset format (PCAP, CSV, etc., that we detail in the text), percentage of attacks inside the dataset, or evaluation metrics (F1-score, accuracy and precision). We deemed the latter not relevant for a survey, as those values depend on the detection algorithm. The authors also detailed good practices for researchers who want to design a testbed. The paper discusses in detail ICS-specific aspects, such as commercial PLCs, HMI stations, and SCADA protocols, which are less useful in the IoT context.

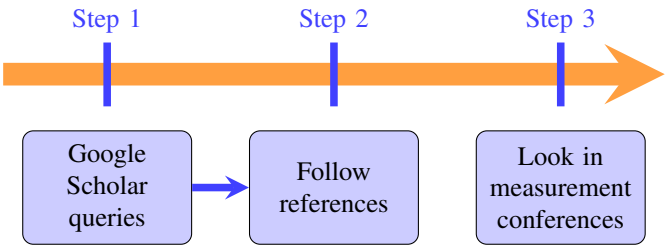


Fig. 3: Methodology for datasets gathering

Rawassizadeh *et al.* [47] surveyed dataset repositories, which are online platforms gathering publicly available datasets. They focused on repositories containing datasets related to mobile, wearable and IoT research. They listed four repositories, namely UCI Machine Learning Repository, UCR Time Series Archive, PhysioBank and CRAWDAD. Other well-known repositories with more general datasets were also mentioned, such as KDnuggets, DRYAD, Datahub, etc. The authors did not detail the characteristics of the individual datasets found in those repositories.

IV. Methodology

This section presents the methodology that we used to find the datasets for this survey. An overview is shown in Figure 3.

In the first step, we queried Google Scholar for articles with general keywords such as "IoT dataset". We selected the most recent ones, going from 2022 back to 2017. For this first step, we limited ourselves to 2017 for two reasons. Firstly, as computer science is a quickly evolving topic, datasets that are only a few years old might already be outdated with respect to current standards and use cases. Secondly, since there are still two steps (see below) after the first one, yielding other datasets to cover, we had to set an arbitrary limit to not end up with an intractable amount of search results. We then made sure the selected articles were relevant by reading the abstract and skimming over the text.

As a second step, we looked for articles referenced by the articles we found in the first step. This led us to older datasets that were used, for example, by the authors as a starting point for their own creations, and to less cited articles that were relevant, but did not appear in the first results of a Google Scholar query.

Finally, as a last step, we explicitly searched for publications in three renowned network measurement conferences, namely the Internet Measurement Conference (IMC) [48], the Passive and Active Measurement (PAM) conference [49], and the Network Traffic Measurement and Analysis (TMA) conference [50]. This allowed us to gather a last set of datasets that did not appear in the first and second steps.

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

Survey	Year	Research Area	Devices/Network	Datasets Mentioned	Description
Buczak *et al.* [43]	2016	Methods for DM&ML in IDS	All kinds of network	✓	DM&ML methods for cyber analytics Well-known datasets for IDS. Challenges for using ML/DM for cyber security.
Chaabouni *et al.* [42]	2019	NIDS	IoT network	✓	Classification for IoT security threats and challenges. ML-NIDS for IoT networks.
Conti *et al.* [46]	2021	Industrial Control System & security mechanisms	Industrial network	✓	Architecture, devices, security protocols for ICS, testbeds and datasets. Best performing ICS algorithm.
Grajeda *et al.* [12]	2017	Digital forensics	Emails, RAM dumps, disk images, network traffic, file sets and etc.	✓	Available datasets presented for digital forensics, along with their origin, testbed and types.
Jing *et al.* [41]	2018	Data categorization	All kinds of network	✗	Security-related data collection and analytics. Requirements proposed for security-related data. Detection methods reviewed for DDoS flooding.
Rawassizadeh *et al.* [47]	2017	Datasets for mobile, wearable and IoT	IoT network	✓	Repositories of datasets for mobile, wearable and IoT research.
Ring *et al.* [44]	2019	Datasets for intrusion detection	All kinds of network	✓	Collection of datasets for intrusion detection. Identified 15 features to assess the datasets.
Sánchez *et al.* [45]	2021	Device behavior fingerprinting	All kinds of network	✓	Review of device types, behavioral data, and processing and evaluation techniques for device identification and device misbehavior detection.
De Keersmaeker *et al.*	2023	General IoT network security	All kinds of network	✓	Collection of the datasets that could be found published or utilized in other papers across multiple domains, along with 13 features for better categorization.

TABLE I: Related surveys (sorted by name of first author)

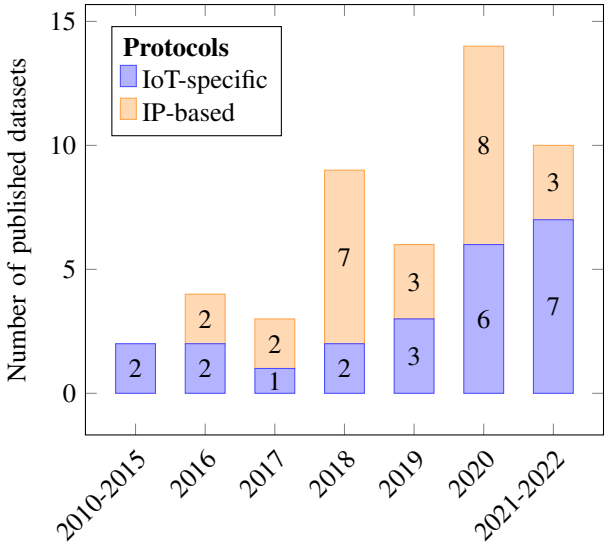


Fig. 4: Number of published IoT-related datasets per year

During the above steps, we also found several dataset repositories, i.e., collections of datasets. When we present the found datasets in the next sections, those repositories will appear in two forms: major datasets hosted by them will be presented in detail (with a reference to the repository). In addition, we will also list the entire repository separately. We consider this an efficient method to include the repositories for those who might be interested, while not spending too much space listing each dataset in each repository.

Figure 4 shows the number of published datasets containing IoT traffic, split into two categories: the ones that cover IoT-specific protocols, and the ones that only cover classic IP-based protocols. We can see an increasing trend over time in the total number of datasets. Yet, their absolute number is still underwhelming, compared to the usage explosion of IoT in the last years. Furthermore, the datasets covering IoT-specific protocols are still in the minority. Given the large number of different protocols, this means that some protocols are studied very little, or not at all. We further discuss this in Section IX-A.

V. Dataset Characteristics

This section will present the characteristics defined and used to classify the surveyed datasets along different axes. In the forthcoming sections, those characteristics will be evaluated on each dataset to highlight their strengths and weaknesses and identify their potential use cases. Figure 5 shows the characteristics, presented as a feature diagram complying with Feature-Based Domain Analysis (FODA) [51], [52].

A. Traffic type, data content, attack scenario

The following three characteristics are used to describe the content of the dataset.

The *Traffic type* indicates if the dataset has been created by instrumenting general purpose devices (e.g., PCs and phones) or IoT devices. If instrumenting IoT devices, the traffic type can be further split based on whether the recorded network protocols include classical IP-based or IoT-specific ones. The IoT-specific protocols have been described in Section II-B.

The *Data content* refers to the nature of the data contained in the dataset, which can reside at different abstraction layers of the communication. This can range from data about the physical electromagnetic signal, to high-level environmental data measured by IoT sensors, or covering traffic data from various layers of the OSI network layer model [17]. In the latter case, the data can be of various precision levels: it can contain the full network packets (including the full payload) in the form of packet captures, only the header of the exchanged packets, or not contain information about individual packets but traffic descriptive features which have been computed on the packets and flows.

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

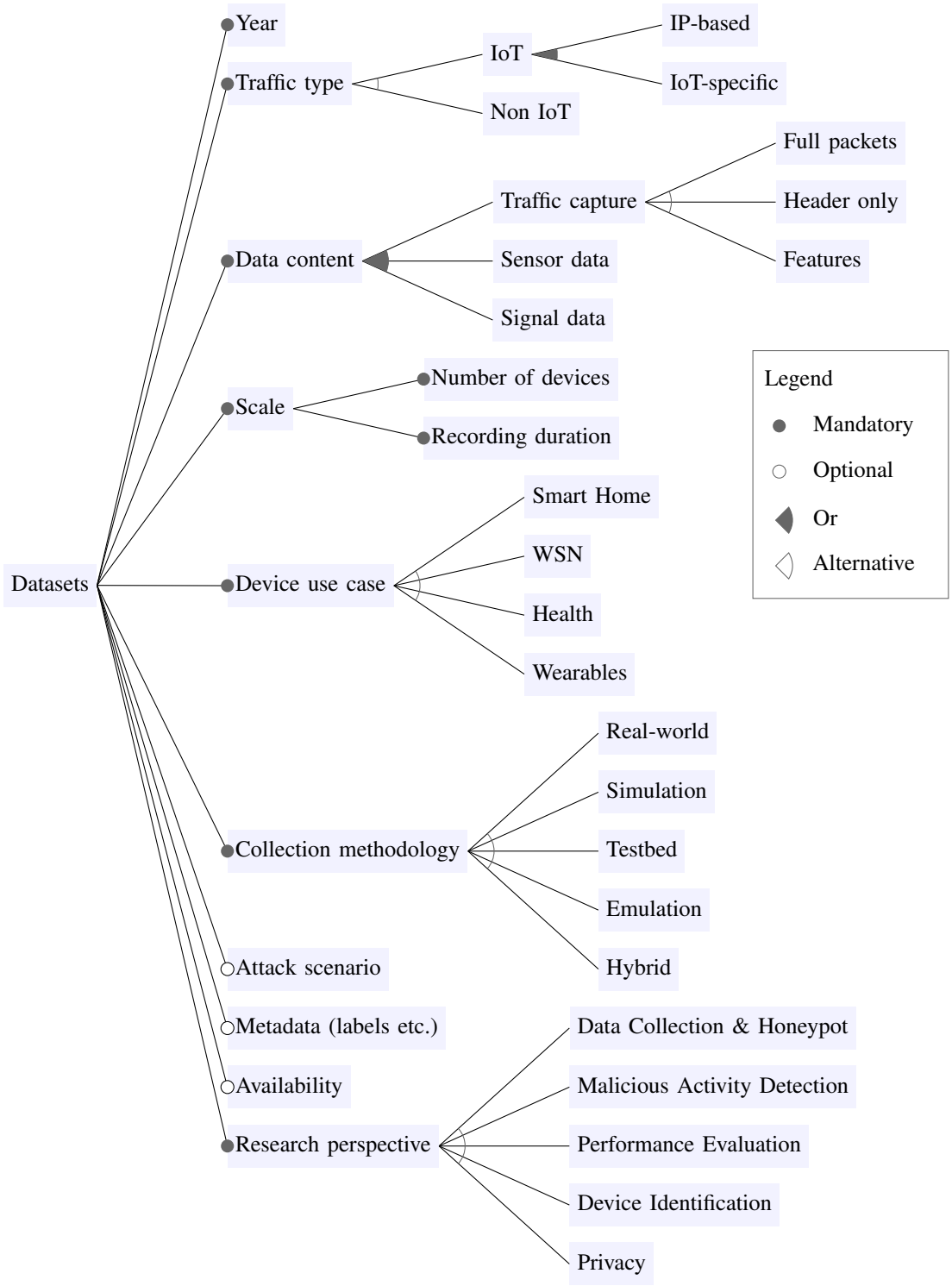


Fig. 5: Dataset characteristics as feature diagram

The *Attack scenario*, finally, is a Boolean value indicating if the covered dataset contains attack patterns or only benign traffic.

B. Scale

The scale of a dataset can be described more explicitly by the number and types of monitored devices and the duration of the data collection.

C. Collection methodology

This feature embodies the methodology the researchers used to collect their dataset. Multiple methodologies are possible, with different accuracies in representing real-world traffic:

- **Real-world network.** The most accurate with respect to real networks, but also the most difficult one, is to directly collect measurements from an existing network. This methodology might require a collaboration with an external organization, e.g., an Internet Service Provider (ISP), to obtain access to an existing real-world network.
- **Simulation.** The researchers used a network simulator, and collected packets from the simulation. This method has the advantages of being easier and cheaper to setup, and also to reproduce. However, the traffic being simulated, it less accurately portrays real-world traffic. Existing network simulators for this purpose include Riverbed Modeler [53] (previously known as OPNET), NS-2 [54] and NS-3 [55], or Cooja [56] for WSNs.
- **Testbed with physical devices.** The researchers constructed an experimental network containing real devices (hosts, servers, routers, switches, etc.), and the dataset was collected by monitoring this network. The advantage of this method is that real devices are being monitored in a controlled environment. It therefore captures well the behavior of the used network protocols. Its drawback is that the deployment of physical devices is expensive and requires a physical place where the testbed will be set up.
- **Emulation.** Instead of using real, physical devices in their testbed, which can be expensive, researchers could emulate the necessary devices with the help of more common hardware. For example, they can instrument general purpose devices or virtual machines to emulate IoT devices, virtual switches/routers by using Software Defined Networking (SDN), or even generate relevant network traffic by using specific traffic generation tools.
- **Hybrid.** To combine advantages of different methodologies, researchers can also instrument a hybrid testbed, composed of real, emulated and simulated devices.

D. Year

As networking and security are two quickly evolving topics, a dataset recorded 10 years ago probably contains traffic patterns that will not occur as much today, or attacks which are not possible to perpetrate anymore due to security updates. As such, the age of the dataset is an important metric to assess the representativeness of a dataset compared to real-world traffic.

E. Presence of metadata

To provide more insight into the network traffic contained in a dataset, some authors process the raw network traces to extract traffic-descriptive *features*, e.g., timestamps and nature of benign or attack traffic, number of bytes or packets exchanged, addresses and ports used, etc. This additional information can help further analysis of the data.

Nowadays, a majority of the collected datasets are used to train and/or evaluate Machine Learning (ML) models for intrusion detection or traffic classification. For such use cases, the data must be *labelled*, i.e., the data points in the dataset must be marked as benign or malicious (for intrusion detection) or as belonging to a specific device type, protocol, or application (for traffic classification).

F. Availability and usage by others

Availability tells whether a dataset is available to other researchers.

In addition, we have checked whether the dataset has been used by other researchers than the ones that have produced it. Although this is not an intrinsic property of the dataset by itself (and therefore not shown in Figure 5), it can indicate whether the dataset is of good quality or has interesting content, and can be used in other cases in a relevant way. However, note that several older datasets, e.g., KDD Cup 99 [57], have been used a lot in the past, not because they are of good quality or relevance, but because researchers did not have a wide corpus of datasets to choose from. Vice versa, recent datasets might not have been used by others, not because they are of bad quality, but because they are too recent and not yet well known. For this characteristic, we searched for papers that cite the paper where the dataset is presented. If one of those papers indicated that it used the dataset in question, there is a match. If there is no match, we indicate it by a question mark. It does not mean that nobody has ever used it, but that we could not detect any usage.

G. Related research perspective

The datasets discussed in this survey were created by their authors for different motives. Therefore, their usefulness to other researchers will depend on the targeted research area. Based on the content of the datasets and the authors' descriptions of their goals and intentions, we identified five main research perspectives that could benefit from the datasets.

- **Data Collection & Honeypot** includes datasets that have been built with the specific intent of creating a dataset to fuel further research, or to fill a gap in the state-of-the-art. Honeypots are intentionally vulnerable systems, deployed specifically to collect data from attackers that will potentially attack the system [58]. As such, they also fit in this category.
- **Malicious Activity Detection** refers to datasets built for the evaluation of security systems which detect or prevent malicious activity, such as Intrusion Detection Systems [59] or firewalls [60]. Those datasets might have been produced to directly evaluate the author's system, or for future usage by other researchers.

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

- **Performance Evaluation** indicates datasets which have been created along the evaluation of a system or protocol, but not necessarily related to security.
- **Device Identification** relates to captures of device activities which have been used or can be used for the development and evaluation of device identification algorithms (e.g. [61]).
- **Privacy** is associated to datasets resulting from privacy analysis of devices, i.e., analysis concerning the disclosure of sensitive user data during device operation [62].

It should be noted that, even if datasets have been created for a specific purpose, they might be used in other research domains if relevant.

VI. Datasets Overview

Before presenting in detail each of the surveyed datasets, we provide a comprehensive and summarized overview of them in tabular form, based on the dataset characteristics introduced in Section V. The idea is that readers who are looking for datasets with specific characteristics first consult the below tables and then read, for the datasets they are interested in, the detailed description in Section VII for the IoT-related datasets or Section VIII for the non-IoT ones.

Tables II and III list the IoT-related datasets, i.e., those datasets that were created by instrumenting IoT devices, following our *Traffic type* characteristic described in Section V-A. They are further split by protocol: Table II presents the datasets containing traffic using IP-based protocols, whereas Table III contains the datasets where IoT-specific protocols have been used, such as the ones presented in section II-B. Table IV presents datasets not directly related to IoT and datasets that were not recorded by instrumenting IoT devices. The relation between the datasets and specific research perspectives or domains are shown in Table V.

The information in the aforementioned tables is further visualized in the form of plots. Figure 6 shows the number of IoT-related datasets containing a specific protocol. As already observed from Figure 4, the majority of datasets contain IP based traffic. Note that we do not include the non-IoT datasets in this plot.

For many use cases, the types of devices and the number of devices or hosts appearing in a dataset are important characteristics. Obviously, researchers building an IDS for Smart Homes are interested in datasets containing traffic from such devices. On the other hand, datasets with a large number of sensors are more likely to be suitable for large-scale experiments. Figure 7 shows the number of devices appearing in each dataset, sorted by general device type. The value on the x-axis of the plot refers to the dataset ID indicated for each dataset in the right-most column of the summary tables II, III and IV. We observe that most datasets contain data from networks with one to 50 devices. Note that some datasets contain more than one type of devices.

VII. IoT Datasets

Even if there exist various communication protocols specifically designed for IoT environments, many IoT deployments use traditional IP-based protocols. The description of the IoT-related datasets is therefore split into datasets containing classic IP traffic and datasets involving various IoT-specific communication protocols, namely Bluetooth, LoRa, Zigbee, etc. In each group, the datasets are presented in chronological order.

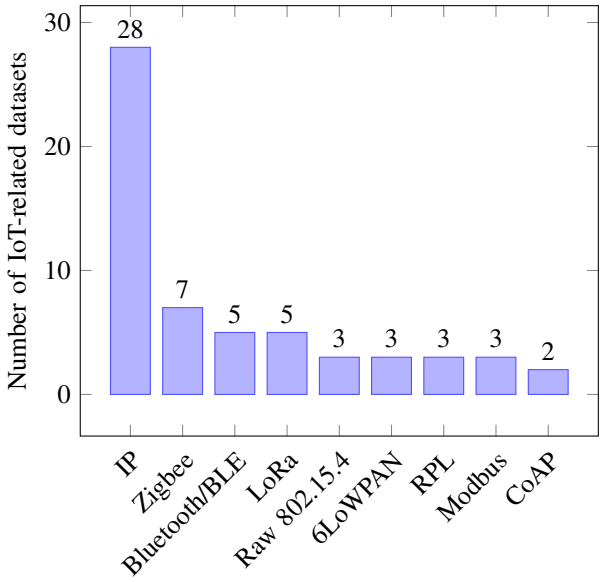


Fig. 6: Distribution of protocols in IoT-related datasets

Some of the presented datasets contain malicious traffic caused by botnet malware that the authors deployed in their test environment. Although the authors used different concrete malware implementations, the base operation of a botnet malware is similar among all malware families: the malware infects poorly-secured devices by brute-forcing default credentials, then waits for further instructions from a *command & control* server, while duplicating itself over the network as a worm [137]. This mode of operation is valid for all the botnet malware covered in the following sections.

A. Datasets covering IP traffic

AWID2 (2015): Researchers at the University of the Aegean, Greece, produced the AWID2 and AWID3 datasets [138], which are of interest for IoT research because they focus on Wi-Fi traffic. The first dataset [63] was produced by Kolias *et al.* [139]. The authors deployed a testbed emulating a home environment. It comprises seven physical devices (one desktop, two laptops, two smartphones, one tablet and one Smart TV) which connect to a single Wi-Fi access point (AP) using unreliable WEP security. Network traffic was generated by web browsing, Internet calls (VoIP), file downloading, and video streaming. Traffic was recorded for one hour by a separate device, the first 35 minutes containing only benign traffic, and the 25 last minutes containing attack traffic as well. Attacks include: flooding attacks, where the network is flooded with a high number of control frames to disrupt legitimate traffic; injection attacks, where the attacker sends a large amount of valid data in a short amount of time; impersonation

Dataset	Protocol	Data content	Attack Scenario	Device Type & Number	Year	Duration	Methodology	Features	Labelled	Availability	Used by others?	ID
AWID2 [63]	Wi-Fi	Features	✓	7 GP	2015	1 h	T	✓(155)	✓	✓	✓	6
Sivanathan *et al.* [64]	IP	Full packets	✓	4+ SH 3 GP	2016	18 min	T	✗	✗	✗	?	34
USC [65]	U	Sensor data	✓	4060 sen.	2016	1 mo	R	✗	✓	✓	?	17
IoT SENTINEL [66]	IP	Full packets	✗	31 SH	2017	U	T	✗	✗	✓	✓	35
Bezerra *et al.* [67]	IP	Full packets Hardware	✓	5 GP	2018	9+ h	E	✓(6)	✓	✓	✓	12
N-BaIoT [68]	IP	Full packets	✓	9 SH 3 GP	2018	U	T	✓(23)	✓	✓	✓	36
Doshi *et al.* [69]	IP	Full packets	✓	3 SH 1 GP	2018	Benign: 10 min Attack: U	T	✓(11)	✓	✗	?	37
DS2OS [70]	Custom	Sensor data	✓	U	2018	24 h	S (DS2OS)	✓(11)	✓	✓	✓	/
UNSW benign [71]	IP	Full packets	✗	28 SH	2016-2017	6 mo 13 d	T	✗	✗	✓	✓	39
UNSW attacks [72]	IP	Full packets	✓	10 SH 4 GP	2018	16 d	T	✓(20)	✓	✓	✓	40
UNSW MUD [73]	MUD	Device profiles	/	28 SH	2018	6 mo	T	/	/	✓	✓	41
Ren *et al.* [74]	IP	Full packets	✗	81 SH 4 GP	2018-2019	4 d 16 h	T	✗	✓	✓	✓	42
Bot-IoT [75]	IP, MQTT	Full packets	✓	U	2019	U	E	✓(32)	✓	✓	✓	/
Kang *et al.* [76]	IP	Full packets	✓	2 SH	2019	U	T	✗	✓	✓	✓	38
Saidi *et al.* [77]	IP	Device signatures	✓	96 SH	2019	7 d	T	/	/	✓	?	43
IoT-23 [78]	IP	Full packets	✓	3 SH 1 GP	2020	21.5 d	T	✗	✓	✓	✓	44
DAD [79]	IP, MQTT	Full packets	✓	16 sen.	2020	7 d	E	✗	✓	✓	✓	18
MedBIoT [80]	IP	Full packets	✓	83 SH 6 GP	2020	U	H	✓(100)	✓	✓	✓	45
Sarica & Angin [81]	IP	Full packets	✓	10 SH 8 GP	2020	~ 1.5 d	S (Mininet)	✓(33)	✓	✓	?	46
UNSW IPFIX [82]	IPFIX	Flow stats	✗	29 SH	2020	3 mo	T	✓(28)	✗	✓	?	47
TON_IoT [83]	IP	Full packets	✓	U	2021	U	E	✓(44)	✓	✓	✓	/
CCD-INID-V1 [84]	IP	Full packets	✓	4 sen.	2021	U	E	✓(83)	✓	✗	?	19
IoTLS [85]	IP, TLS	Full packets	✗	40 SH	2018-2020	~ 2 y	T	✗	✗	✓	?	48
ECU-IoHT [86]	IP	Header only	✓	3 sen. 1 GW 2 GP	2020	~ 3 mo 20 d	T	✓(18)	✓	✓	?	20
MQTTset [87]	IP, MQTT	Full packets	✓	8 sen. 1 GW 1 GP	2020	1 w	E	✓(34)	✓	✓	✓	21
Kolcun *et al.* [88]	IP	Full packets	✗	41 SH 1 S	2021	27 w	T	✓(37)	✓	✓	?	49
27-Botnet [89]	IP	Full packets	✓	10 SH (27 BN)	2021	18 d	T	✓(23)	✓	✗	?	50
ECU-IoFT [90]	Wi-Fi	Features	✓	1 drone 2 GP	2022	3 min	T	✓(10)	✓	✓	?	55

TABLE II: Classification for IoT-related datasets with IP-based protocols

U: Unspecified; /: Not relevant
Protocols BT: Bluetooth; BLE: Bluetooth Low Energy; ZB: Zigbee
Devices type SH: Smart Home; sen.: sensors (& actuators); GP: General Purpose (PCs, servers, smartphones, etc.); HP: honeypot; GW: gateway; BN: botnet
Duration min: minutes; h: hours; d: days; w: weeks; mo: months; y: years
Methodology R: Real-world; T: Testbed; E: Emulation; S: Simulation; H: Hybrid
Features If available, the number of features provided by the authors is indicated.
ID refers to the ID of the dataset on the x-axis of Figure 7.

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

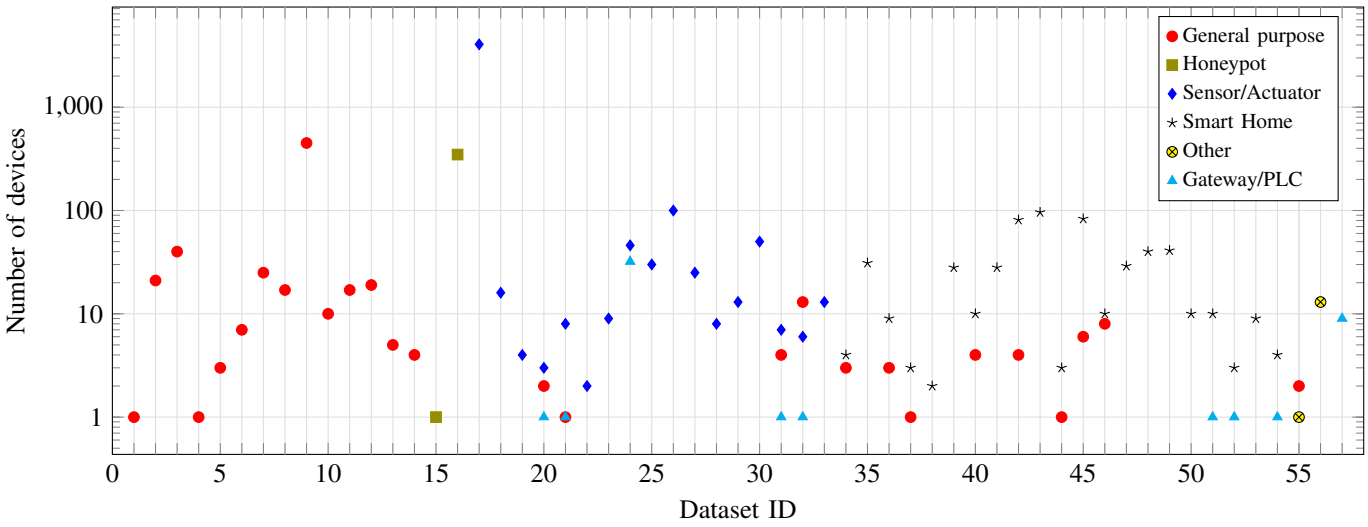


Fig. 7: Number of devices and device types in the datasets. "General purpose" comprises PCs, servers and phones. The "Dataset ID" on the x-axis will be used in Tables II, III, and IV to refer to a specific dataset.

attacks, where a false, malicious AP is introduced in the network.

The dataset, published as CSV files, was divided into two sets containing identical data, only differing in the way attacks are labelled. Both contain a large version and a reduced version, which are further split into a training set and a test set. The training sets of the reduced versions contain 1,795,575 records (1,633,190 benign and 162,385 malicious). Each record has 155 attributes, such as signal or Wi-Fi parameters, timestamps and lengths.

The second dataset, AWID3 [134], will be discussed in Section VIII, as it does not include IoT devices.

Sivanathan et al. dataset (2016): To validate their flow-based network-level Smart Home security system, Sivanathan *et al.* [64] produced a dataset by building a testbed emulating a typical Smart Home environment, including a SmartThings kit, a camera, a smart switch and a motion detector.

Normal traffic from and to the devices was recorded over an 18-minute period. In parallel, a laptop acting as an infected host inside the network launched an attack script executing a three-stage attack. It started with a Simple Service Discovery Protocol (SSDP) scan to discover devices vulnerabilities. Then, remote access to the device was set up with port forwarding. Finally, an external host connected to the compromised devices, which gave it the capacity of launching an attack from the device. The collected dataset takes the form of a raw PCAP traffic trace, comprising 334,088 packets and 327 MB. It was not released publicly.

Urban_IoT_DDoS_Data by USC (2016): Hekmati *et al.* [140] stated that the most important step to defend against DDoS attacks is to detect them successfully. The training of a Neural Network (NN) model for detection will be challenging in the future due to the massive portion of infected IoT networks and the complexity of botnet attacks. The existence of large-scale, relevant datasets is thus critical, but most of the existing ones are not focused on IoT environments. To overcome this, the authors presented a dataset based on a large-scale real-world testbed: an urban IoT system in a large city, which contains 4060 spatially distributed event-driven sensors. The data consists of the binary activity status of each node every 30 seconds, recorded during a month, along with three metadata fields: the node ID, the geolocation of the node (latitude and longitude), and timestamps of the activity status of the node. Additionally, the authors provide a script to generate attacks in the dataset. Whenever the activity status of a node changes, the activity will be recorded as "0" for benign activity, and "1" for attacks.

The dataset is available at [65] as CSV files, along with their attack-generation script.

IoT SENTINEL devices captures (2017): Miettinen *et al.* [141] developed a security system called *IoT SENTINEL*, which identifies the devices present in a network and monitors the traffic from vulnerable devices to ensure the security of the network. To evaluate their system, the authors built a testbed comprising several IoT devices and recorded network traffic to obtain a dataset that they used for their device classification models. They released it publicly as raw packet captures [66].

The testbed consisted of 31 consumer IoT devices covering a wide range of types, such as health, lighting, appliances, cameras, and home automation. Traffic was recorded at device setup, with each setup process repeated 20 times; the dataset therefore does not contain any attack traces. The collected traffic is IP traffic, either directly from the device for the majority, or indirectly from a gateway for those devices that communicate using Zigbee or Z-Wave.

Bezerra et al. (2018): The goal was to create a dataset illustrating the behaviour of IoT devices when they are infected with botnet malware, not only on the networking aspect, but also with observations on their internal functioning [67]. With this intent, the authors instrumented one Raspberry Pi and used it to emulate various IoT devices. They implemented three profiles that represent different types of Smart Home devices:

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

Dataset	Protocol	Data content	Attack Scenario	Device Type & Number	Year	Duration	Methodology	Features	Labelled	Availability	Used by others?	ID
LWSNDR [91]	802.15.4	Sensor data	✗	2 router 4 sen.	2010	6 h	T	✗	✓	✓	✓	/
Fu *et al.* [92]	802.15.4	Perf. metrics	✗	2 sen.	2015	6 mo	T	✗	✗	✓	✓	22
WSN-DS [93]	LEACH	Full packets	✓	100 GP	2016	1 h	S (ns-2)	✓(23)	✓	✓	✓	/
Dowling *et al.* [94]	Zigbee	Attack logs	✓	1 HP	2017	3 mo	T	✗	✗	✗	?	15
Purnama *et al.* [95]	IP, Zigbee	Full packets	✓	5 sen.	2018	50 min	T	✗	✗	✗	?	23
BLEBeacon [96]	BLE	Signal data	✗	46 sen. 32 scanner(RPi)	2016	33 d	T	✗	✗	✓	✓	24
LP-WAN local. [97]	LoRa, Sigfox	Signal data Position	✗	20+ sen.	2017-2018	2 mo 21 d	T	✗	✗	✓	✓	25
RPL-NIDDS17 [98]	6LoWPAN, RPL	Full packets	✓	U	2019	U	S (NetSim)	✓(20)	✓	✗	✓	/
IoT-DDoS [99]	802.15.4, RPL, 6LoWPAN, UDP	Full packets	✓	29 sen.	2019	1 d	S (Cooja)	✓(16)	✓	✗	?	/
LoED [100]	LoRa	Full packets Signal data	✗	8503 dev. 9 GW	2019-2020	2-4 mo	R	✗	✗	✓	✓	57
LoRa radio 1 [101]	LoRa	Signal data	✗	100 sen.	2020	1+ d	T	✗	✓	✓	?	26
LoRa radio 2 [102]	LoRa	Signal data	✗	25 sen.	2021	5 d	T	✗	✓	✓	?	27
cmu/zigbee-smarthome [103]	Zigbee	Full packets	✓	10 SH 1 GW	2020	36 h	T	✗	✗	✓	✓	51
Sadikin *et al.* [104]	Zigbee	Full packets	✓	3 SH 1 GW	2020	U	T	✗	✓	✗	?	52
GHOST-IoT [105]	IP, BT, ZB, RF869	Header only	✓	9 SH 2 GW 3 GP	2019	10 d	T	✓(112)	✗	✓	✓	53
Chen *et al.* [106]	IP, BT, ZB, LoRa	Header only	✓	15 sen.	2020	73 min	T	✗	✓	✓	?	28
Barman *et al.* [107]	BT/BLE	Full packets	✓	13 wear.	2020	98 h	T	✗	✗	/	✓	52
Lacava *et al.* [108]	BLE Mesh	Full packets	✓	13 sen.	2021	6s h	T	✗	✗	✓	?	29
cmu/zigbee-eda [109]	Zigbee	Full packets	✓	4 SH 1 GW	2021	4 d	T	✗	✗	✓	?	54
IoT-CIDDS [110]	UDP, RPL, 6LoWPAN	Features	✓	10-100 sen.	2021	24 h	S (Cooja)	✓(21)	✓	✓	?	30
WUSTL-IIOT-2021 [111]	Modbus	Features	✓	7 sen. 1 PLC 4 GP	2021	53 h	T	✓(41)	✓	✓	✓	31
X-IIoTID [112]	Modbus, CoAP, MQTT, etc.	Features	✓	6 sen. 1 PLC 13 GP	2021	114 d	H	✓(67)	✓	✓	✓	32
CoAP-DoS [113]	CoAP	Full packets	✓	4 GP	2022	16 h	E	✓(17)	✓	✓	?	13
Edge-IIoTset [114]	Modbus, MQTT, etc.	Full packets	✓	13 sen.	2022	50 d	H	✓(61)	✓	✓	✓	33

TABLE III: Classification for IoT-related datasets with IoT-specific protocols

U: Unspecified; /: Not relevant
Protocols BT: Bluetooth; BLE: Bluetooth Low Energy; ZB: Zigbee
Devices type SH: Smart Home; sen.: sensors (& actuators); wear.: wearables; GP: General Purpose; HP: honeypot; GW: gateway
Duration min: minutes; h: hours; d: days; w: weeks; mo: months; y: years
Methodology R: Real-world; T: Testbed; E: Emulation; S: Simulation; H: Hybrid
Features If available, the number of features provided by the authors is indicated.
ID refers to the ID of the dataset on the x-axis of Figure 7.

multimedia centre (a device with the purpose of adding capabilities to a regular TV to make it a smart TV); surveillance camera; surveillance camera with additional traffic, which is an augmented version of the previous one, and can be interacted with by using SSH or Telnet. Besides the Raspberry Pi, the testbed also includes four PCs.

To produce the dataset, the network traffic as well as internal metrics such as the CPU and memory usage of the device were recorded. First, recording occurred without any attack, then three botnet scenarios were introduced to generate malicious behaviour. In the first scenario, the Hajime [142], Aidra [143], and BASHLITE [144] botnets were deployed one at a time on the device. The second scenario focused on the Mirai botnet [5]. For the last scenario, the device was infected with four botnets at the same time to observe their interactions. Behaviour was recorded, for each scenario and device profile, for one hour, adding up to a total recording time of nine hours under infection.

The dataset is labelled, indicating if the data was gathered during infection or not. Flow statistics were extracted as NetFlow records. The dataset, containing CSV files describing the device's internal behaviour, labelled flows, and PCAP files

This article has been accepted for publication in IEEE Communications Surveys & Tutorials. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/COMST.2023.3288942

Dataset	Protocol	Data content	Attack Scenario	Device Type & Number	Year	Duration	Methodology	Features	Labelled	Availability	Used by others?	ID
Dartmouth [115]	IP	Header only	✗	Thousands	2009	5+ y	R	✗	✗	✓	✓	/
CAIDA [116]	IP	Header only	✓	U	2007	1 h	R	✗	✗	✓	?	/
NSL-KDD [117]	IP	Full packets	✓	U	2009	U	S	✗	✓	✓	✓	/
CDX 2009 [118]	IP	Full packets	✓	U	2009	4 d	E	✗	✓	✓	✓	/
DARPA 2009 [119]	IP	Full packets	✓	U	2009	10 d	S	✗	✗	✓	✓	/
Waikato [120]	IP	Header only	✗	U	Various	Various	R	✗	✗	✓	✓	/
Kyoto 2006+ [121]	IP	Header only	✓	348 HP	2006-2015	9 y	E	✓(24)	✓	✓	✓	16
SSENet-2011 [122]	IP	Full packets	✓	1 S	2011	4 h	T	✓(19)	✓	✗	✓	1
ISCXIDS2012 [123]	IP	Flow data	✓	21 PCs	2012	7 d	T	✗	✓	✓	✓	2
TUIDS [124]	IP	Full packets	✓	40 PCs	2012	4 w	T	✓(74)	✓	✗	✓	3
ADFA [125]	/	System calls	✓	1 S	2012	U	E	✗	✗	✓	✓	4
CTU-13 [126]	IP	Full packets	✓	10+ GP	2013	100+ h	E	✓(22)	✓	✓	✓	/
UNSW-NB15 [127]	IP	Full packets	✓	3 S	2015	31 h	E	✓(61)	✓	✓	✓	5
UGR'16 [128]	IP	Flow info	✓	25 PCs	2016	21 w	R	✗	✓	✓	✓	7
CIC-IDS-2017 [129]	IP	Full packets	✓	14 PCs 3 S	2017	5 d	T	✓(80)	✓	✓	✓	8
CSE-CIC-IDS-2018 [130]	IP	Full packets	✓	420 PCs 30 S	2018	10 d	T	✓(83)	✓	✓	✓	9
CICDDoS2019 [131]	IP	Full packets	✓	1S 8 PCs	2019	14 h 40 min	T	✓(80+)	✓	✓	✓	/
CIRA-CIC-DoHBrw-2020 [132]	HTTPS	Flow info	✓	14 S	2020	U	T	✓(28)	✓	✓	✓	10
LITNET [133]	IP	Flow info	✓	10+ S	2020	10 mo	R	✓(85)	✓	✓	✓	/
AWID3 [134]	Wi-Fi	Full packets	✓	11 GP 6 S	2021	2 h 20 min	T	✓(254)	✓	✓	✓	11
NF-UQ-NIDS [135]	IP	NetFlow	✓	/	2021	/	/	✓(83)	✓	✓	✓	/
H23Q [136]	HTTP QUIC	Full packets	✓	13 PCs 6 S	2022	1 h 47 min	T	✓(200)	✓	✓	✗	12

TABLE IV: Classification for non-IoT datasets

U: Unspecified; /: Not relevant
Devices type GP: General Purpose; S: server; HP: honeypot
Duration min: minutes; h: hours; d: days; w: weeks; mo: months; y: years
Methodology R: Real-world; T: Testbed; E: Emulation; S: Simulation
Features If available, the number of features provided by the authors is indicated.
ID refers to the ID of the dataset on the x-axis of Figure 7.

Research perspective	Datasets
Data Collection & Honeypot	[76], [94], [96], [97], [100], [101], [102], [103], [105], [109], [111], [113], [115], [116], [118], [119], [120], [121], [124], [126], [127], [136]
Malicious Activity Detection	[64], [65], [67], [68], [69], [70], [72], [75], [78], [79], [80], [81], [83], [84], [86], [87], [89], [90], [91], [93], [98], [99], [104], [106], [108], [110], [112], [114], [117], [122], [125], [123], [128], [129], [130], [131], [132], [133], [135]
Performance Evaluation	[63], [85], [92], [134]
Device Identification	[66], [71], [73], [77], [82], [88], [95]
Privacy	[74], [107]

TABLE V: Datasets related to each research perspective

Data Collection & Honeypot: Pure data collection without a specific purpose. May contain both benign and attack traffic.
Malicious Activity Detection: Any usage related to intrusion/anomaly detection, e.g., ML model training and testing, or IDS evaluation.
Performance Evaluation: Evaluating the performance of different protocols, network architectures, etc.
Device Identification: Categorizing or identifying different devices according to their behavior.
Privacy: Protection of sensitive user data.

gathering the raw traffic, is available on demand [145].

N-BaIoT (2018): Meidan *et al.* [10] created a public dataset containing benign traffic of IoT devices, as well as malicious traffic by infecting the devices with a botnet [68]. The data was collected from nine commercial IoT devices infected by the *Mirai* [5] and BASHLITE [144] botnets. The dataset contains 7,068,606 instances, from which the authors extracted 115 descriptive features, summarizing traffic flows. Those features were extracted by Kitsune, an autoencoder tool [146]. They implemented several attacks including scanning, UDP and TCP flooding, and spamming.

Doshi dataset (2018): Doshi *et al.* [69] generated attack classifier training data in a testbed IoT network containing three consumer IoT devices, that is, a camera, a smart switch, and a blood pressure monitor. The devices were connected to a Raspberry Pi which acted as a Wi-Fi access point.

The collected data consists of benign traffic received from all three devices during 10 minutes, and recorded in a PCAP file. TCP SYN flooding, UDP flooding and a HTTP GET flooding attack were performed by the authors. Those attacks were chosen because they are the most common types of DDoS attacks launched from Mirai botnets. The process produced a dataset amounting to 491,855 packets, where 459,565 packets are malicious and 32,290 packets are benign.

Features extracted from the network traces can be divided into two classes: Stateless features can be derived from flow-independent characteristics, while stateful features capture how network traffic evolves over time.

DS2OS dataset (2018): For the validation of their system for anomaly detection in a Smart Home environment, Aubet [147] built their own public dataset [70] using a simulated environment composed of eight different devices, such as movement sensors, a washing machine, and a thermostat. They then recorded the application layer traffic of the network for 24 hours. The testbed was simulated by using the DS2OS IoT middleware from the Technical University of Munich [148], and thus uses a custom application layer protocol. As a consequence, the dataset only contains high level information about the actions of the devices.

UNSW IoT traffic traces (2019): A research group at the University of New South Wales (UNSW), Australia, publicly released several datasets based on their work [1]. Those datasets will be described in detail hereafter.

The first one is a record of IoT benign traffic traces. Sivanathan *et al.* [149] deployed a testbed of 28 Smart Home devices (cameras, lights, plugs, etc.). A vast majority of their devices use IP networking, either wired (through Ethernet) or wireless (through Wi-Fi). The only exception is the Philips Hue light bulb [150] that communicates using Zigbee with its gateway, the Philips Hue Bridge, which is wired to the IP network using Ethernet and allows the light bulb to connect to the Internet. The data collection lasted 26 weeks, from 10/2016 to 04/2017, forming an extensive, large-scale dataset.

The dataset contains the headers as well as the payload of the traffic records and has been published on the research group's website [71] in the PCAP and CSV format. A drawback of this dataset is that it is not labelled.

UNSW IoT attack traces (2019): The previous dataset only contains normal operational traffic without attacks. The UNSW research group also released a dataset [72] containing attack traffic besides the benign traffic. Hamza *et al.* [151] deployed a testbed of 10 IoT devices, including smart plugs, cameras and light bulbs, and recorded their traffic for 16 days. In parallel, they launched attacks on their IoT testbed. To know which attack to launch, they instrumented an attack tool that first scans for vulnerabilities on the devices, then launches the appropriate attacks, among the following ones:

- Direct attacks: ARP spoofing, TCP SYN flooding, Fraggle (UDP flooding), Ping of Death.
- Reflection attacks: SNMP, SSDP, TCP SYN, and ICMP (Smurf) reflection. attacks

Depending on the maximum traffic rate that the devices can handle, the attacks could be launched at different rates: 1, 10 or 100 packets per second. Interestingly, besides the attacks the researchers performed themselves, other attacks were detected, due to the fact that the devices were intentionally left unsecured. This provides a good depiction of a real-world attack scenario, since literal real-world attacks were recorded.

Alongside the raw packet traces dataset (PCAP files), the researchers also released a derived dataset containing the count of unidirectional network flows and their IP addresses and port numbers, computed every minute. The researchers also provided ground truth in the form of two accompanying files. The first one describes the start and end time of the attacks, influenced flows, attack type, and bit rate. The second one details the PCAP file number containing the attack, and the attacker's and victim's IP address.

IoT MUD profiles (2019): The Internet Engineering Task Force (IETF) developed the Manufacturer Usage Description (MUD) standard [152]. The objective of this specification is to provide a standardized manner for manufacturers to describe the intended network behaviour of IoT devices, such that network administrators can, based on this specification, implement the necessary security measures, e.g., firewall rules. The underlying idea is that IoT traffic, as a special case of M2M communication, is more predictable than the communication patterns of general-purpose computing devices.

Hamza *et al.* [153] developed the open-source tool *MUDgee*[2] that generates such MUD behavioural profiles from packets captures of IoT devices' network activity. Additionally, they recorded network traffic of a testbed composed of 28 consumer IoT devices for a duration of six months, and applied their tool on the collected traces to generate MUD profiles for those devices. The profiles are, once again, available on the research group's website [73]. This dataset, unlike most of the others presented in this paper, does not contain network traffic, but can still be of great use for IoT security research.

Ren et al. information exposure dataset (2019): Ren *et al.* [154] analyzed the exposure of user data in a Smart Home context, i.e., what information is disclosed when Smart

[1]https://iotanalytics.unsw.edu.au/index

[2]https://github.com/ayyoob/mudgee

Home devices communicate with other services in the Internet. They deployed a large testbed of 81 devices in two countries (US and UK) in order to analyze regional differences. The authors selected the devices by choosing the most popular ones on retail websites from six categories: cameras, smart hubs (e.g., gateways for non IP devices such as smart lamps), home automation (e.g., plugs, motion sensors, lamps), TVs, audio (i.e., connected speakers), and appliances (e.g., fridge, microwave, dishwasher). They all communicate over IP.

Network traffic data was collected at a server that acted as a network access point. For each device, multiple collection scenarios were implemented:

- Power experiments: traffic captured for two minutes when the device is powered on.
- Interaction experiments: traffic captured when a user interacts with the device. The interaction can take different forms depending on the device: local actions (e.g., physical interaction or voice command directly to the device), LAN app action (using a companion smartphone app in the same network as the device), cloud app action (using a companion app connected to a different network), and voice command using an intermediary device, such as Amazon's Alexa.
- Idle experiments: traffic captured when the device is idle, to analyze the device's background traffic.
- Uncontrolled experiments: the devices were deployed in a publicly accessible testbed on a university campus, where people could come and use the devices for their needs, while the devices' traffic was recorded.

In total, 34,586 controlled experiments were performed, as well as 112 hours of uncontrolled experiments, between September 2018 and February 2019. The dataset, containing network traffic recorded in the aforementioned scenarios, is available upon request to the researchers. Instructions are available on their website [74].

Bot-IoT (2019)*:* Koroniotis *et al.* [155] produced a dataset, *Bot-IoT* [75], specifically aimed towards attacks perpetrated by botnets in IoT environments. Their emulated testbed was deployed on the university computing infrastructure. It contains VMs to emulate the legitimate and malicious hosts, whereas the IoT devices were emulated using the *Node-RED* tool [156]. Those emulated devices acted as Smart Home devices, namely smart fridge, smart garage door, weather station, smart lights, and smart thermostat. The IoT traffic consisted of MQTT messages exchanged over classic IP networks.

Alongside the simulated IoT devices, the Ostinato tool [157] was used to simulate benign traffic between the legitimate VMs. Moreover, the malicious VMs were used to simulate an attacking botnet. The launched attacks were dispatched into three types, namely probing attacks (port scanning and OS fingerprinting), (distributed) denial of service, and information theft (data theft and keylogging). Finally, descriptive features were extracted from the recorded traffic using the *Argus* tool [158].

Kang et al. IoT Network Intrusion Dataset (2019)*:* Kang *et al.* produced a small-scale dataset containing benign and malicious IoT network traffic from a testbed composed of two Smart Home devices, namely a smart speaker and a camera [76]. Besides the normal traffic of the devices, four attack types were implemented: Man-in-the-Middle, Denial-of-Service, scanning, and Mirai botnet. The provided dataset takes the form of 42 PCAP files containing the raw network traffic, with an accompanying metadata file detailing the attacks types, involved devices, timestamps and packets.

Saidi et al. IoT devices signatures (2020)*:* As part of their research about detection and identification of consumer IoT devices "in the wild", i.e., from an ISP perspective without proper information about the address and the type of devices, Saidi *et al.* [61] developed a method for identifying and producing a signature of IoT devices, based on sampled information about the destination of their network traffic. The set of signatures they produced has been released publicly [77], with each signature consisting of the following fields: the destination domain as well as the domain group (where multiple similar domains are grouped), the destination port, the category and name of the device, all gathered into a CSV file. The IP address corresponding to the domains can be determined by using DNS.

The testbed instrumented for those measurements consisted of 96 consumer IoT devices from 40 vendors, 56 of which are unique ones, covering home appliances, surveillance, audio, video, etc. Network communication was recorded at European ISP and IXP vantage points, for a total duration of seven days in November 2019, the first four days being used for active experiments, and the last three ones for idle measurements.

IoT-23 (2020)*:* Garcia *et al.* developed another publicly available dataset containing IoT network traffic, benign and malicious, named *IoT-23* [78]. The dataset was built with the specific intent of being used to train and evaluate ML algorithms for malware detection. It is divided into 23 different scenarios: 20 of them are the "attack" scenarios that contain malicious traffic, and the three remaining ones only contain benign traffic. All communication uses the IP protocol.

For the benign traffic, a small-scale testbed was instrumented, comprising three Smart Home devices, namely a Philips Hue smart lamp, an Amazon Echo, and a Somfy smart door lock. Network traffic was recorded. The recording duration was 1.4 hours (8,276 packets) for the Somfy lock, 5.4 hours (398,000 packets) for the Amazon Echo, and 24 hours (21,000 packets) for the Philips Hue. For the malicious traffic, the authors infected a Raspberry Pi with different malwares, including Mirai [5], Torii [159], and BASHLITE [144]. They recorded the network traffic for periods ranging from one to 112 hours, depending on the network activity of the malware, i.e., a malware producing a high packet rate needs less recording time than a quiet malware.

As the dataset is intended to be used by ML algorithms, network flows have been labelled to indicate whether they are benign or malicious and to which malware type they belong.

DAD (2020)*:* Vigoya *et al.* [160] produced the public *DAD* (Dataset for Anomaly Detection) containing labelled IoT network traffic data [79]. Their intent was to produce a sufficiently large-scale dataset, annotated and with diverse scenarios, that can be used by ML algorithms to detect anomalies

in IoT sensor networks. The dataset was collected in a virtual environment emulating the 16 IoT temperature sensors of a data center. The sensors send measurement samples every five minutes, and communicate through MQTT over IP. Network traffic was recorded at the MQTT broker for seven days.

Attack traffic was created by altering the sensors in three different ways on five of the seven recorded days of traffic: *Interception* (some temperature measurements are not sent), *Modification* (temperature measurements are changed), *Duplication* (sending more packets than normal).

The full dataset contains 101,583 packets, which were accordingly labelled as *normal* or *anomaly*. 63.3% of the packets are MQTT packets, among which 16% are malicious.

MedBIoT (2020): Guerra-Manzanares *et al.* [161] argued that the datasets aimed for ML-based IDSs for IoT environments were scarce and small-sized. In order to address this issue, they created the public *MedBIoT* dataset [80]. The labeled dataset combines real and emulated IoT devices in a network where 83 IoT devices are deployed, 80 of which are emulated in Docker containers, and the three others are physical devices. The latter comprise two smart switches and one smart bulb. Malware (Mirai [5], BASHLITE [144] and Torii [159]) was deployed in order to generate attack traffic.

Data was collected from all the endpoints and servers during the botnet propagation. The researchers focused on the early stage of botnet deployment, i.e., spreading and C&C communication. The network was divided into three parts: Internet network, monitoring network and IoT LAN network. The Internet network provides Internet connectivity, while the monitoring network provides storage and processing for the data collected from the network switch. The IoT LAN network allows to let the malware spread in a contained manner. ML-helping features were derived from the raw data (PCAP files) by using the Kitsune auto-encoder tool [146]. In total, 100 statistical features were extracted, computed over different time windows.

Sarica and Angin dataset (2020): Sarica and Angin [162] argued that no dataset existed for a scenario where IoT is used along Software Defined Networking (SDN), and that this use case is quite different from the case of standard, non-IoT networks. As such, they produced their own dataset, instrumenting IoT devices in an SDN environment. In fact, they released two versions of their dataset as CSV files [81], differing only in the number of devices used in the testbed. The purpose of having two separate datasets is to use the second one to test an ML model that was trained on the first one.

The *Mininet* tool [163] was used to create a virtual network where simulated IoT devices periodically send small amounts of data. Their first dataset makes use of five such simulated devices, and their second one initially contains ten devices, with two of them being disconnected after a predetermined time. The environment also includes emulated benign and malicious hosts, as well as a server acting as the IoT data sink. All devices are connected to an Open vSwitch, managed by an SDN controller running Open Networking Operating System (ONOS) [164].

The datasets include traces containing only benign traffic, as well as traces with malicious traffic. For the benign traffic, the authors recorded the network traffic for 30 to 35 minutes, and repeated the experiment 20 times. The 15 first experiments contain IoT traffic and background traffic, whereas the five last ones only contain IoT traffic. For the malicious traffic, five different attack types were implemented and launched by four hosts inside the network, targeting the server or the IoT devices. The different attack types are DoS, DDoS, Port scanning, OS fingerprinting, and Fuzzing.

In parallel to the traffic capture, the SDN controller processed the traffic to create descriptive features. The outcome is a set of 33 features, which are also available and detailed alongside the raw datasets.

IoT IPFIX records (2021): The Internet Protocol Flow Information Export (IPFIX) protocol [165] has been developed by the IETF to provide monitoring devices a standardized way to export information about IP flows. As Internet Service Providers (ISP) might have difficulties identifying devices in their network, more precisely Smart Home devices that hide behind Network Address Translation (NAT), Pashamokhtari *et al.* [166] proposed methods to identify such devices based on their IPFIX records. To evaluate their solution, they produced a public dataset [82].

The authors instrumented a testbed comprising 29 physical Smart Home devices, including cameras, speakers, smart TV, light bulbs, vacuum cleaner, etc. The captured data spans over January, March, and April 2020, and contains network flows caused by user interaction (e.g., changing a light bulb color or voice interacting with a smart speaker), as well as autonomous background device traffic. The total number of recorded IPFIX records is over nine million. The authors also derived 28 flow-descriptive features, amongst others the total number of packets, the maximum packet size, and statistics about the packet inter-arrival time for both directions (inbound and outbound).

TON_IoT (2021): The *TON_IoT* (Telemetry of IoT devices, Operating Systems and Network traffic) dataset [83], [167] is a very broad dataset collected in a the testbed simulating a real-world smart city network, with Edge, Fog, and Cloud computing services. Software Defined Networking (SDN), Service Orchestration (SO) and Network Function Virtualization (NFV) were used to implement the testbed. IoT devices were emulated using Node-RED [156] in the same way as the Bot-IoT dataset [155] (see above). The simulated devices comprise the same set of devices as in the *Bot-IoT* dataset, as well as a Modbus device (industrial control device) and a GPS tracker.

The dataset includes benign and attack traffic. The attacks cover a broad range of nine families: Scanning, Denial of Service, Distributed Denial of Service, Ransomware, Backdoor, Injection, Cross-site scripting, Password cracking, and Man-in-the-Middle. Flow features were produced using Zeek [168]. 44 features were extracted, and divided into four groups:

- Connection features (e.g., source and destination IP addresses and ports, protocols, number of exchanged bytes);
- Statistical features (e.g. number of packets);

- User-centric features, i.e. information about the application-level services (e.g., DNS, SSL and HTTP features);
- Violation features, i.e., anomalies.

CCD-INID-V1 (2021): Liu *et al.* [84] showed that existing intrusion detection technologies have presented several drawbacks, such as the difficulty to be applied on a dynamic and scalable environment like modern IoT, or the compromise between the resource constraints of IoT devices and the ever-growing amount of traffic data. They created the CCD-INID-V1 dataset and used it, along with two other ones, to evaluate their lightweight ML model for feature selection and attack detection and classification in IoT networks.

CCD-INID-V1 contains network data generated in a testbed which aims to emulate real-world traffic in two scenarios, namely Smart Homes and smart labs. The testbed consists of four Raspberry Pis with attached sensors, acting as IoT sensor devices. The IoT devices were configured to collect temperature measurements in the Smart Home use case, and both temperature and pressure measurements in the smart lab use case. The readings are then transmitted to a cloud server using HTTPS over a Wi-Fi IP connection.

The authors selected and implemented five attacks, namely ARP Poisoning, ARP DoS, UDP Flood, Hydra Bruteforce with Asterisk protocol, and SlowLoris. The dataset records have been labeled accordingly as benign or malicious. Additionally, the authors used NFStream [169] to extract 83 features, including source and destination IP and MAC addresses, packet and bytes count, and various timestamps. This dataset has not been publicly released yet, but can be retrieved upon request to the authors.

IoTLS (2021): Paracha *et al.* [170] studied the TLS behaviour of consumer IoT devices to analyze their compliance with security standards and their potential vulnerabilities to TLS attacks. To this end, they used a set of 40 Smart Home devices supporting TLS, including cameras, smart TVs, audio devices, smart hubs, home automation devices, and appliances, disposed into an apartment environment.

Traffic was recorded on a gateway providing Internet access to the testbed. The measurements were divided into *passive* and *active* measurements. The former indicates that the network traffic generated by the devices was recorded when they were left idle or when study participants interacted with them in an unsupervised way. For the latter, the researchers impersonated a TLS server to interact with the devices, and intercepted the traffic with *mitmproxy* [171]. The passive measurements lasted approximately 2 years, from January 2018 to March 2020, depending on the device, whereas the active measurements took place during one month in March. They made their recorded data publicly available [85].

ECU-IoHT (2021): As argued by Ahmed *et al.* [172], there is a lack of datasets with smart health devices. Therefore, they developed the public dataset *ECU-IoHT* [86], which covers attack traffic against the Internet of Health Things (IoHT). The testbed used to produce this dataset was built using the MySignals eHealth kit [173] (now discontinued). It contains a broad range of IoHT devices, and allows to store and process data on the MySignals cloud platform. Specifically, the authors used three sensors in their test setup:

- a temperature sensor, which sends data every 10 seconds;
- a blood pressure sensor, which sends data more sparsely, but generates large quantities of values;
- a heart rate sensor, which sends data extremely frequently.

Those sensors send their data to the MySignals kit, either via IP or Bluetooth, then the latter acts as a gateway and transfers the data to the cloud platform over IP.

To represent attack traffic, six different attacks were implemented: vulnerability scans with *nmap* [174], ARP poisoning, DoS using Ettercap [175] (an open-source tool for MITM attacks), DDoS Smurf, script injection using the MITMf framework [176] (now discontinued).

Traffic from the MySignals kit was recorded, including the aforementioned attacks, between February 24 and June 12, 2020. 19 statistical features were extracted from the raw captures by using the Argus tool [158], such as the number of flows, bytes, and packets, and the different protocols used. Moreover, the packets were labelled as benign or with the attack they belonged to.

MQTTset (2020): Vaccari *et al.* [177] produced the public MQTTset dataset [87], a dataset dedicated to the MQTT protocol. The authors leveraged a network composed of eight Smart Home sensors (temperature, humidity, smoke, etc.), emulated with the IoT-Flock IoT traffic generator [178], connected to an MQTT broker with Eclipse Mosquitto [179]. Full traffic was recorded at the broker for a duration of one week. It comprises legitimate traffic, as well as different attacks against the MQTT protocol. The dataset is available either as raw packet captures or in CSV format presenting 34 summarizing features, including a label indicating if the traffic is legitimate or malicious.

Kolcun et al. dataset (2021): Kolcun *et al.* [180] studied the identification of IoT devices based on their network behavior. In order to compare the accuracy of different machine learning methods proposed by other researchers, they built a dataset containing data collected over a period of 27 weeks from a large-scale IoT testbed.

Their testbed consisted of 41 different Smart Home devices from six categories: surveillance (Blink camera, Bosiwo Camera etc.), media (Apple TV, Fire TV etc.), audio (Allure speaker, Echodot etc.), hubs (Insteon hub, Lightify hub etc.), appliances (Smart kettle coffee machine etc.) and home automation (Honeywell thermostat, Nest thermostat etc.). Their network traffic was recorded with tcpdump [181] and filtered based on the MAC address into different PCAP files.

Every IoT device in this network was assigned a unique device ID, and features were extracted from TCP/UDP flows, identified by the 5-tuple referencing the source and destination IP address, source and destination port, and the transport protocol. The extracted statistical features depend on the ML algorithm that was used, and include, among others, the number of exchanged packets and bytes, statistics about packet sizes, and time intervals between requests. As the dataset was built for identification and fingerprinting algorithms, it does

not include any attack traffic. The authors released their dataset publicly [88].

27-Botnet (2021): Gandhi *et al.* [89] developed BOND, a deep learning model that detects IoT botnets. They produced their own dataset for the evaluation of their tool. The dataset contains benign and botnet labelled traffic and is practically split in two versions. The base version, 27-Botnet, contains full network logs of 25 botnet families and eight physical IoT devices spanning a 18-day period. The second version is intended to incorporate zero day attacks and is accordingly named 27-Botnet-0Day. It contains traffic generated by sources that are not present in the base version: two botnet families, Mirai and Satori, and two IoT devices. The authors extracted 23 features from the raw traffic data.

ECU-IoFT (2022): Ahmed *et al.* [182] created the ECU-IoFT dataset [90], which contains the network traffic of a wireless drone from Tello [183]. To record the dataset, the drone was connected to a Wi-Fi access point, and controlled by a user using the corresponding smartphone app, while an attacker was launching attacks against the drone in parallel. Three types of attacks were executed: Wi-Fi deauthentication, WPA2-PSK Wi-Fi cracking, and a Tello API exploit. Wireless traffic was recorded on the attacker's machine for a total duration of around 3 minutes (54,492 packets). The dataset takes the form of a CSV file containing 10 features which summarize the recorded traffic, and includes a label that indicates if the traffic is benign or related to an attack.

B. Datasets Covering IoT-specific protocols

In the following, we will review IoT datasets collected in non-IP networks. Such networks are typically used by IoT devices that do not implement an IP protocol stack due to resource constraints.

LWSNDR (2010): Sutaharan *et al.* [184] argued that existing synthetic datasets from WSNs lacked labelling. Therefore, they implemented a single-hop and multi-hop sensor-data network. Data generated from this network is labelled and identified for different types of anomalies. The data consists of measurements collected from real humidity-temperature sensors based on Crossbow TelosB motes [185] during a 6-hour period at intervals of five seconds.

Several controlled anomalous situations were generated. Following [186], anomalies can occur in sensor networks at different scales:

- Individual measurements or network traffic attributes at a sensor node.
- All of the data at a sensor node with respect to neighboring nodes.
- A set of sensor nodes in the network.

The LWSNDR dataset is available at [91].

Fu et al. dataset (2015): In order to have a better understanding of data delivery performance over a point-to-point IEEE 802.15.4 wireless link, Fu *et al.* [187] set up an indoor WSN link, where the joint effect of multiple stack parameters on the performance was studied. An experimental dataset was created during the study and published on [92].

The authors employed a sender-receiver pair of TelosB motes [185], each equipped with a TI CC2420 radio [188] using the IEEE 802.15.4 stack implementation in TinyOS [189]. Seven key parameters residing at different layers are considered:

- PHY layer: distance, transmission power level;
- MAC layer: maximum number of transmissions, retry delay time, maximum queue size;
- Application layer: packet inter-arrival time, packet payload size.

Close to 50,000 parameter configurations were tested and more than 200 million packets were collected over six months. The dataset contains only benign traffic data.

WSN-DS (2016): Almomani *et al.* [190] created a dataset containing normal and attack traffic for the validation of an IDS tailored for a WSN [93]. The authors simulated, with the ns-2 network simulator [54], a network composed of multiple wireless sensor devices using the Low Energy Aware Cluster Hierarchy (LEACH) protocol [191], which is a hierarchical MAC protocol for WSNs. To collect data over the whole WSN, every node monitors the traffic of five of its neighbors, and sends a report to the sink. To include malicious traffic, four types of DDoS attacks were implemented: blackhole attack, where a malicious node gathers and drop all traffic from its neighbours; grayhole attack, where a malicious node gathers and selectively drops traffic from its neighbours; flooding attack, where a large number of routing messages are sent to disrupt the network; and scheduling attack, where the attacker configures the network such that every device tries to transmit at the same time.

The full dataset contains 374,661 labelled records. 23 features were extracted from the dataset, such as the node ID, the number of packets sent and received, and topological information about the WSN.

Dowling dataset (2017): Dowling *et al.* [94] deployed a honeypot that emulates a vulnerable Zigbee gateway, with the intent of recording and analyzing SSH attacks. The honeypot was deployed for three months on an unfiltered network and was only accessible via SSH on port 22. All activities on the honeypot like probes, attempted logins, successful logins and failed attempts were recorded. 368 log files were generated with 6 million lines (368 MB of raw text data). A large variety of attacks were recorded, the most prominent being: dictionary attacks, where attackers try to gain access to the honeypot by trying generic credentials; brute-force attacks, where attackers try to compromise the honeypot with various batch files; recon attacks, which simply try to probe the honeypot for information; botnet attacks, where a botnet tries to compromise the device to add it to the zombies set; and finally launch attacks, which try to make the honeypot launch attacks against other hosts. Even if the honeypot acted as a Zigbee gateway, most of the recorded attacks did not show any interest in the Zigbee capabilities of the target. The recorded attack logs have not been released publicly.

Purnama et al. dataset (2018): Purnama *et al.* [95] discussed proposed network analytics to characterize the typical

behaviour of IoT devices. A dataset was generated in a small-scale IoT testbed, which consists of nine wireless Arduino devices, each being attached to a sensor. The network is star-shaped, with a router in the center, and the devices either communicate over IP using Wi-Fi, or with Zigbee through a gateway. Ten five-minute periods of traffic were recorded (50 minutes in total). The dataset was divided into two parts, the first one containing only benign traffic, whereas the second one also contains DDoS attacks. Based on the collected data, the authors analyzed the vulnerabilities of Zigbee transmissions, namely their encryption, by using the KillerBee tool [192]. The authors have not released this dataset publicly.

BLEBeacon (2018): Sikeridis *et al.* [193] created the public BLEBeacon dataset [96] as part of their research on Bluetooth Low Energy Beacons (small power-restricted Bluetooth devices periodically broadcasting information) . 46 participants carried one Gimbal Series 10 Beacon [194] each, which broadcasted a BLE beacon every second. The beacons were received by 32 Raspberry Pi-based BLE gateways, disposed into the test environment, a three-floor campus building. Those gateways forwarded the received messages to a central server. The experiment took place from September 15 to October 17, 2016. The dataset contains, for each message received by a gateway, its RSSI, as well as the timestamps when a certain beacon arrived inside the coverage of a gateway and left it. This data can be used to derive the presence and position of individuals in an specific environment, as the authors analyzed in [195]–[197].

Sigfox and LoRaWAN datasets for Fingerprint Localization (2018): Algorithms have been designed to derive the position of resource-constrained devices based on LPWAN measurements, but those need comprehensive datasets to evaluate their precision. Aernouts *et al.* [198] conducted a data collection campaign with LoRaWAN and Sigfox devices. They produced three datasets, containing signal characteristics, i.e., Received Signal Strength Indicator (RSSI), the base stations that received the message, and the GPS position of the devices. Two of the datasets were collected with Sigfox devices, one of which in an urban environment and the other one in a rural environment, whereas the last one instrumented LoRaWAN devices in an urban environment. They are all publicly available [97].

Data collection took place from November 16, 2017 to February 5, 2018. The urban datasets were captured in a 52.97 km^2 area in Antwerp, Belgium, whereas the rural ones were recorded in a 1068 km^2 area between Antwerp and Ghent, Belgium. For the Sigfox rural dataset, ten Sensolus Stickntrack devices [199] were used, and placed inside commuting cars between Ghent and Antwerp. Those devices broadcasted their GPS coordinates every ten minutes. This dataset contains 25,638 records. The urban Sigfox dataset contains 14,378 messages, obtained with 20 hybrid SigFox/LoRaWAN sensor devices [200]. The GPS data of those devices were transmitted every 12 minutes. Finally, the LoRaWAN dataset used the same devices. However, thanks to the higher data rate and payload size of the LoRaWAN protocol, the messages contained more information, such as the Horizontal Dilution of Precision (HDOP), a value expressing the error of geographical coordinates. Transmission could happen every minute, resulting in a total of 123,539 messages.

RPL-NIDDS17 (2019): Verma and Ranga [98] proposed a dataset containing network traffic from IoT devices using RPL over 6LoWPAN. The dataset was obtained from a network of IoT devices simulated with the NetSim Simulation Platform [201], containing several sensor nodes, one wired sink node, one gateway, and one router. To include attack traffic, the authors implemented seven types of attack that could be combined:

- Sinkhole: a malicious node attracts all its neighbors' traffic.
- Blackhole: a malicious node drops every packet that it receives.
- Hello flooding: similar to sinkhole, a malicious node floods the network with a good routing information such that multiple nodes consider it as a neighbor.
- Clone ID: a malicious node takes over the same identifier as a legitimate node, to receive the packets destined to it.
- Local repair: a malicious node continuously broadcasts bad routing information, triggering a network topology update on every node.
- Sybil: a malicious node impersonate multiples legitimate nodes and receives the traffic destined to them.
- Selective forwarding (grayhole attack): a malicious node randomly drops packets.

The network traces have been labelled as malicious (with attack type) or normal. Furthermore, 20 features have been extracted from the raw dataset in three categories: basic (size of the payload at different layers), flow-based (related to communication flows between a source and a destination node), and time-based (transmission and reception timestamps). The authors did not release this dataset publicly.

IoT-DDoS (2020): Al-Hadhrami et al. [99] pinpointed the lack of datasets focused on IoT environments for the evaluation of IDS, and argued that irrelevant, outdated datasets, such as the KDD Cup 99 [202] (which will be described in Section VIII), are still used. As a solution, they introduced their own real-time data collection framework, intended for the construction of IoT-related datasets for IDS evaluation.

The authors instrumented a testbed comprising several IoT sensors (Zolertia Z1 mote [203] with humidity and temperature sensors) simulated in the Cooja environment [56], and a border router. They communicate their measurements using UDP over 6LoWPAN, and use RPL for routing, both with IEEE802.15.4 as MAC. In order to simulate the noise level of a real IoT network, two distributor nodes emitting a noise signal were deployed. They recorded network traffic on their testbed for 96 simulated hours.

To emulate a real-world IoT network, four operational scenarios have been implemented. The first one contains only benign traffic, whereas the three others contain three attack variants. The flooding attacks consist of the transmission of a large amount of control messages to a mote for a short duration, disrupting the normal functioning of the device. With selective forwarding attacks, infected nodes choose to drop

certain packets, causing the malfunctioning of the network. Finally, in a blackhole attack, the infected nodes drop every packet and make sure the network disruption is maximal by attracting a large portion of the traffic.

Data labelling was an issue for the authors due to the sheer amount of data transferred from each node and the disparities in the protocols and data transmission. They decided to label the data based on the simulation vector and the timing, as well as the type of attack occurring. Additionally, 12 features were extracted from the collected traffic on the physical (e.g. signal characteristics), network (e.g. number of control messages), and application layers (e.g. temperature and humidity values). The dataset is not public.

LoED (2020): The LoRaWAN at the Edge Dataset (LoED), produced by Bhatia *et al.* [204], is an open dataset [100] with network data from a LoRaWAN sensor network in the urban center of London, UK. The authors recorded, during four months spanning 2019 and 2020, LoRaWAN packets received by nine gateways, positioned in diverse urban locations (five outdoor and four indoor). For each packet, the collected data included, among others, the raw payload, the sender's address, and signal modulation parameters. Additionally, they performed a preliminary analysis to extract information from the recorded traffic, such as the number of packets per day or per node, or the distribution of some traffic characteristics, e.g., packet type, frequency, or LoRa spreading factor.

LoRa radio data (2020): With the intent of fingerprinting seemingly identical LoRa devices, i.e. devices sharing the same model, Al-Shawabka *et al.* [205] collected a dataset containing radio data of a great number of LoRa devices, recorded over several days in two different environments. The dataset is available on request [101].

Their testbed comprises of 100 *a priori* identical sensor devices, consisting of a Pysense sensor [206] connected to a FiPy radio board [207]. For data collection, every device has been configured to transmit ten consecutive bursts, with each burst containing 100 consecutive measurements of temperature, humidity, and device voltage, separated by 10 ms. Each burst is separated from the previous one by one second. The data collection was repeated twice, once in an indoor testbed, and once in an outdoor environment. This amounts to 2000 collected packets per device, which thus makes a total number of 200,000 packets. The datasets were labelled using an additional SigMF metafile [208], a file format describing radio data in a JSON format, which was extended to include LoRa-specific features.

A similar research was pursued by Elmaghbub and Hamdaoui [209], who used a testbed with 25 identical LoRa Pycom devices for radio fingerprinting. Their dataset [102] was recorded in numerous experimental scenarios, varying in the location and distance of the sender and receiver, and recording times. It includes radio information, also presented in the form of SigMF files, as well as descriptive metadata.

cmu/zigbee-smarthome (2020): The Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD) is an online public resource gathering a wide range of datasets covering network wireless data, created by the Dartmouth College [210]. It includes many datasets from several contributing locations along with useful tools for data collection, anonymization, and analysis. Akestoridis *et al.* contributed the *cmu/zigbee-smarthome* dataset [103]. It was captured to analyze the security of Smart Home devices using Zigbee with their Zigator tool [211].

The testbed instrumented to generate the dataset comprises ten commercial Smart Home Zigbee devices, including smart plugs, motion sensors and light bulbs. As a gateway between the Zigbee and the outside networks, a SmartThings Hub was used, of different version depending on the experiment. Network traffic was recorded during 35 hours with an Ettus USRP N210 software-defined radio [212] running the GNU Radio software [213], resulting in 571,509 valid Zigbee packets.

For two experiments, the authors introduced a Personal Area Network (PAN) ID conflict, which can be considered as a type of attack. It consists of introducing a forged beacon in the network that advertises the same PAN ID as the existing network, but a different Extended PAN ID. This might cause some devices to become vulnerable for a short amount of time. Nevertheless, the dataset is provided as raw, unlabeled PCAP files.

Sadikin dataset (2020): Sadikin *et al.* [104] proposed various methods to detect known attacks and new types of attacks on Zigbee-based IoT systems by introducing a novel hybrid IDS. For the evaluation of their security system, they relied on an empirically produced dataset containing Zigbee traffic data. The authors used a Zigbee smart lighting system consisting of three light bulbs and a gateway as intermediate between the Zigbee network and the Internet. All the Zigbee traffic was recorded by a sniffer. The resulting dataset contains benign traffic, produced by interaction between the system and its companion smartphone application, as well as malicious traffic. To craft the latter, the researchers implemented four types of attacks: reconnaissance, where an attacker simply probes the network to gather information; Denial of Service, which tries to disrupt the network's correct functioning; malicious control, where the attacker aims at gaining control over the network by impersonating a benign device; and, finally, device hijacking, where an attacker takes control of a device in the network. After data collection, the authors analyzed the data to distinguish malicious and benign behaviour, based on several traffic characteristics. The recorded dataset was not released publicly.

GHOST-IoT-data-set (2020): In the GHOST EU project [214], Anagnostopoulos *et al.* [215] produced the public *GHOST-IoT-data-set* [105] containing network traffic from a real-world Smart Home environment deployed in a residence with two inhabitants. It comprises nine devices, that is, four motion and two door sensors, an emergency button, a weight scale and a blood pressure meter, that use different network communication protocols, namely IP, Zigbee, Z-Wave, Bluetooth, and RF869. The latter is a proprietary protocol used by the emergency button, implemented over the Industrial, Scientific and Medical (ISM) bandwidth around 869 MHz. Additionally, one gateway is present, to allow communication between the Bluetooth and Zigbee devices and the Internet. It

uses the Point-to-Point Protocol (PPP) [216], which directly connects it to a router in order to access the Internet.

The devices' network traffic was recorded for ten days in October 2019. During the nine first days, the normal behaviour of the devices was recorded, whereas the last day was used to implement seven attack scenarios. Those scenarios included:

- Flooding the network with a high quantity of successive measurements.
- Removing the devices' batteries, to simulate physical battery drain or firmware modification.
- Connection of unknown devices.
- Modification of the devices' location.

The full dataset contains a total of 3,811,419 packets. Alongside the raw packet headers, the dataset also contains auxiliary files aggregating traffic into flows, with descriptive statistical features. However, the attacks are not labelled.

Chen et al. smart city dataset (2020)*:* Chen *et al.* [106] developed methods to detect DDoS attacks in IoT environments by using Machine Learning techniques and to block the malicious traffic using SDN. For the evaluation, they built a public dataset [3] by recording the network traffic of eight "smart poles" placed on their campus. The poles are equipped with sensors, such as cameras, temperature and humidity, air quality, and wind speed sensors, and a Raspberry Pi, which provides multiple network interfaces, i.e., Wi-Fi, Ethernet, Bluetooth, Zigbee and LoRa. The exact distribution of sensors and network capabilities is pole-specific. One of the poles is equipped with a heterogeneous Raspberry Pi-based gateway that can aggregate the measurements from any other pole, using any of the aforementioned network protocols. Additionally, every pole also has a Wi-Fi access point (AP) attached to it, to which pedestrians can connect to and use to browse the Internet.

For a duration of 73 minutes, they recorded sensor network traffic, collected at the gateway pole, and network traffic related to passersby connecting to the APs. Afterwards, three of the smart poles were configured to launch DDoS attacks by flooding the pole gateway with irrelevant sensor data. Among the 31,521 collected sensor packets, 4,864 were attack packets. Besides, classic DDoS attacks, comprising ICMP, TCP SYN, and UDP flood attacks, were launched against the network infrastructure of the testbed. The dataset is labeled.

"Every Byte Matters" - Bluetooth wearable devices (2020)*:* Barman *et al.* [217] produced a dataset with traffic data from 13 wearable Bluetooth connected devices [107], including smart watches, headphones, etc. Seven of them used classic Bluetooth, and seven used BLE. The devices were configured to perform typical actions for such devices — from syncing with the smartphone for the simplest devices to using applications for the smart watches. The first part of the dataset contains 10,700 Bluetooth captures, each having a duration of approximately 30 seconds (total duration: 98 hours). The second part, which reflects more closely a real-world usage of the devices, consists of longer captures of approximately 20 minutes (total duration: 38 hours). The raw dataset is available upon request. It contains neither attack traffic data nor traffic-descriptive features.

Lacava et al. BLE Mesh dataset (2021)*:* As an alternative to classic star-shaped BLE networks, the emerging BLE mesh networks allow for short burst multihop networks, potentially having a broader coverage [218]. Lacava *et al.* [108] proposed an IDS based on pattern classification and recognition of DoS attacks on BLE mesh networks. To validate their system, the authors recorded a dataset in a BLE testbed, that they released together with the source code for their IDS [219].

The testbed is composed of 13 BLE devices, implemented with ESP32 IoT development boards [220], which form a BLE mesh network. The collected dataset is divided into three scenarios, the first one representing a totally benign behaviour of the network, whereas the two others include attacks, respectively a targeted grey hole attack and a black hole attack. Both rely on a malicious node dropping some packets instead of forwarding them. For the former attack, the dropped packets are the ones coming from only one victim node, whereas for the latter, all incoming packets are dropped, resulting in a denial of service in a certain area of the network. For each scenario, data collection lasted for six hours, resulting in a total recording time of 18 hours. The published dataset contains the raw BLE packets in the form of CSV files.

cmu/zigbee-eda (2021)*:* As follow-up work to their aforementioned Zigator tool for security analysis of Zigbee networks [211], Akestoridis and Tague developed HiveGuard, a distributed security monitoring system for the same networks [221]. They produced a public dataset for the evaluation of their system [109]. This dataset contains Zigbee traffic recorded during an energy depletion attack, which intents to quickly consume the battery of IoT devices in the network.

To collect the dataset, the authors instrumented a testbed composed of four SmartThings Zigbee devices, namely two multi-purpose sensors, one motion sensor and one button, connected to one SmartThings hub acting as a gateway. They ran their experiments on one device at a time, making sure the power consumption of the devices was the lowest possible, to precisely evaluate the effect of the energy depletion attack. Traffic was recorded by two Raspberry Pis running Zigator. The energy depletion attack was launched in parallel, and data collection lasted until one hour after the battery of each device had been completely consumed. The released dataset is divided among the four monitored devices, and takes the form of raw PCAP files.

IoT-CIDDS (2021)*:* For the evaluation of their cross-layer IDS "IoT-Sentry", Kamaldeep *et al.* [110] created a public dataset [222]. It contains benign as well as attack traffic data related to five different attacks acting on multiple layers, namely UDP flooding, Blackhole, Selective Forwarding, RPL DIS flooding, and ICMPv6 flooding. To create the dataset, traffic was recorded for 24 hours in a network simulated with Cooja [223], which contains from 10 to 100 Zolertia motes running RPL over 6LoWPAN, the exact number depending on the scenario. The labeled dataset takes the form a CSV file presenting 21 features extracted from the authors' raw dataset.

[3]http://smartcity.cs.nthu.edu.tw/dataset/download.php

WUSTL-IIOT-2021: Zolanvari *et al.* [224], [225] highlighted the lack of datasets specifically built for IIoT network security. To fill this gap, they proposed a small-scale testbed, which emulates a typical ICS: a water level monitoring system. It includes three sensors and four actuators, one PLC, and four other general-purpose devices, one being the attacker. The sensors and actuators communicate using the Modbus protocol. The authors instrumented this testbed and recorded its network traffic during normal operation as well as under attack for a total duration of 53 hours. To generate attack traffic, four attack types were implemented, namely command injection, DoS, reconnaissance, and a backdoor. They publicly released a subset of their recordings as the WUSTL-IIOT-2021 dataset [111]. The dataset takes the form of a CSV file containing 41 features, one being the attack label, that summarize the recorded network traffic inside the testbed. The authors used this dataset in their work on Explainable AI applied to IIoT security [226].

X-IIoTID (2021): Al-Hawawreh *et al.* developed a testbed of brownfield IIoT devices [227] to provide a platform which could be used for subsequent research on IIoT security, in a similar way to the aforementioned WUSTL-IIOT-2021 dataset. Their testbed includes various IP-based protocols, some being general-purpose ones (e.g., HTTP, DNS, WebSocket), whereas others are designed for IoT/IIoT environments (Modbus, MQTT, CoAP). They leveraged the testbed to create the X-IIoTID publicly available dataset [112], [228]. It contains benign and malicious network traffic recorded on the testbed, as well as logs and hardware metrics. Data was recorded during 114 days, but not 24 hours per day. 18 different attacks from nine attack families were implemented to generate attack traffic. As post-processing step, the authors extracted 67 features from their collected data, including 31 related to network traffic, and 3 corresponding to the attack type labels.

CoAP-DoS (2022): The most recent dataset that will be covered is the publicly released CoAP-DoS, by Mathews *et al.* [113], [229]. It provides a dataset fit for ML-based IDSs, and is the only one mentioned covering the CoAP protocol, and more precisely denial of service attacks targeting this protocol. The authors instrumented four devices, namely one Raspberry Pi acting as a server and three computers emulating clients, two being malicious. Those generated normal and DDoS traffic, which was recorded for 16 hours. The dataset is available as a full packet capture containing the total of 661,304 packets, or as 17 traffic features extracted in a CSV file. The malicious traffic is labeled in separate JSON files.

Edge-IIoT (2022): Ferrag *et al.* [230] built and extensive hybrid testbed for the generation and recording of IoT and IIoT network traffic. It spans multiple abstraction layers, from the top cloud layer with orchestrating services, to the bottom perception layer of IoT and IIoT sensors and actuators, passing through different mediator layers. The perception layer contains 13 sensors and actuators, e.g., sound, temperature, and water level sensors, servo and DC motors, etc. With the incentive of creating an IIoT dataset suited for Federated Learning, they instrumented their testbed to produce the Edge-IIoTset dataset [114]. It contains both benign and malicious traffic covering multiple IP-based protocols, the application-layer ones being MQTT, DNS, HTTP, and Modbus.

The traffic collection phase lasted for 50 days, but was not continuous. The full packet captures are provided in the dataset in PCAP format. A total of 14 attacks were implemented to produce attack traffic along the normal device traffic. After the data collection phase, the authors derived 61 features from the traffic, including the attack label.

VIII. Non-IoT Datasets

Datasets without IoT-specific data can be interesting for multiple reasons. First, several of such datasets were used in the past for the evaluation of IDSs for IoT networks at a time when IoT-specific datasets did not yet exist. The most well-known examples of such datasets are DARPA 98 [231] and KDD Cup 99 [202], which are two aged and outdated datasets [232] that have been frequently used by lack of better options. Second, some IoT devices use generic protocols such as TCP/IP and HTTP and are vulnerable to the same malware as general-purpose computers. Datasets containing such traffic can therefore still be useful for IoT research.

In this section, we discuss datasets that do not contain IoT traffic and that were published or updated after 2000.

UCI Machine Learning dataset Repository (Since 1987): The UCI dataset repository [233] is a well-known repository of datasets and data generators for machine learning. The datasets cover a wide range of domains. Most datasets from this repository are labelled and generated with features, and are stored in CSV format.

Dartmouth College campus-wide data (2004): To analyze the evolution of network usage in a campus environment, Henderson *et al.* [234] recorded network traces inside the Dartmouth College campus. Multiple versions of this dataset exist that are publicly available [115]. The last version dating from 2009 includes four different subsets, related to four different data collection methodologies:

- Syslog: Every access point sent messages to a central server whenever a client was (de)authenticated and (de)associated.
- SNMP: AP statistics were collected every five minutes with the Simple Network Management Protocol (SNMP).
- Ethernet sniffers: Network sniffers were installed on multiple routers within the campus, and *tcpdump* [181] was used to record traffic passing through the routers' wired interfaces.
- User movement: For every wireless card seen, the location of the associated AP was recorded to illustrate user mobility.

CAIDA (2007): The CAIDA "DDoS Attack 2007" dataset consists of a one-hour traffic trace from a DDoS attack on August 4, 2007 [116]. For privacy reasons, packet payloads were removed, and the headers were anonymized using CryptoPAn [235]. The dataset therefore only contains anonymized packet headers, presented as 5-minute PCAP files, and is available upon request. The attacks are not labelled, and no features were extracted.

***NSL-KDD (2009)**:* Tavallaee *et al.* [232] proposed an improved version of the KDD Cup '99 dataset [202], the *NSL-KDD* dataset [117]. They removed all the redundant records that biased the prediction results of ML algorithms. Then, they increased the difficulty of subsets of the dataset by designing a subset population algorithm that favours records having a low prediction success rate. With those solutions, the authors generated a training set and a test set containing 125,973 and 22,544 records, respectively, as well as a new test set containing 11,850 records which were wrongly classified by at least one classifier in the evaluation of the initial KDD Cup '99 dataset. However, this dataset still suffers from the inherent problems of the initial DARPA 98 dataset.

***CDX 2009**:* Cyber-security competitions can generate interesting network traffic that can be used by security systems, as argued by [236]. One of such competitions is the Cyber Defense Exercise (CDX) hosted by the National Security Agency (NSA) of the United States [237], where students from military colleges in the US have to build secure networks and defend them against attacks crafted by the NSA.

The traffic occurring during CDX 2009, which took place for four days in April 2009, was recorded and released publicly [118]. It contains labelled, benign and malicious traffic that the students must defend against, targeted towards vulnerable VMs intentionally placed inside the network. The attacks included a 15-minute DDoS SYN flood attack, constant scanning of hosts and services, as well as various manual attacks. The benign traffic was generated by automatic web crawling scripts, and by manual interaction with the web, email, and DNS services of the network. One shortcoming of this dataset is that it has a quite small scale, only limited to the network where the competition took place, and does not including real background Internet traffic.

***DARPA 2009**:* The *DARPA 2009* dataset [238] was developed to provide a dataset containing modern attacks at the time. It comprises benign IP network traffic, interacting with HTTP, SMTP and DNS services. Besides, it also contains traffic related to various DDoS attacks and worms, captured between a simulated /16 network and the Internet, between November 3 and November 12, 2009. A subset of this dataset is available at [119].

***Waikato Internet Traffic Storage**:* The WAND Network Research Group of the University of Waikato, New Zealand, has released several publicly accessible network datasets in the Waikato Internet Traffic Storage (WITS) repository [120]. Their datasets contain IP network traffic, recorded from different networks in New Zealand, for example:

- The Waikato dataset, the last version being Waikato VIII [239]: It contains packet header traces collected at the edge between the University of Waikato and the public Internet between April and November 2011, for a total of 87 days, amounting to 27,870 million packets (corresponding to 16,576 GB of traffic).
- The ISPDSL dataset, the last version being ISPDSL-II [240]: This dataset contains a contiguous recording of packet headers, collected from a New Zealand ISP in January 2010 during 11 days. The total number of packets is 17,798 million, which corresponds to 8,122 GB of traffic.
- The IPLS dataset, also known as the Abilene dataset, the last version being IPLS III [241]: This mostly contiguous dataset contains network traces collected at the Indianapolis router on the Abilene network, a retired American network connecting several US universities [242], for four hours in January 2004. The total number of recorded packets is 3,310 million, corresponding to 2,733 GB of traffic.

Those datasets are provided in the ERF format, and should be processed using the libtrace tool [243], [244].

***Kyoto 2006+**:* Song *et al.* [245] produced the public *Kyoto 2006+* dataset [121] for IDS research. It was updated continuously, and contains now nine years of network traffic capture (2006–2015) . The authors used a wide range of devices, intentionally left vulnerable such that they act as honeypots and attract attack traffic, amounting to 348 honeypots. The authors also deployed a server in the network, acting as a mail and DNS server, to generate benign traffic. The recorded raw traffic was first converted into network sessions with Zeek [168], then analyzed with three distinct security tools (one IDS, one antivirus, and one shellcode detection tool). 24 statistical features were extracted, 14 of them derived from the original 41 features of the KDD Cup '99 dataset (e.g. duration of the connection, service, byte count), and the remaining 10 being new additional features (e.g., attack label, source and destination addresses and ports, detection or not by the security tools).

***SSENet-2011**:* The *SSENet-2011* dataset was created by Vasudevan *et al.* [122] for IDS research. Their testbed was constituted of five geographically separated sites, connected together with a VPN. In one of the location, a vulnerable Windows 2003 server was configured, and all traffic to and from this machine was recorded. In parallel, attacks were launched from the four other locations, including probing, flooding, and privilege escalation attacks. Data collection lasted for four hours. The Tstat tool [246] was used to extract flow information from the raw captures.

***ISCXIDS2012**:* Alongside their research to design a framework for generating relevant and extensible network traffic datasets by following a set of five guidelines, Shiravi *et al.* [247] produced an illustrative public sample dataset [123] using 21 vulnerable Windows workstations. It contains seven days of labelled benign and malicious IP traffic with full payloads.

***TUIDS (2012)**:* Gogoi *et al.* [124] set up a real-world testbed able to generate network traffic, both benign and malicious. Traffic from this testbed was recorded, filtered and preprocessed to produce a labelled dataset. The testbed comprised one router, three switches, one server, two workstations and 40 simple PCs. Six Virtual Local Area Networks (VLAN) were formed with the devices, and connected to each other with a Layer-3 switch. Attack and normal traffic was gathered during a 4-week period. 16 attacks were implemented to form the malicious traffic, such as the well-known Fraggle and

Smurf attacks. The attacks were launched from one VLAN to another, as well as from a remote LAN through the Internet. 50 packet-level and 24 flow-level features were extracted from the raw data, such as addresses and protocols, packet flags and length, as well as time-based and connection-based features. Attack traffic was accordingly labelled. The dataset was split into a training part and a testing part, where the benign and attack traffic each roughly account for half of the data. A subset of the dataset seemed to have been publicly released, but is unfortunately not reachable anymore.

***ADFA (2013)**:* Similar to the above datasets, Creech and Hu [248] created the public *ADFA* dataset [125] as a better alternative to the flawed KDD datasets. A Linux and a Windows version of the dataset exist; the Linux version will be discussed here.

The testbed used for data collection consists of only one Linux web server hosting services such as Apache Web Server, FTP, SSH and MySQL. The server was intentionally left vulnerable by installing TikiWiki [249] version 8.1, which contains a known PHP code injection vulnerability [250]. The dataset contains raw system call traces of benign as well as malicious activities on the web server. Several attacks were implemented, including password brute-force against the FTP and SSH ports of the web server using *Hydra* [251], adding a new superuser by privilege escalation, and injection of the Meterpreter shell [252].

***CTU-13 (2014)**:* García *at al.* [253] argued that botnet detection methods are rarely compared to each other. Therefore, they compared three such methods by using a dataset containing benign and botnet network traffic that they produced and released publicly [126]. The testbed instrumented to collect the dataset contains several Windows XP VMs running on a single Debian host. Those VMs were infected with different botnet malware, and 13 malicious scenarios were implemented, including DDoS attacks, mail spam, and network scans. Traffic was recorded on a router inside the network and labelled accordingly with one among three possible labels: background, normal, or botnet. The *Argus* tool [158] was also used to convert the raw data to *NetFlow* [254] records and to extract flow-descriptive features.

***UNSW-NB15 (2015)**:* Moustafa and Slay [255] produced a dataset containing benign and attack traffic for security research [127]. The traffic was produced using a traffic generator tool, the Keysight PerfectStorm (formerly IXIA PerfectStorm) [256]. This tool generates network traffic imitating real-world interactions, as well as known attack patterns, and is used to test network infrastructures. For the dataset, the tool was configured with two servers producing the benign traffic, and one other server acting as the attacker. Traffic was captured on a router on the path between the servers and the client devices for a total duration of 31 hours (non contiguous), producing a 100 GB dataset. For the first 50 GB, the PerfectStorm tool generated one attack per second, whereas for the last 50 GB, it generated ten attacks per second. The dataset contains 2,218,761 normal records, and 321,283 attack records, including fuzzing, reconnaissance, and DoS attacks, which were labelled accordingly with their start and end times, attack type, addresses, ports and protocols. 49 packet- and flow-descriptive features were generated using Argus [158] and Zeek [168], as well as 12 custom algorithms for additional features.

***UGR'16 (2016)**:* Maciá-Fernández *et al.* [257] were interested in traffic cyclostationarity, a property describing repeating pattern in traffic, such as the difference in network usage between weekdays and weekends. They proposed the public *UGR'16* dataset built with real-world traffic and up-to-date attacks [128].

It consists of two main parts: a calibration and a test set. The calibration set only contains benign network traffic. For this baseline to correctly embed cyclostationarity, traffic capture should last for several cycles (i.e., weeks), and be as representative of the real-world case as possible. The calibration set was collected during 17 weeks, from March to June 2016. The test set combines both normal and attack traffic, and was collected for six weeks between July and August 2016. The network infrastructure used to collect the data is a tier-3 ISP, providing various web services. The authors argued this network is representative for a very wide subset of Internet users. It involves a total of 20 victim and five attacker hosts.

The attacks are limited to network attacks:low-rate DoS, port scanning, and Botnet traffic. The attack traffic was split in two-hour batches, with each attack being perpetrated within a batch, either at a planned or at a random timestamp. Similar to the CTU-13 dataset [253], traffic was labeled as "attack", "normal", or "background" when it is unclear whether it corresponds to an attack or not. The published dataset consists of CSV files with extracted NetFlow flow features.

***CIC datasets**:* The Canadian Institute for Cybersecurity (CIC) at the University of New Brunswick released several datasets with network attack traffic [258]. Four of them will be covered in the following.

***CIC-IDS-2017**:* Sharafaldin *et al.* [259] argued that widely known, existing datasets lacked traffic and attack diversity, lost information due to anonymization, or were simply too small. They created a publicly available dataset [129] collected over five days in July 2017 in a testbed separating the victim network from the attack network. The former comprised two switches, one firewall, three servers and ten PCs. The latter contains one router, one switch, and four PCs running Windows 8.1 or Kali Linux. Benign background traffic was generated by using the authors' B-Profiles [260], which are abstract models for human interaction with the Internet. Seven attack types were implemented:

- Bruteforce: FTP-Patator and SSH-Patator on an Ubuntu web server.
- DoS: Hulk, GoldenEye, Slowloris and Slowhttptest on an Ubuntu web server.
- Heartbleed [261] towards an Ubuntu web server.
- Web Attack: XSS and Brute-Force automated with Selenium towards an Ubuntu server.
- Infiltration Attack: Using Metasploit to verify the vulnerabilities and Nmap for port scanning after the Ubuntu and Mac victims downloaded the infected file.
- Botnet attack: Attack against various Windows OS using Ares, a Python based Botnet.

- DDoS and port scan: The DDoS attack was generated by sending UDP, TCP and HTTP requests using Low Orbit Ion Canon (LOIC) by a group of Windows systems towards the victim servers. The port scan is launched by Nmap over all Windows machines.

Attacks were spread over the four last days of traffic recording. 80 flow-based features were extracted from the raw traffic with CICFlowMeter [262] and labelled with the attack type.

CSE-CIC-IDS-2018*:* This dataset [130], [263] is similar to the previously mentioned CIC-IDS-2017 dataset. Compared to the latter, it has a longer capture period of ten days and was recorded with a larger network layout, still divided among the victim and the attacking networks (420 PCs and 30 servers in the victim network and 50 PCs in the attacking network). The implemented attacks are the same as in CIC-IDS-2017.The authors once more used their network profiles for automatic traffic generation [260]. In addition to the B-Profiles for benign traffic, M-Profiles describe attack traffic.

CICDDoS2019*:* Sharafaldin *et al.* [264], alongside their DDoS attack taxonomy and new detection and classification methods, released a 15-hour dataset [131], which was intended to solve the issues of existing datasets covering DDoS attacks in classic IP networks. The testbed used for traffic generation follows the same network layout as the previous two datasets. The victim network is composed of one server, one firewall, two switches, and four PCs. The attack network is not detailed. To generate benign traffic, the authors relied on their B-profiles [260] to represent the interactions of 25 users with Internet services. For attack traffic, they implemented 12 different DDoS attack variants, based on their taxonomy, including SYN and UDP flood, or NTP and DNS reflection. Attack traffic was labelled, and features were extracted with CICFlowMeter [262].

CIRA-CIC-DoHBrw-2020*:* Designed by MontazeriShatoori *et al.* [265], this dataset [132] focuses on securing DNS traffic, and includes **D**NS-**o**ver-**H**TTPS (DoH) traffic, which is an IETF standard to overcome DNS vulnerabilities by using encryption [266]. The dataset is composed of HTTPS traffic flows, which have been classified with a two-level label. The first level indicates if the flow corresponds to DoH traffic or not, and the second level indicates if the flow is benign or malicious. Benign HTTPS traffic has been recorded while visiting popular websites, with the browsers configured to use DoH instead of standard DNS. In parallel, malicious DoH traffic was generated. A set of ten servers was instrumented to record the dataset. The authors then used the DoHMeter tool, which is part of their DoHlyzer suite [267], to extract 28 statistical features from the recorded traffic flows, including the number and rate of packets, and statistics about the packet length and request/response times.

LITNET-2020*:* Damasevicius *et al.* [268] proposed a new public, annotated real-world network dataset for N-IDS validation [133]. It contains both benign and attack network flows and includes 85 network flow features and 12 types of attack. The dataset was collected during 10 months in the Lithuanian Research and Education Network (LITNET) that connects multiple universities in different cities in the country. High-performance next-generation firewalls analyzed the data passing through the network and sent NetFlow records to collector machines that were responsible for receiving, filtering, storing the data. To generate attack traffic, 12 types of attack were implemented, including ICMP, UDP, TCP SYN flood and HTTP flood, as well as Smurf and different types of worms. Flows were labelled as benign or with the corresponding attack. A total of 85 features were extracted from the network flows, by following the description scheme suggested by Moustafa and Slay [269].

AWID3 *(2021):* Chatzoglou *et al.* [270] proposed a new version [134] of the *AWID2* dataset [63], instrumenting WPA2-certified devices. Their configuration assumes an enterprise 802.11 environment and requires Protected Management Frames (PMF) introduced with the IEEE 802.11w amendment [271].

The used testbed represented an enterprise Wi-Fi network comprising 17 devices in total: ten general purpose physical clients (computers, smartphones and tablet), six servers, and one attacker device. The clients executed benign Internet access activities, selected randomly among a list including watching YouTube videos, downloading files or send and receive emails. Network traffic was captured by one of the servers, for a total duration of around 140 minutes. 21 different attacks were implemented, representing different attack families including deauthentication, disassociation, rogue AP (false illegitimate access point), etc. The dataset is available in both the PCAP and CSV formats. The former provides the raw recorded packets, split into 13 files containing around 10 minutes of traffic to keep them easily analyzable, whereas the latter follows the format of AWID2, with the number of attributes increased to 254.

NF-UQ-NIDS *(2021):* Sarhan *et al.* [272], [273] argued that, even if several datasets are publicly available for the validation of IDSs, they are often impossible to compare to each other, because they do not use the same set of features. Making a step towards the unification of dataset features, they analyzed four aforementioned datasets, namely UNSW-NB15 [255], Bot-IoT [155], TON_IoT [167], and CSE-CIC-IDS2018 [130], and converted them into NetFlow [254] format, producing four new datasets available publicly [135]. Those new datasets share the same set of NetFlow features and are available in two versions with eight basic features or 43 extended features, respectively. Additionally, they produced a new comprehensive dataset which merges the four previous datasets together.

H23Q *(2022):* After the aforementioned AWID3 dataset, the University of the Aegean developed a new network traffic dataset, H23Q [136], proposed by Chatzoglou *et al.* [274], which includes attacks against services using the three modern protocols HTTP/2 [275], HTTP/3 [276], and QUIC [277]. As noted by Lars Eggert [278], the QUIC protocol might become primordial for IoT networks, thanks to its inherent speed and security compared to TCP.

The authors instrumented 13 physical clients, one being malicious, split in three different local networks, and six servers

in an Azure cloud, each exposing a service implementing the studied protocols. The server part of the testbed, as well as the QUIC-oriented attacks implemented for malicious traffic collection, were described by the same authors in their preliminary analysis of QUIC vulnerabilities [279]. The dataset represents benign and malicious traffic, for a total recording duration of 107 minutes, and is available in CSV and PCAP format. The CSV files contain 200 descriptive features, one of them being the label indicating the attack class, whilst the PCAP files contain full packets.

IX. Challenges and Future Directions

In this second to last section, we highlight the challenges related to the datasets we covered, and propose further research directions. Namely, we point out the lack of datasets including IoT-specific protocols, the relative small scale of the presented datasets, and the challenges associated with data privacy and the generation of realistic traffic.

A. IoT-specific protocols and attack traffic

In Figure 6, our survey has highlighted the prominence of IP-based datasets and the lack of datasets involving IoT-specific protocols. We have found five datasets with Bluetooth traffic, seven with Zigbee, and five with LoRa. On the application layer, the CoAP protocol seems interesting in theory, but appears in only two datasets, and with a limited variety of attacks targeting it. This could indicate that the protocol is not widely used in practice, or has not drawn enough attention from the industry or researchers. To summarize, for many protocols at all networking layers, there is only little public data available that would allow researchers to see how those protocols are actually used in the wild and that would allow to perform experiments that can be reproduced by others.

In addition, some of those datasets have limitations that might make them less useful for security researchers. Taking the example of LP-WANs, i.e., LoRa and Sigfox, most of the pertinent datasets were created for device fingerprinting and localization [198], [205], [209], [280], [281]. As such, those datasets only contain physical layer data, as well as potentially the geolocation of the devices, but they do not contain network data such as the payload or port used for data transmission. Moreover, to the best of our knowledge, no public dataset exists that contains attack traffic targeting such networks.

We therefore hope that future researchers working in this area will consider generating and releasing datasets that contain such protocols. The incentive should be given to the research community to produce more network datasets specifically for IoT, such as LP-WANs and CoAP, and consider including attack traffic, preferably of diverse types. Publicly releasing such datasets can prove very helpful for research on the network security of IoT devices.

In some cases, researchers do not want to release their collected data publicly. This can be caused by privacy and confidentiality concerns or by constraints imposed by the funding source (industry collaboration, etc.). As a partial solution to this problem, Cermak *et al.* [282] introduced a methodology to be able to share datasets while keeping private information hidden. Their methodology proposes the usage of small portions of data, or "annotated units" as they call them, which represent a subset of a full, real-world traffic trace. Those units can be anonymized more easily, and thus be shared with the community. For example, such a traffic unit can be the subset of the full network transmission that corresponds to one specific type of attack. Researchers willing to use those can then enrich the small units with potentially private self-collected traffic to form a dataset of a relevant size.

B. Data privacy and confidentiality

IoT devices have become popular in consumer and industrial applications, and given the nature and amount of data exchanged in IoT environments, data privacy and confidentiality have become important concerns.

Data privacy refers to the property that no information related to an individual should be disclosed without their explicit consent. An example of why this property must be enforced is in healthcare applications. IoT devices collect data such as the user's heartbeat rate, their blood pressure, and their physical activity. With such data, sensitive information about the user can be retrieved, e.g., the presence of a disease, and be used in a harmful way [283]. *Data confidentiality* ensures that unauthorized people cannot access the data. The difference between data privacy and confidentiality is that the former protects individuals while the latter protects the data [284].

Ensuring privacy and confidentiality is important in IoT environments, however, there exist challenges that hinder this. For example, a fundamental aspect of data privacy is that users understand what data is collected about them and where it is processed. This is rarely the case in current consumer-oriented IoT applications for various reasons. One of them is the simple fact that users do not always have the skills and knowledge to understand all the technical processes in IoT environments, and the manufacturers do not always make the effort to provide simple and clear explanations [285]. As far as data confidentiality is concerned, the main challenge is the limited resources of current IoT devices. In traditional networks, confidentiality is ensured by cryptography, but implementing complex cryptographic mechanisms is difficult on IoT devices due to their limited computing power. In addition, the heterogeneity of devices and communication protocols makes it difficult to design solutions that are compatible with any type of IoT environment.

The goal of the datasets presented in this survey is mostly to evaluate anomaly detection techniques, as shown in Table V. Indeed, only two datasets were created with a privacy-related research purpose. Privacy is a relatively new topic compared to data confidentiality and computer security in general. Clearly, there is a lack of interesting datasets for research on privacy.

However, researchers are often not willing or allowed to publicly release their datasets because of privacy and confidentiality concerns. When releasing a dataset, complex measures must be taken to not violate privacy. If the dataset is empirical and therefore the result of the involvement of individuals, it is necessary to inform them which data are collected and what they are used for, and the necessary steps must be taken to

get their explicit consent. Data confidentiality can be also problematic, especially in datasets that contain traffic payload, as necessary for the evaluation of IDSs using Deep Packet Inspection. If such a dataset is released, the authors must be careful about the information that can be retrieved in the dataset.

To avoid these challenges, two approaches exist: one is to rely on test beds or simulations without human involvement, for example by using traffic generators (see Section IX-D). The other is the usage of anonymization techniques. However, when releasing a public dataset, a trade-off must be found between its usefulness and the level of anonymization applied to it [286]–[288].

C. Scale and complexity

As shown in Figure 7, most IoT datasets are relatively small in terms of number of monitored devices. This is understandable for applications such as Smart Homes that only involve small networks. However, it's less understandable for other application areas where the scalability of IoT to networks of thousands of devices has been demonstrated by the industry. We therefore appeal to the research community to collect and publish more datasets from such large-scale applications.

Another open issue is the complexity of existing datasets. Most of them focus on a specific communication technology. However, modern IoT applications are often deployed in heterogeneous systems, even on a small scale. For the latter, a typical example is a Smart Home containing devices using various technologies, from short-range personal-area networks to long-range and cellular networks. In such a system, it would be wrong to see the individual networks as isolate entities since the devices are interdependent. Future datasets have to reflect this complexity.

D. Realistic traffic generation

Recording a dataset from a testbed or operative network requires considerable effort, not only to configure and instrument the infrastructure but also to guarantee that the monitored data is representative and pertinent, i.e., it contains realistic network traffic caused by benign as well as malicious activities. To alleviate this task, one may draw support from *traffic generators*, which are tools able to synthesize benign and malicious traffic.

In the past years, a large number of tools for the generation of general IP traffic have been published, including the open-source D-ITG [289], [290], Harpoon [291], [292], fudp [293], and TRex [294]. However, those well-known tools are more inclined towards classical Internet traffic, and might not excel at simulating IoT traffic. Recently, researchers started developing traffic generators oriented towards IoT networks and protocols, six of which will be presented below.

IoT-Flock [178], [295] is an open-source framework able to generate CoAP and MQTT traffic. It can produce benign and malicious traffic. Use cases can be defined in XML files, or by using the graphical or command line interfaces. The authors evaluated their tool by simulating their IoT sensor testbed, as well as malicious traffic. Both types of traffic were discriminated very accurately by ML classifiers.

The hybrid IoT traffic generator by Hsu *et al.* [296] is based on an analysis of human-to-human activities. It can generate Machine Type Communication (MTC) traffic with geographic information attached for cell and network loading level assessment. To showcase its usability, the authors applied their generator on two related examples: mobile network performance assessment and improvement.

Hui *et al.* [297] used a condition mechanism to incorporate semantic and network structure knowledge and device category in IoT traffic generation with a generative adversarial network. The generator consists of three sub-generators to satisfy three of the requirements for IoT traffic generation mentioned in their paper. Their evaluation shows that their traffic generator comply with state-of-the-art data fidelity baselines.

Nguyen-An *et al.* [298] extended Scapy to generate IoT traffic based on the functionalities and characteristics of different device in Smart Home environments. All the generated packets are gathered into a single trace file, with support for different formats such as PCAP and CSV. For the evaluation, the authors simulated Smart Home traffic, which resulted in a lower entropy in network parameters for their simulated traffic than for real traffic.

The LoRa traffic generator by Lavric *et al.* [299] generates packets that are compatible with the LoRaWAN specifications. The generator is able to generate large volumes of traffic at high speed. The user has access to parameters such as spreading factor, packet payload, and coding rate. The traffic generator can be integrated into other applications. Thus, gateway manufacturers can use it to assess the real throughput of their products.

The LoRa cell traffic generator by Gucciardo *et al.* [300] is able to emulate the behavior of thousands of low-rate sensor nodes deployed in the same cell by using a single Software Defined Radio (SDR) platform. Unlike the traditional generators who focus on creating individual packet flows which emulate specific applications and protocols, this generator generates a combined radio signal given by the super-position of the signals transmitted by multiple sensors simultaneously active on the same channel. This work can be of great help for LoRa network designers who want to avoid inferring signals from becoming impossible to decode.

A challenge that has to be addressed by future traffic generators is the fact that the different networking layers in IoT applications are highly intertwined. IoT application traffic over a LoRa network is not just a slower version of the same traffic over a Wi-Fi network. Resource constraints force developers of IoT applications to adapt the network communication to the limitations of the used devices and network technologies. Therefore, the generation of realistic IoT traffic has to consider the targeted environment in its entirety, from the application down to the type of devices.

In parallel to traffic generation tools, one can retrieve relevant traffic data by instrumenting honeypot networks, which are networks intentionally left vulnerable and open to the internet, with the specific intent of collecting potentially malicious traffic. Franco *et al.* [301], analogously to our

manuscript, surveyed existing honeypot networks for (Industrial) IoT networks, which could be used to gather relevant traffic for IoT network security research.

X. Conclusion

Datasets are critical in many research areas and the existence of a sufficient number of high-quality datasets can not only be the base of new scientific insights but also ensure the accuracy and reliability of the results. In particular, *public* datasets allow to compare solutions developed by different teams and to publish results that can be reproduced and verified by others.

In this paper, we have presented a comprehensive survey of available datasets suitable for the research on IoT network security. Among the 74 surveyed datasets, half of them contain IoT network traffic and, in some but not all cases, also network attacks. We have also included several well-known datasets that do not contain IoT traffic but that have been used in previous research on IoT, sometimes due to the lack of IoT specific datasets. We have classified the datasets based on ten characteristics, such as the observed network protocols, the level of detail of the dataset records, as well as the date, duration, number of devices and methodology of the data collection, in order to help researchers who are looking for a dataset suitable to their particular needs. Finally, we have also briefly described the existing tools to generate IoT traffic.

Our conclusion is that there is still a scarcity or even lack of publicly available datasets covering IoT-specific protocols, such as LoRa or CoAP. Other emerging protocols, such as QUIC, which have initially be designed for other use cases than IoT might present characteristics relevant for IoT communication and should therefore attract research interest as well. In addition, many of the existing datasets do not contain attack traffic. Although benign traffic datasets are still useful for security researchers, it requires additional effort and careful planning to augment them with attack traffic. We therefore encourage future authors of new IoT datasets to also include various attack scenarios in their experiments.

References

[1] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in Internet-of-Things," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1250–1258, 2017.

[2] V. Hassija, V. Chamola, V. Saxena, D. Jain, P. Goyal, and B. Sikdar, "A survey on IoT security: application areas, security threats, and solution architectures," *IEEE Access*, vol. 7, pp. 82 721–82 743, 2019.

[3] J. Mohanty, S. Mishra, S. Patra, B. Pati, and C. R. Panigrahi, "IoT security, challenges, and solutions: a review," *Progress in Advanced Computing and Intelligent Engineering*, pp. 493–504, 2021.

[4] Z. Wang, D. Liu, Y. Sun, X. Pang, P. Sun, F. Lin, J. C. S. Lui, and K. Ren, "A survey on iot-enabled home automation systems: Attacks and defenses," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2292–2328, 2022.

[5] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the Mirai Botnet," in *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, Aug. 2017, pp. 1093–1110. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis

[6] Y. Harbi, Z. Aliouat, A. Refoufi, and S. Harous, "Recent Security Trends in Internet of Things: A Comprehensive Survey," *IEEE Access*, 2021.

[7] J. Habibi, D. Midi, A. Mudgerikar, and E. Bertino, "Heimdall: Mitigating the Internet of Insecure Things," *IEEE Internet of Things Journal*, vol. 4, no. 4, pp. 968–978, 2017.

[8] M. Eskandari, Z. H. Janjua, M. Vecchio, and F. Antonelli, "Passban IDS: An Intelligent Anomaly-Based Intrusion Detection System for IoT Edge Devices," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6882–6897, Aug. 2020, conference Name: IEEE Internet of Things Journal.

[9] L. Metongnon, R. Sadre, and E. C. Ezin, "Distributed Middlebox Architecture for IoT Protection," in *2019 15th International Conference on Network and Service Management (CNSM)*, Oct. 2019, pp. 1–7, iSSN: 2165-963X.

[10] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT—network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, 2018.

[11] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. M. A. Hashem, "Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches," *Internet of Things*, vol. 7, p. 100059, Sep. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660519300241

[12] C. Grajeda, F. Breitinger, and I. Baggili, "Availability of datasets for digital forensics–and what is missing," *Digital Investigation*, vol. 22, pp. S94–S105, 2017.

[13] P. Penrose, R. Macfarlane, and W. J. Buchanan, "Approaches to the classification of high entropy file fragments," *Digital Investigation*, vol. 10, no. 4, pp. 372–384, 2013.

[14] G. K. Ndonda and R. Sadre, "Network trace generation for flow-based IDS evaluation in control and automation systems," *International Journal of Critical Infrastructure Protection*, vol. 31, p. 100385, 2020.

[15] C. Archer. (2021, 29 Mar.) "What's the difference between SCADA and IoT?". WellAware. Accessed: July 17, 2022. [Online]. Available: https://blog.wellaware.us/blog/difference-between-scada-and-iot-scada-vs-iot

[16] H. M. Song, J. Woo, and H. K. Kim, "In-vehicle network intrusion detection using deep convolutional neural network," *Vehicular Communications*, vol. 21, p. 100198, 2020.

[17] H. Zimmermann, "OSI reference model-the ISO model of architecture for open systems interconnection," *IEEE Transactions on communications*, vol. 28, no. 4, pp. 425–432, 1980.

[18] K. Sairam, N. Gunasekaran, and S. R. Redd, "Bluetooth in wireless communication," *IEEE Communications Magazine*, vol. 40, no. 6, pp. 90–96, 2002.

[19] A. F. Harris III, V. Khanna, G. Tuncay, R. Want, and R. Kravets, "Bluetooth low energy in dense IoT environments," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 30–36, 2016.

[20] J. A. Gutiérrez, "On the use of IEEE Std. 802.15. 4 to enable wireless sensor networks in building automation," *International Journal of Wireless Information Networks*, vol. 14, no. 4, pp. 295–301, 2007.

[21] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low Power Wide Area Networks: An Overview," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 855–873, 2017.

[22] C. Milarokostas, D. Tsolkas, N. Passas, and L. Merakos, "A Comprehensive Study on LPWANs With a Focus on the Potential of LoRa/LoRaWAN Systems," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 825–867, 2023.

[23] J. Petajajarvi, K. Mikhaylov, A. Roivainen, T. Hanninen, and M. Pettissalo, "On the coverage of LPWANs: range evaluation and channel attenuation model for LoRa technology," in *2015 14th international conference on its telecommunications (itst)*. IEEE, 2015, pp. 55–59.

[24] J. Haxhibeqiri, E. De Poorter, I. Moerman, and J. Hoebeke, "A survey of LoRaWAN for IoT: From technology to application," *Sensors*, vol. 18, no. 11, p. 3995, 2018.

[25] A. Lavric, A. I. Petrariu, and V. Popa, "SigFox Communication Protocol: The New Era of IoT?" in *2019 International Conference on Sensing and Instrumentation in IoT Era (ISSI)*. IEEE, 2019, pp. 1–4.

[26] K. Mekki, E. Bajic, F. Chaxel, and F. Meyer, "Overview of cellular LPWAN technologies for IoT deployment: Sigfox, LoRaWAN, and NB-IoT," in *2018 ieee international conference on pervasive computing and communications workshops (percom workshops)*. IEEE, 2018, pp. 197–202.

[27] M. B. Yassein, W. Mardini, and A. Khalil, "Smart homes automation using Z-wave protocol," in *2016 International Conference on Engineering & MIS (ICEMIS)*. IEEE, 2016, pp. 1–6.

[28] S. C. Ergen, "ZigBee/IEEE 802.15. 4 Summary," *UC Berkeley, September*, vol. 10, no. 17, p. 11, 2004.

[29] G. Mulligan, "The 6LoWPAN architecture," in *Proceedings of the 4th workshop on Embedded networked sensors*, 2007, pp. 78–82.

[30] R. Alexander, A. Brandt, J. Vasseur, J. Hui, K. Pister, P. Thubert, P. Levis, R. Struik, R. Kelsey, and T. Winter, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," Internet Requests for Comments, RFC Editor, RFC 6550, Mar. 2012. [Online]. Available: https://www.rfc-editor.org/info/rfc6550

[31] D. Soni and A. Makwana, "A survey on MQTT: a protocol of Internet of Things (IoT)," in *International Conference On Telecommunication, Power Analysis And Computing Techniques (ICTPACT-2017)*, vol. 20, 2017, pp. 173–177.

[32] P. T. Eugster, P. A. Felber, R. Guerraoui, and A.-M. Kermarrec, "The many faces of publish/subscribe," *ACM computing surveys (CSUR)*, vol. 35, no. 2, pp. 114–131, 2003.

[33] Z. Shelby, K. Hartke, and C. Bormann, "The Constrained Application Protocol (CoAP)," Internet Requests for Comments, RFC Editor, RFC 7252, Jun. 2014. [Online]. Available: https://www.rfc-editor.org/rfc/rfc7252

[34] "What is Modbus and How does it work?". Schneider Electric. Accessed: Feb. 13, 2023. [Online]. Available: https://www.se.com/us/en/faqs/FA168406

[35] "Modbus Messaging on TCP/IP Implementation Guide V1.0b". Modbus Organization. Accessed: Feb. 13, 2023. [Online]. Available: https://www.modbus.org/docs/Modbus_Messaging_Implementation_Guide_V1_0b.pdf

[36] M. Tabaa, B. Chouri, S. Saadaoui, and K. Alami, "Industrial Communication based on Modbus and Node-RED," *Procedia Computer Science*, vol. 130, pp. 583–588, 2018.

[37] G. Corotinschi and V. G. Găitan, "Enabling IoT connectivity for Modbus networks by using IoT edge gateways," in *2018 International Conference on Development and Application Systems (DAS)*. IEEE, 2018, pp. 175–179.

[38] S.-I. Toc and A. Korodi, "Modbus-OPC UA Wrapper Using Node-RED and IoT-2040 with Application in the Water Industry," in *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 2018, pp. 000 099–000 104.

[39] "Modbus Application Protocol Specification V1.1b3". Modbus Organization. Accessed: Feb. 13, 2023. [Online]. Available: https://www.modbus.org/docs/Modbus_Application_Protocol_V1_1b3.pdf

[40] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[41] X. Jing, Z. Yan, and W. Pedrycz, "Security data collection and data analytics in the internet: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 586–618, 2019.

[42] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network intrusion detection for iot security based on learning techniques," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2671–2701, 2019.

[43] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[44] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Computers & Security*, vol. 86, pp. 147–167, 2019.

[45] P. M. S. Sánchez, J. M. J. Valero, A. H. Celdrán, G. Bovet, M. G. Pérez, and G. M. Pérez, "A survey on device behavior fingerprinting: Data sources, techniques, application scenarios, and datasets," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1048–1077, 2021.

[46] M. Conti, D. Donadel, and F. Turrin, "A Survey on Industrial Control System Testbeds and Datasets for Security Research," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2248–2294, 2021.

[47] R. Rawassizadeh and D. Kotz, "Datasets for mobile, wearable and IoT research," *GetMobile: Mobile Computing and Communications*, vol. 20, no. 4, pp. 5–7, 2017.

[48] "IMC Conference". ACM Sigcomm. Accessed: Feb. 10, 2022. [Online]. Available: https://www.sigcomm.org/events/imc-conference

[49] "Passive and Active Measurement Conference 2022". Accessed: Feb. 10, 2022. [Online]. Available: https://pam2022.nl

[50] "TMA Conference". Accessed: Feb. 10, 2022. [Online]. Available: https://tma.ifip.org

[51] K. C. Kang, S. G. Cohen, J. A. Hess, W. E. Novak, and A. S. Peterson, "Feature-oriented domain analysis (FODA) feasibility study," Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst, Tech. Rep., 1990.

[52] P.-Y. Schobbens, P. Heymans, and J.-C. Trigaux, "Feature diagrams: A survey and a formal semantics," in *14th IEEE International Requirements Engineering Conference (RE'06)*, 2006, pp. 139–148.

[53] "Riverbed Modeler | Discrete Event Simulator for Network Simulation". Riverbed. Accessed: Jan. 11, 2022. [Online]. Available: https://www.riverbed.com/products/npm/riverbed-modeler.html

[54] "ns-2 - User Information". nsnam. Accessed: Feb. 3, 2022. [Online]. Available: http://nsnam.sourceforge.net/wiki/index.php/User_Information

[55] "ns-3 | A discrete-event network simulator for internet systems". Accessed: Jan. 11, 2022. [Online]. Available: https://www.nsnam.org/

[56] F. Osterlind, A. Dunkels, J. Eriksson, N. Finne, and T. Voigt, "Cross-Level Sensor Network Simulation with COOJA," in *Proceedings. 2006 31st IEEE conference on local computer networks*. IEEE, 2006, pp. 641–648.

[57] "Data | KDD Cup 1999: Computer network intrusion detection". KDD. Accessed: Jan. 12, 2022. [Online]. Available: https://www.kdd.org/kdd-cup/view/kdd-cup-1999/Data

[58] L. Spitzner, *Honeypots: tracking hackers*. Addison-Wesley Reading, 2003, vol. 1.

[59] R. G. Bace, P. Mell *et al.*, "Intrusion detection systems," 2001.

[60] M. G. Gouda and A. X. Liu, "Structured firewall design," *Computer networks*, vol. 51, no. 4, pp. 1106–1120, 2007.

[61] S. J. Saidi, A. M. Mandalari, R. Kolcun, H. Haddadi, D. J. Dubois, D. R. Choffnes, G. Smaragdakis, and A. Feldmann, "A Haystack Full of Needles: Scalable Detection of IoT Devices in the Wild," in *IMC '20: ACM Internet Measurement Conference, Virtual Event, USA, October 27-29, 2020*. ACM, 2020, pp. 87–100. [Online]. Available: https://doi.org/10.1145/3419394.3423650

[62] H. Nissenbaum, "Privacy in context," in *Privacy in Context*. Stanford University Press, 2009.

[63] "AWID2". University of the Aegean. Accessed: Jan. 25, 2022. [Online]. Available: https://icsdweb.aegean.gr/awid/awid2

[64] A. Sivanathan, D. Sherratt, H. H. Gharakheili, V. Sivaraman, and A. Vishwanath, "Low-cost flow-based security solutions for smart-home IoT devices," in *2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*. IEEE, 2016, pp. 1–6.

[65] "Urban_IoT_DDoS_Data". ANRG USC. Accessed: Feb. 11, 2022. [Online]. Available: https://github.com/ANRGUSC/Urban_IoT_DDoS_Data

[66] S. Marchal. (2017, 3 Apr.) "IoT devices captures". Aalto University. Accessed: Feb. 8, 2022. [Online]. Available: https://research.aalto.fi/en/datasets/iot-devices-captures

[67] V. H. Bezerra, V. G. T. da Costa, R. A. Martins, S. B. Junior, R. S. Miani, and B. B. Zarpelão, "Providing IoT host-based datasets for intrusion detection research," in *Proceedings of the 18th Brazilian Symposium on Information and Computational Systems Security*. Porto Alegre, RS, Brasil: SBC, 2018, pp. 15–28. [Online]. Available: https://sol.sbc.org.br/index.php/sbseg/article/view/4240

[68] "detection_of_IoT_botnet_attacks_N_BaIoT Data Set". UCI Machine Learning Repository. Accessed: Jan 20, 2022. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/detection_of_IoT_botnet_attacks_N_BaIoT

[69] R. Doshi, N. Apthorpe, and N. Feamster, "Machine learning DDoS detection for consumer Internet of Things devices," in *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2018, pp. 29–35.

[70] F.-X. Aubet and M.-O. Pahl. (2018) "DS2OS traffic traces". Accessed: Feb. 28, 2022. [Online]. Available: https://www.kaggle.com/francoisxa/ds2ostraffictraces

[71] "IoT Traffic Traces". UNSW Sydney. Accessed: Jan. 12, 2022. [Online]. Available: https://iotanalytics.unsw.edu.au/iottraces.html

[72] "IoT Benign and Attack Traces". UNSW Sydney. Accessed: Jan. 1, 2022. [Online]. Available: https://iotanalytics.unsw.edu.au/attack-data

[73] "MUD Profiles". UNSW Sydney. Accessed: Jan. 12, 2022. [Online]. Available: https://iotanalytics.unsw.edu.au/mudprofiles

[74] "IoT Information Exposure (IMC '19)". Mon(IoT)r Research Group. Accessed: Jan. 20, 2022. [Online]. Available: https://moniotrlab.ccis.neu.edu/imc19

[75] "The Bot-IoT Dataset". UNSW Sydney. Accessed: Jan 19, 2022. [Online]. Available: https://research.unsw.edu.au/projects/bot-iot-dataset

[76] H. Kang, D. H. Ahn, G. M. Lee, J. D. Yoo, K. H. Park, and H. K. Kim. (2019) "IoT network intrusion dataset". IEEE Dataport. DOI: http://dx.doi.org/10.21227/q70p-q449.

[77] "IoT Detection in the Wild (IMC '20)". Mon(IoT)r Research Group. Accessed: Feb. 2, 2022. [Online]. Available: https://moniotrlab.ccis.neu.edu/imc20

[78] S. Garcia, A. Parmisano, and M. J. Erquiaga. (2020, Jan.) "IoT-23: A labeled dataset with malicious and benign IoT network traffic". Zenodo. DOI: http://dx.doi.org/10.5281/zenodo.4743746.

[79] L. Vigoya, D. Fernandez, V. Carneiro, and F. Cacheda. "DAD | DATASET for Anomaly Detection in IoT Networks". Accessed: Jan. 24, 2022. [Online]. Available: https://github.com/dad-repository/dad

[80] A. Guerra-Manzanares, J. Medina-Galindo, H. Bahsi, and S. Nõmm. "MedBIoT - Medium-sized IoT Botnet Data Set". Accessed: Jan. 24, 2022. [Online]. Available: https://cs.taltech.ee/research/data/medbiot

[81] Sarica, Alper Kaan. "SDN-Dataset". Accessed: Feb 17, 2022. [Online]. Available: https://github.com/AlperKaan35/SDN-Dataset

[82] "IoT IPFIX Records". UNSW Sydney. Accessed: Jan. 27, 2022. [Online]. Available: https://iotanalytics.unsw.edu.au/iotipfixrecords

[83] "The TON_IoT Datasets". UNSW Sydney. Accessed: Jan 20, 2022. [Online]. Available: https://research.unsw.edu.au/projects/toniot-datasets

[84] Z. Liu, N. Thapa, A. Shaver, K. Roy, M. Siddula, X. Yuan, and A. Yu, "Using Embedded Feature Selection and CNN for Classification on CCD-INID-V1—A New IoT Dataset," *Sensors*, vol. 21, no. 14, p. 4834, 2021.

[85] "IoTLS (IMC '21)". Mon(IoT)r Research Group. Accessed: Feb. 2, 2022. [Online]. Available: https://moniotrlab.ccis.neu.edu/iotls

[86] M. Ahmed, S. Byreddy, A. Nutakki, L. F. Sikos, and P. Haskell-Dowland. "ECU-IoHT". Accessed: Jan. 24, 2022. [Online]. Available: https://ro.ecu.edu.au/datasets/48

[87] I. Vaccari, G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso. "MQTTset - A new dataset for machine learning techniques on MQTT". Accessed: June 22, 2022. [Online]. Available: https://www.kaggle.com/datasets/cnrieiit/mqttset

[88] "Data and code for TMA 2021 paper "Revisiting IoT Device Identification"". Accessed: Feb. 28, 2022. [Online]. Available: https://github.com/DADABox/revisiting-iot-device-identification

[89] H. Gandhi, M. Mehra, and V. Ribeiro, "BOND: Efficient and Frugal DL Model Co-Design for Botnet Detection on IoT Gateways," in *The First International Conference on AI-ML-Systems*, ser. AIMLSystems 2021. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3486001.3486237

[90] M. Ahmed, D. Cox, B. Simpson, and A. Aloufi. "ECU-IoFT: A Dataset for Analysing Cyber-Attacks on Internet of Flying Things". Accessed: June 22, 2022. [Online]. Available: https://github.com/iMohi/ECU-IoFT

[91] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami. "Labelled Wireless Sensor Network Data Repository (LWSNDR)". Accessed: Jan. 27, 2022. [Online]. Available: https://home.uncg.edu/cmp/downloads/lwsndr.html

[92] S. Fu and Y. Zhang. (2015) "the due/packet-delivery dataset (v. 2015-04-01)". Accessed: Mar. 20, 2023. [Online]. Available: https://ieee-dataport.org/open-access/crawdad-duepacket-delivery

[93] "WSN-DS". Security Engineering Lab. Accessed: Feb. 3, 2022. [Online]. Available: https://sel.psu.edu.sa/Research/datasets/2016_WSN-DS.php

[94] S. Dowling, M. Schukat, and H. Melvin, "A ZigBee honeypot to assess IoT cyberattack behaviour," in *2017 28th Irish signals and systems conference (ISSC)*. IEEE, 2017, pp. 1–6.

[95] B. Purnama, Sharipuddin, Kurniabudi, R. Budiarto, D. Stiawan, and D. Hanapi, "Monitoring Connectivity of Internet of Things Device on Zigbee Protocol," in *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2018, pp. 351–356.

[96] "BLEBeacon Dataset". Accessed: Feb. 3, 2022. [Online]. Available: https://github.com/dimisik/BLEBeacon-Dataset

[97] M. Aernouts, R. Berkvens, K. Van Vlaenderen, and M. Weyn. (2019, Jul.) "Sigfox and LoRaWAN Datasets for Fingerprint Localization in Large Urban and Rural Areas". Zenodo. DOI: http://dx.doi.org/10.5281/zenodo.3904158.

[98] A. Verma and V. Ranga, "Evaluation of network intrusion detection systems for RPL based 6LoWPAN networks in IoT," *Wireless Personal Communications*, vol. 108, no. 3, pp. 1571–1594, 2019.

[99] Y. Al-Hadhrami and F. K. Hussain, "Real time dataset generation framework for intrusion detection systems in IoT," *Future Generation Computer Systems*, vol. 108, pp. 414–423, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X19322678

[100] L. Bhatia, M. Breza, R. Marfievici, and J. A. McCann. (2020, Sep.) "LoED: The LoRaWAN at the Edge Dataset". Zenodo. DOI: http://dx.doi.org/10.5281/zenodo.4121430.

[101] "LoRa Radio Data". InterDigital. Accessed: Jan. 26, 2022. [Online]. Available: https://www.interdigital.com/data_sets/lora-radio-data

[102] A. Elmaghbub and B. Hamdaoui. (2021, Jul.) "Release Note: Comprehensive LoRa RF Datasets for Device Fingerprinting Using Deep Learning". Accessed: Feb. 16, 2022. [Online]. Available: http://research.engr.oregonstate.edu/hamdaoui/sites/research.engr.oregonstate.edu.hamdaoui/files/release_note_2021.pdf

[103] D.-G. Akestoridis, M. Harishankar, M. Weber, and P. Tague. (2020, May) "CRAWDAD dataset cmu/zigbee-smarthome (v. 2020-05-26)". Downloaded from https://ieee-dataport.org/open-access/crawdad-cmuzigbee-smarthome.

[104] F. Sadikin, T. van Deursen, and S. Kumar, "A ZigBee Intrusion Detection System for IoT using Secure and Efficient Data Collection," *Internet of Things*, vol. 12, p. 100306, 2020.

[105] A. Collen, N. A. Nijdam, J. Augusto-Gonzalez, S. K. Katsikas, K. M. Giannoutakis, G. Spathoulas, E. Gelenbe, K. Votis, D. Tzovaras, N. Ghavami *et al.* "GHOST-IoT-data-set | Smart-home network traffic IoT dataset". Accessed: Jan. 24, 2022. [Online]. Available: https://github.com/gspathoulas/ghost-iot-dataset

[106] Y.-W. Chen, J.-P. Sheu, Y.-C. Kuo, and N. Van Cuong, "Design and implementation of IoT DDoS attacks detection system based on machine learning," in *2020 European Conference on Networks and Communications (EuCNC)*. IEEE, 2020, pp. 122–127. [Online]. Available: http://smartcity.cs.nthu.edu.tw/dataset/download.php

[107] L. Barman, A. Dumur, A. Pyrgelis, and J.-P. Hubaux. (2021) "Repository for the paper "Every Byte Matters: Traffic Analysis of Bluetooth Wearable Devices"". Accessed: Jan. 25, 2022. [Online]. Available: https://github.com/ldsec/every_byte_matters

[108] A. Lacava, E. Giacomini, F. D'Alterio, and F. Cuomo, "Intrusion Detection System for Bluetooth Mesh Networks: Data Gathering and Experimental Evaluations," in *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 2021, pp. 661–666.

[109] D.-G. Akestoridis and P. Tague. (2021, Oct.) "CRAWDAD dataset cmu/zigbee-eda (v. 2021-10-22)". CRAWDAD. DOI: http://dx.doi.org/10.15783/t8mt-a674.

[110] Kamaldeep, M. Malik, M. Dutta, and J. Granjal, "IoT-Sentry: A Cross-Layer-Based Intrusion Detection System in Standardized Internet of Things," *IEEE Sensors Journal*, vol. 21, no. 24, pp. 28 066–28 076, 2021.

[111] M. Zolanvari, Z. Yang, K. Khan, R. Jain, and N. Meskin. (2021, Oct.) "WUSTL-IIOT-2021 Dataset for IIoT Cybersecurity Research". Accessed: Feb. 13, 2023. [Online]. Available: https://www.cse.wustl.edu/~jain/iiot2/index.html

[112] M. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, "X-IIoTID: A Connectivity- and Device-agnostic Intrusion Dataset for Industrial Internet of Things," 2021. [Online]. Available: https://dx.doi.org/10.21227/mpb6-py55

[113] J. Mathews, P. Chatterjee, and S. Banik. "CoAP-DDoS - DDoS Dataset for CoAP Protocol in IoT". Accessed: Sept. 28, 2022. [Online]. Available: https://www.kaggle.com/datasets/jaredalanmathews/coapddos

[114] M. A. Ferrag. (2022) "Edge-IIoTset Cyber Security Dataset of IoT & IIoT". Accessed: Feb. 13, 2023. [Online]. Available: https://www.kaggle.com/datasets/mohamedamineferrag/edgeiiotset-cyber-security-dataset-of-iot-iiot

[115] D. Kotz, T. Henderson, I. Abyzov, and J. Yeo. (2009, Sep.) "CRAWDAD dataset dartmouth/campus (v. 2009-09-09)". CRAWDAD. DOI: http://dx.doi.org/10.15783/C7F59T.

[116] "The CAIDA "DDoS Attack 2007" Dataset". CAIDA. Accessed: Jan. 1, 2022. [Online]. Available: https://www.caida.org/catalog/datasets/ddos-20070804_dataset

[117] "NSL-KDD dataset". Canadian Institute for Cybersecurity. Accessed: Jan. 27, 2022. [Online]. Available: https://www.unb.ca/cic/datasets/nsl.html

[118] "Cyber Research Center - Data Sets". United States Military Academy West Point. Accessed: Feb. 1, 2022. [Online]. Available: https://www.westpoint.edu/centers-and-research/cyber-research-center/data-sets

[119] (2009) "DARPA_2009_DDoS_attack-20091105". Impact Cyber Trust. Accessed: Feb. 8, 2022. [Online]. Available: https://www.impactcybertrust.org/dataset_view?idDataset=742

[120] "WITS: Waikato Internet Traffic Storage". WAND Network Research Group. Accessed: Jan. 28, 2022. [Online]. Available: https://wand.net.nz/wits

[121] "Traffic Data from Kyoto University's Honeypots". Accessed: Jan. 31, 2022. [Online]. Available: http://www.takakura.com/Kyoto_data

[122] A. Vasudevan, E. Harshini, and S. Selvakumar, "SSENet-2011: A Network Intrusion Detection System dataset and its comparison with KDD CUP 99 dataset," in *2011 Second Asian Himalayas International Conference on Internet (AH-ICI)*, 2011, pp. 1–5.

[123] "Intrusion detection evaluation dataset (ISCXIDS2012)". Canadian Institute for Cybersecurity. Accessed: Feb. 9, 2022. [Online]. Available: https://www.unb.ca/cic/datasets/ids.html

[124] P. Gogoi, M. H. Bhuyan, D. Bhattacharyya, and J. K. Kalita, "Packet and flow based network intrusion dataset," in *International Conference on Contemporary Computing*. Springer, 2012, pp. 322–334.

[125] "ADFA IDS Datasets". UNSW Sydney. Accessed: Feb. 3, 2022. [Online]. Available: https://research.unsw.edu.au/projects/adfa-ids-datasets

[126] "The CTU-13 Dataset. A Labeled Dataset with Botnet, Normal and Background traffic.". Stratosphere Lab. Accessed: Feb. 1 ,2022. [Online]. Available: https://www.stratosphereips.org/datasets-ctu13

[127] "The UNSW-NB15 Dataset". UNSW Sydney. Accessed: Jan. 26, 2022. [Online]. Available: https://research.unsw.edu.au/projects/unsw-nb15-dataset

[128] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, and R. Therón. "UGR'16: A New Dataset for the Evaluation of Cyclostationarity-Based Network IDSs". Accessed: Jan. 24, 2022. [Online]. Available: https://nesg.ugr.es/nesg-ugr16

[129] (2017) "CIC-IDS-2017". Canadian Institute for Cybersecurity. Accessed: Jan. 24, 2022. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html

[130] (2018) "CSE-CIC-IDS2018". Canadian Institute for Cybersecurity. Accessed: Jan. 24, 2022. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2018.html

[131] (2019) "CICDDoS2019". Canadian Institute for Cybersecurity. Accessed: Jan. 24, 2022. [Online]. Available: https://www.unb.ca/cic/datasets/ddos-2019.html

[132] (2020) "CIRA-CIC-DoHBrw-2020". Canadian Institute for Cybersecurity. Accessed: Feb. 4, 2022. [Online]. Available: https://www.unb.ca/cic/datasets/dohbrw-2020.html

[133] R. Damasevicius, A. Venckauskas, S. Grigaliunas, J. Toldinas, N. Morkevicius, T. Aleliunas, and P. Smuikys. "LITNET-2020: an annotated real-world network flows dataset for network intrusion detection". Accessed: Jan. 24, 2022. [Online]. Available: https://dataset.litnet.lt

[134] "AWID3". University of the Aegean. Accessed: Jan. 25, 2022. [Online]. Available: https://icsdweb.aegean.gr/awid/awid3

[135] "Machine Learning-Based NIDS Datasets". The University of Queensland. Accessed: Mar. 30, 2022. [Online]. Available: https://staff.itee.uq.edu.au/marius/NIDS_datasets

[136] "H23Q". University of the Aegean. Accessed: Feb. 7, 2023. [Online]. Available: https://icsdweb.aegean.gr/awid/other-datasets/H23Q

[137] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "Ddos in the iot: Mirai and other botnets," *Computer*, vol. 50, no. 7, pp. 80–84, 2017.

[138] "AWID - Aegean Wi-Fi Intrusion Dataset". University of the Aegean. Accessed: Feb. 14, 2022. [Online]. Available: https://icsdweb.aegean.gr/awid

[139] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, "Intrusion Detection in 802.11 Networks: Empirical Evaluation of Threats and a Public Dataset," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 184–208, 2016.

[140] A. Hekmati, E. Grippo, and B. Krishnamachari, "Large-scale Urban IoT Activity Data for DDoS Attack Emulation," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 2021, pp. 560–564.

[141] M. Miettinen, S. Marchal, I. Hafeez, N. Asokan, A.-R. Sadeghi, and S. Tarkoma, "IoT SENTINEL: Automated Device-Type Identification for Security Enforcement in IoT," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 2177–2184.

[142] S. Herwig, K. Harvey, G. Hughey, R. Roberts, and D. Levin, "Measurement and analysis of hajime, a peer-to-peer iot botnet," in *Network and Distributed Systems Security (NDSS) Symposium*, 2019.

[143] (2016, 03 Nov.) Aidra Botnet. New Jersey Cybersecurity & Communications Integration Cell. Accessed: February 7, 2023. [Online]. Available: https://www.cyber.nj.gov/threat-center/threat-profiles/botnet-variants/aidra-botnet

[144] C. Cimpanu, "There's a 120,000-Strong IoT DDoS Botnet Lurking Around," *Softpedia News*, 30 Aug. 2016, https://news.softpedia.com/news/there-s-a-120-000-strong-iot-ddos-botnet-lurking-around-507773.shtml. Accessed: 21/01/2022.

[145] "Datasets - IoT Security". Sec-MQ. Accessed: Mar. 4, 2022. [Online]. Available: http://www.uel.br/grupo-pesquisa/secmq/dataset-iot-security.html

[146] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: an ensemble of autoencoders for online network intrusion detection," *arXiv preprint arXiv:1802.09089*, 2018.

[147] F.-X. Aubet, "Machine learning-based adaptive anomaly detection in smart spaces," Bachelor's Thesis, Technische Universität München, Apr. 2018.

[148] "DS2OS – the Distributed Smart Space Orchestration System". Technische Universität München. Accessed: Feb. 28, 2022. [Online]. Available: https://s2labs.org/?site=projects/__DS2OS

[149] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman, "Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics," *IEEE Transactions on Mobile Computing*, vol. 18, no. 8, pp. 1745–1759, Aug 2019.

[150] "Smart lighting". Philips Hue. Accessed: Jan. 12, 2022. [Online]. Available: https://www.philips-hue.com

[151] A. Hamza, H. H. Gharakheili, T. A. Benson, and V. Sivaraman, "Detecting Volumetric Attacks on IoT devices via SDN-Based Monitoring of MUD Activity," in *Proceedings of the 2019 ACM Symposium on SDN Research*, 2019, pp. 36–48.

[152] E. Lear, R. Droms, and D. Romascanu, "Manufacturer Usage Description Specification," Internet Requests for Comments, RFC Editor, RFC 8520, Mar. 2019. [Online]. Available: https://rfc-editor.org/rfc/rfc8520

[153] A. Hamza, D. Ranathunga, H. H. Gharakheili, M. Roughan, and V. Sivaraman, "Clear as MUD: Generating, validating and applying IoT behavioral profiles," in *Proceedings of the 2018 Workshop on IoT Security and Privacy*, 2018, pp. 8–14.

[154] J. Ren, D. J. Dubois, D. Choffnes, A. M. Mandalari, R. Kolcun, and H. Haddadi, "Information Exposure From Consumer IoT Devices: A Multidimensional, Network-Informed Measurement Approach," in *Proceedings of the Internet Measurement Conference (IMC)*, 2019, pp. 267–279.

[155] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[156] "Node-RED". OpenJS Foundation & Contributors. Accessed: Jan 19, 2022. [Online]. Available: https://nodered.org

[157] "Ostinato Traffic Generator for Network Engineers". Accessed: Jan 19, 2022. [Online]. Available: https://ostinato.org

[158] "Argus". Qosient. Accessed: Jan 19, 2022. [Online]. Available: https://openargus.org

[159] (2018, 27 Sep.) Torii botnet - Not another Mirai variant. Avast Threat Intelligence Team. Accessed: February 7, 2023. [Online]. Available: https://blog.avast.com/new-torii-botnet-threat-research

[160] L. Vigoya, D. Fernandez, V. Carneiro, and F. Cacheda, "Annotated dataset for anomaly detection in a data center with IoT sensors," *Sensors*, vol. 20, no. 13, p. 3745, 2020.

[161] A. Guerra-Manzanares, J. Medina-Galindo, H. Bahsi, and S. Nõmm, "MedBIoT: Generation of an IoT Botnet Dataset in a Medium-sized IoT Network," in *ICISSP*, 2020, pp. 207–218.

[162] A. K. Sarica and P. Angin, "A Novel SDN Dataset for Intrusion Detection in IoT Networks," in *2020 16th International Conference on Network and Service Management (CNSM)*. IEEE, 2020, pp. 1–5.

[163] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: rapid prototyping for software-defined networks," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, 2010, pp. 1–6.

[164] "Open Networking Operating System (ONOS) SDN Controller for SDN/NFV Solutions". Open Networking Foundation. Accessed: Jan 19, 2022. [Online]. Available: https://opennetworking.org/onos

[165] J. Quittek, T. Zseby, B. Claise, and S. Zander, "Requirements for IP flow information export (IPFIX)," Internet Requests for Comments, RFC Editor, RFC 3917, Oct. 2004. [Online]. Available: https://www.rfc-editor.org/info/rfc3917

[166] A. Pashamokhtari, N. Okui, Y. Miyake, M. Nakahara, and H. H. Gharakheili, "Inferring Connected IoT Devices from IPFIX Records in Residential ISP Networks," in *2021 IEEE 46th Conference on Local Computer Networks (LCN)*. IEEE, 2021, pp. 57–64.

[167] N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets," *Sustainable Cities and Society*, vol. 72, p. 102994, 2021.

[168] "The Zeek Network Security Monitor". Accessed: Jan 20, 2022. [Online]. Available: https://zeek.org

[169] "NFStream: Flexible Network Data Analysis Framework". NFStream Developers. Accessed: May 11, 2021. [Online]. Available: https://www.nfstream.org

[170] M. T. Paracha, D. J. Dubois, N. Vallina-Rodriguez, and D. R. Choffnes, "IoTLS: understanding TLS usage in consumer IoT devices," in *IMC '21: ACM Internet Measurement Conference, Virtual Event, USA, November 2-4, 2021*, D. Levin, A. Mislove, J. Amann, and M. Luckie, Eds. ACM, 2021, pp. 165–178. [Online]. Available: https://doi.org/10.1145/3487552.3487830

[171] "mitmproxy - an interactive HTTPS proxy". Mitmproxy Project. Accessed: Feb. 2, 2022. [Online]. Available: https://mitmproxy.org

[172] M. Ahmed, S. Byreddy, A. Nutakki, L. F. Sikos, and P. Haskell-Dowland, "ECU-IoHT: A dataset for analyzing cyberattacks in Internet of Health Things," *Ad Hoc Networks*, vol. 122, p. 102621, 2021.

[173] "MySignals - eHealth and Medical IoT Development Platform". Libelium. Accessed: Jan. 24, 2022. [Online]. Available: http://www.my-signals.com

[174] "Nmap: the Network Mapper - Free Security Scanner". Accessed: Jan. 24, 2022. [Online]. Available: https://nmap.org

[175] "Ettercap Home Page". Accessed: Jan. 24, 2022. [Online]. Available: https://www.ettercap-project.org

[176] byt3bl33d3r. "MITMf | Framework for Man-In-The-Middle attacks". Accessed: Jan. 24, 2022. [Online]. Available: https://github.com/byt3bl33d3r/MITMf

[177] I. Vaccari, G. Chiola, M. Aiello, M. Mongelli, and E. Cambiaso, "MQTTset, a New Dataset for Machine Learning Techniques on MQTT," *Sensors*, vol. 20, no. 22, p. 6578, 2020.

[178] S. Ghazanfar, F. Hussain, A. U. Rehman, U. U. Fayyaz, F. Shahzad, and G. A. Shah, "IoT-Flock: An Open-source Framework for IoT Traffic Generation," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 2020, pp. 1–6.

[179] "Eclipse Mosquitto™- An open source MQTT broker". Eclipse Foundation. Accessed: June 22, 2022. [Online]. Available: https://mosquitto.org

[180] R. Kolcun, D. A. Popescu, V. Safronov, P. Yadav, A. M. Mandalari, R. Mortier, and H. Haddadi, "Revisiting IoT Device Identification," in *5th Network Traffic Measurement and Analysis Conference, TMA 2021, Virtual Event, September 14-15, 2021*. IFIP, 2021. [Online]. Available: http://dl.ifip.org/db/conf/tma/tma2021/tma2021-paper6.pdf

[181] "Tcpdump & Libpcap". The Tcpdump Group. Accessed: Feb. 1, 2022. [Online]. Available: https://www.tcpdump.org

[182] M. Ahmed, D. Cox, B. Simpson, and A. Aloufi, "ECU-IoFT: A Dataset for Analysing Cyber-Attacks on Internet of Flying Things," *Applied Sciences*, vol. 12, no. 4, p. 1990, 2022.

[183] "tello". Ryze Tech. Accessed: June 22, 2022. [Online]. Available: https://www.ryzerobotics.com/tello

[184] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, "Labelled data collection for anomaly detection in wireless sensor networks," in *2010 sixth international conference on intelligent sensors, sensor networks and information processing*. IEEE, 2010, pp. 269–274.

[185] "TelosB Mote Platform". Crossbow Technology. Accessed: Feb. 28, 2022. [Online]. Available: https://www.willow.co.uk/TelosB_Datasheet.pdf

[186] S. Rajasegarar, C. Leckie, and M. Palaniswami, "Detecting Data Anomalies in Wireless Sensor Networks," in *Security in Ad Hoc And Sensor Networks*. World Scientific, 2010, pp. 231–259.

[187] S. Fu, Y. Zhang, Y. Jiang, C. Hu, C.-Y. Shih, and P. J. Marrón, "Experimental study for multi-layer parameter configuration of WSN links," in *2015 IEEE 35th International Conference on Distributed Computing Systems*. IEEE, 2015, pp. 369–378.

[188] "CC2420 - Single-Chip 2.4 GHz IEEE 802.15.4 Compliant and ZigBee™ Ready RF Transceiver". Texas Instruments. Accessed: Feb. 28, 2022. [Online]. Available: https://www.ti.com/product/CC2420

[189] P. Levis, S. Madden, J. Polastre, R. Szewczyk, K. Whitehouse, A. Woo, D. Gay, J. Hill, M. Welsh, E. Brewer *et al.*, "TinyOS: An operating system for sensor networks," in *Ambient intelligence*. Springer, 2005, pp. 115–148.

[190] I. Almomani, B. Al-Kasasbeh, and M. Al-Akhras, "WSN-DS: a dataset for intrusion detection systems in wireless sensor networks," *Journal of Sensors*, vol. 2016, 2016.

[191] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd annual Hawaii international conference on system sciences*. IEEE, 2000, pp. 10–pp.

[192] "KillerBee - IEEE 802.15.4/ZigBee Security Research Tool". River Loop Security. Accessed: Feb. 28, 2022. [Online]. Available: https://github.com/riverloopsec/killerbee

[193] D. Sikeridis, I. Papapanagiotou, and M. Devetsikiotis, "BLEBeacon: A real-subject trial dataset from mobile Bluetooth low energy beacons," *arXiv preprint arXiv:1802.08782*, 2018.

[194] "Series 10 Beacon". Gimbal. Accessed: Feb. 3, 2022. [Online]. Available: https://store.gimbal.com/collections/beacons/products/s10

[195] M. Inaya, M. Meli, D. Sikeridis, and M. Devetsikiotis, "A real-subject evaluation trial for location-aware smart buildings," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2017, pp. 301–306.

[196] D. Sikeridis, M. Devetsikiotis, and I. Papapanagiotou, "Occupant tracking in smart facilities: An experimental study," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2017, pp. 818–822.

[197] ——, "A cloud-assisted infrastructure for occupancy tracking in smart facilities," in *IBM Cloud Academy Conference (ICA CON)*, 2017.

[198] M. Aernouts, R. Berkvens, K. Van Vlaenderen, and M. Weyn, "Sigfox and LoRaWAN datasets for fingerprint localization in large urban and rural areas," *Data*, vol. 3, no. 2, p. 13, 2018.

[199] "Sensolus tracker - SNT 3 Ultra". Sigfox Partner Network. Accessed: Jan. 26, 2022. [Online]. Available: https://partners.sigfox.com/products/stickntrack

[200] S. Latre, P. Leroux, T. Coenen, B. Braem, P. Ballon, and P. Demeester, "City of things: An integrated and multi-technology testbed for IoT smart city experiments," in *2016 IEEE international smart cities conference (ISC2)*. IEEE, 2016, pp. 1–8.

[201] "NetSim for Researchers". Tetcos. Accessed: Feb. 4, 2022. [Online]. Available: https://www.tetcos.com/netsim-std.html

[202] "Introduction | KDD Cup 1999: Computer network intrusion detection". KDD. Accessed: Jan. 12, 2022. [Online]. Available: https://kdd.org/kdd-cup/view/kdd-cup-1999

[203] M.-P. Uwase, N. T. Long, J. Tiberghien, K. Steenhaut, and J.-M. Dricot, "Outdoors Range Measurements with Zolertia Z1 Motes and Contiki," in *Real-world wireless sensor networks*. Springer, 2014, pp. 79–83.

[204] L. Bhatia, M. Breza, R. Marfievici, and J. A. McCann, "LoED: The LoRaWAN at the edge dataset: dataset," in *Proceedings of the Third Workshop on Data: Acquisition To Analysis*, 2020, pp. 7–8.

[205] A. Al-Shawabka, P. Pietraski, S. B. Pattar, F. Restuccia, and T. Melodia, "DeepLoRa: Fingerprinting LoRa Devices at Scale Through Deep Learning and Data Augmentation," in *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2021, pp. 251–260.

[206] "Pysense 2.0 X - Pycom Multi-network Dev Board Sensor Shield". Pycom. Accessed: Jan. 26, 2022. [Online]. Available: https://pycom.io/product/pysense-2-0-x

[207] "FiPy - Pycom -Five Network Development Board for IoT". Pycom. Accessed: Jan. 26, 2022. [Online]. Available: https://pycom.io/product/fipy

[208] B. Hilburn, N. West, T. O'Shea, and T. Roy, "SigMF: the signal metadata format," in *Proceedings of the GNU Radio Conference*, vol. 3, no. 1, 2018.

[209] A. Elmaghbub and B. Hamdaoui, "LoRa device fingerprinting in the wild: Disclosing RF data-driven fingerprint sensitivity to deployment variability," *IEEE Access*, vol. 9, pp. 142 893–142 909, 2021.

[210] D. Kotz and T. Henderson, "CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth," *IEEE Pervasive Computing*, vol. 4, no. 4, pp. 12–14, 2005.

[211] D.-G. Akestoridis, M. Harishankar, M. Weber, and P. Tague, "Zigator: Analyzing the security of zigbee-enabled smart homes," in *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2020, pp. 77–88.

[212] "USRP N210 Software Defined Radio (SDR)". Ettus Research. Accessed: Jan. 24, 2022. [Online]. Available: https://www.ettus.com/all-products/un210-kit

[213] "GNU Radio – The Free & Open Source Radio Ecosystem". GNU Radio Project. Accessed: Jan. 24, 2022. [Online]. Available: https://www.gnuradio.org

[214] A. Collen, N. A. Nijdam, J. Augusto-Gonzalez, S. K. Katsikas, K. M. Giannoutakis, G. Spathoulas, E. Gelenbe, K. Votis, D. Tzovaras, N. Ghavami *et al.*, "GHOST-safe-guarding home IoT environments with personalised real-time risk control," 2018.

[215] M. Anagnostopoulos, G. Spathoulas, B. Viaño, and J. Augusto-Gonzalez, "Tracing your smart-home devices conversations: A real world IoT traffic data-set," *Sensors*, vol. 20, no. 22, p. 6600, 2020.

[216] W. A. Simpson, "The Point-to-Point Protocol (PPP)," Internet Requests for Comments, RFC Editor, RFC 1661, Jul. 1994. [Online]. Available: https://www.rfc-editor.org/info/rfc1661

[217] L. Barman, A. Dumur, A. Pyrgelis, and J.-P. Hubaux, "Every Byte Matters: Traffic Analysis of Bluetooth Wearable Devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 2, jun 2021. [Online]. Available: https://doi.org/10.1145/3463512

[218] S. M. Darroudi and C. Gomez, "Bluetooth low energy mesh networks: A survey," *Sensors*, vol. 17, no. 7, p. 1467, 2017.

[219] A. Lacava, E. Giacomini, F. D'Alterio, and F. Cuomo. "Intrusion Detection System for Bluetooth Low Energy Mesh Networks: data gathering and experimental evaluations". Accessed: Mar. 2, 2022. [Online]. Available: https://github.com/netlab-sapienza/BLEMeshIDS

[220] "ESP32 Series - Datasheet". Espressif Systems. Accessed: Mar. 2, 2022. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_en.pdf

[221] D.-G. Akestoridis and P. Tague, "HiveGuard: A Network Security Monitoring Architecture for Zigbee Networks," in *2021 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2021, pp. 209–217.

[222] Kamaldeep, M. Malik, M. Dutta, and J. Granjal. "IoT-CIDDS". Accessed: June 22, 2022. [Online]. Available: https://github.com/manishamalik53/IoT-CIDDS

[223] "Cooja Simulator". Contiki. Accessed: Jan. 11, 2022. [Online]. Available: https://anrg.usc.edu/contiki/index.php/Cooja_Simulator

[224] M. Zolanvari, M. A. Teixeira, and R. Jain, "Effect of Imbalanced Datasets on Security of Industrial IoT Using mMchine Learning," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2018, pp. 112–117.

[225] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, "Machine Learning-Based Network Vulnerability Analysis of Industrial Internet of Things," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6822–6834, 2019.

[226] M. Zolanvari, Z. Yang, K. Khan, R. Jain, and N. Meskin, "TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security," *IEEE internet of things journal*, 2021.

[227] M. Al-Hawawreh and E. Sitnikova, "Developing a Security Testbed for Industrial Internet of Things," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5558–5573, 2020.

[228] M. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, "X-iiotid: A connectivity-agnostic and device-agnostic intrusion data set for industrial internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3962–3977, 2021.

[229] J. Mathews, P. Chatterjee, and S. Banik, "CoAP-DoS: An IoT Network Intrusion Data Set," in *2022 6th International Conference on Cryptography, Security and Privacy (CSP)*. IEEE, 2022, pp. 91–95.

[230] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40 281–40 306, 2022.

[231] "1998 DARPA Intrusion Detection Evaluation Dataset". MIT Lincoln Laboratory. Accessed: Feb. 02, 2022. [Online]. Available: https://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset

[232] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.

[233] "UCI Machine Learning Repository". University of California, Irvine, School of Information and Computer Sciences. Accessed: June 28, 2022. [Online]. Available: https://archive.ics.uci.edu/ml/index.php

[234] T. Henderson, D. Kotz, and I. Abyzov, "The changing usage of a mature campus-wide wireless network," in *Proceedings of the 10th annual international conference on Mobile computing and networking*, 2004, pp. 187–201.

[235] J. Xu, J. Fan, M. H. Ammar, and S. B. Moon, "Prefix-preserving IP address anonymization: measurement-based security evaluation and a new cryptography-based scheme," in *10th IEEE International Conference on Network Protocols, 2002. Proceedings.* IEEE, 2002, pp. 280–289.

[236] B. Sangster, T. J. O'Connor, T. Cook, R. Fanelli, E. Dean, W. J. Adams, C. Morrell, and G. Conti, "Toward instrumenting network warfare competitions to generate labeled datasets," in *Proceedings of the 2nd Conference on Cyber Security Experimentation and Test*, ser. CSET'09. USA: USENIX Association, 2009, p. 9.

[237] "Cyber Defense Exercise (CDX)". National Security Agency | Central Security Service. Accessed: Feb. 1, 2022. [Online]. Available: https://apps.nsa.gov/iaarchive/programs/cyber-defense-exercise/index.cfm

[238] N. Moustaf and J. Slay, "Creating novel features to anomaly network detection using DARPA-2009 data set," in *Proceedings of the 14th European Conference on Cyber Warfare and Security. Academic Conferences Limited*, 2015, pp. 204–212.

[239] "WITS: Waikato VIII". WAND Network Research Group. Accessed: Jan. 28, 2022. [Online]. Available: https://wand.net.nz/wits/waikato/8

[240] "WITS: ISPDSL-II". WAND Network Research Group. Accessed: Jan. 28, 2022. [Online]. Available: https://wand.net.nz/wits/ispdsl/2

[241] "WITS: IPLS III". WAND Network Research Group. Accessed: Jan. 28, 2022. [Online]. Available: https://wand.net.nz/wits/ipls/3

[242] "Abilene Network". Abilene. Accessed: Jan. 28, 2022. [Online]. Available: https://web.archive.org/web/20120324103518/http://www.internet2.edu/pubs/200502-IS-AN.pdf

[243] S. Alcock, P. Lorier, and R. Nelson, "Libtrace: A packet capture and analysis library," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 2, pp. 42–48, 2012.

[244] "libtrace". WAND Network Research Group. Accessed: Jan. 28, 2022. [Online]. Available: https://research.wand.net.nz/software/libtrace.php

[245] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, "Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation," in *Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, ser. BADGERS '11. New York, NY, USA: Association for Computing Machinery, Apr. 2011, pp. 29–36. [Online]. Available: https://doi.org/10.1145/1978672.1978676

[246] "Tstat - TCP STatistic and Analysis Tool". Telecommunication Networks Group - Politecnico di Torino. Accessed: Feb. 9, 2022. [Online]. Available: http://tstat.polito.it

[247] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Computers & Security*, vol. 31, no. 3, pp. 357–374, 2012.

[248] G. Creech and J. Hu, "Generation of a new IDS test dataset: Time to retire the KDD collection," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2013, pp. 4487–4492.

[249] "HomePage". Tiki Wiki CMS Groupware. Accessed: Feb. 3, 2022. [Online]. Available: https://tiki.org/HomePage

[250] "Tiki Wiki CMS Groupware 8.2 - 'snarf_ajax.php' Remote PHP Code Injection". Exploit Database. Accessed: Feb. 3, 2022. [Online]. Available: https://www.exploit-db.com/exploits/18265

[251] Maciejak, David and Dlabal, Jan. "Hydra". Accessed: Feb. 3, 2022. [Online]. Available: https://github.com/vanhauser-thc/thc-hydra

[252] "What is Meterpreter ? - Security Wiki". Secret Double Octopus. Accessed: Feb. 3, 2022. [Online]. Available: https://doubleoctopus.com/security-wiki/threats-and-tools/meterpreter

[253] S. García, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," *Computers & Security*, vol. 45, pp. 100–123, 2014.

[254] B. Claise, "Cisco Systems NetFlow Services Export Version 9," Internet Requests for Comments, RFC Editor, RFC 3954, Oct. 2004. [Online]. Available: https://www.rfc-editor.org/info/rfc3954

[255] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, Nov 2015, pp. 1–6.

[256] "PerfectStorm". Keysight Technologies. Accessed: Jan. 26, 2022. [Online]. Available: https://www.keysight.com/be/en/products/network-test/network-test-hardware/perfectstorm.html

[257] G. Maciá-Fernández, J. Camacho, R. Magán-Carrión, P. García-Teodoro, and R. Therón, "UGR '16: A new dataset for the evaluation of cyclostationarity-based network IDSs," *Computers & Security*, vol. 73, pp. 411–424, 2018.

[258] "Datasets". Canadian Institute for Cybersecurity. Accessed: Mar. 3, 2022. [Online]. Available: https://www.unb.ca/cic/datasets/index.html

[259] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," *ICISSp*, vol. 1, pp. 108–116, 2018.

[260] I. Sharafaldin, A. Gharib, A. H. Lashkari, and A. A. Ghorbani, "Towards a reliable intrusion detection benchmark dataset," *Software Networking*, vol. 2018, no. 1, pp. 177–200, 2018.

[261] Z. Durumeric, F. Li, J. Kasten, J. Amann, J. Beekman, M. Payer, N. Weaver, D. Adrian, V. Paxson, M. Bailey *et al.*, "The Matter of Heartbleed," in *Proceedings of the 2014 conference on internet measurement conference*, 2014, pp. 475–488.

[262] "CICFlowMeter". Canadian Institute for Cybersecurity. Accessed: Jan. 24, 2022. [Online]. Available: https://www.unb.ca/cic/research/applications.html

[263] "A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018)". Registry of Open Data on AWS. Accessed: Mar. 03, 2021. [Online]. Available: https://registry.opendata.aws/cse-cic-ids2018

[264] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *2019 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2019, pp. 1–8.

[265] M. MontazeriShatoori, L. Davidson, G. Kaur, and A. Habibi Lashkari, "Detection of DoH Tunnels using Time-series Classification of Encrypted Traffic," in *2020 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)*, 2020, pp. 63–70.

[266] P. E. Hoffman and P. McManus, "DNS Queries over HTTPS (DoH)," Internet Requests for Comments, RFC Editor, RFC 8484, Oct. 2018. [Online]. Available: https://www.rfc-editor.org/info/rfc8484

[267] M. MontazeriShatoori, L. Davidson, G. Kaur, and A. H. Lashkari. (2019) "DoHlyzer". Accessed: Feb. 4, 2022. [Online]. Available: https://github.com/ahlashkari/DoHlyzer

[268] R. Damasevicius, A. Venckauskas, S. Grigaliunas, J. Toldinas, N. Morkevicius, T. Aleliunas, and P. Smuikys, "LITNET-2020: An annotated real-world network flow dataset for network intrusion detection," *Electronics*, vol. 9, no. 5, p. 800, 2020.

[269] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," *Information Security Journal: A Global Perspective*, vol. 25, no. 1-3, pp. 18–31, 2016.

[270] E. Chatzoglou, G. Kambourakis, and C. Kolias, "Empirical Evaluation of Attacks Against IEEE 802.11 Enterprise Networks: The AWID3 Dataset," *IEEE Access*, vol. 9, pp. 34 188–34 205, 2021.

[271] "IEEE Standard for Information technology - Telecommunications and information exchange between systems - Local and metropolitan area networks - Specific requirements. Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 4: Protected Management Frames". IEEE Standards Association. Accessed: Feb. 3, 2022. [Online]. Available: https://standards.ieee.org/ieee/802.11w/3736

[272] M. Sarhan, S. Layeghy, N. Moustafa, and M. Portmann, "NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems," in *Big Data Technologies and Applications*, Z. Deze, H. Huang, R. Hou, S. Rho, and N. Chilamkurti, Eds. Cham: Springer International Publishing, 2021, pp. 117–135.

[273] M. Sarhan, S. Layeghy, and M. Portmann, "Towards a standard feature set for network intrusion detection system datasets," *Mobile Networks and Applications*, pp. 1–14, 2021.

[274] E. Chatzoglou, V. Kouliaridis, G. Kambourakis, G. Karopoulos, and S. Gritzalis, "A hands-on gaze on HTTP/3 security through the lens of HTTP/2 and a public dataset," *Computers & Security*, vol. 125, p. 103051, 2023.

[275] M. Belshe, R. Peon, and M. Thomson, "Hypertext Transfer Protocol Version 2 (HTTP/2)," Internet Requests for Comments, RFC Editor, RFC 7540, May 2015. [Online]. Available: https://www.rfc-editor.org/rfc/rfc7540

[276] M. Bishop, "HTTP/3," Internet Requests for Comments, RFC Editor, RFC 9114, Jun. 2022. [Online]. Available: https://www.rfc-editor.org/rfc/rfc9114

[277] J. Iyengar and M. Thomson, "QUIC: A UDP-Based Multiplexed and Secure Transport," Internet Requests for Comments, RFC Editor, RFC 9000, May 2021. [Online]. Available: https://www.rfc-editor.org/rfc/rfc9000

[278] L. Eggert, "Towards Securing the Internet of Things with QUIC," in *Workshop on Decentralized IoT Systems and Security (DISS)*, 2020.

[279] E. Chatzoglou, V. Kouliaridis, G. Karopoulos, and G. Kambourakis, "Revisiting QUIC attacks: A comprehensive review on QUIC security and a hands-on study," *International Journal of Information Security*, pp. 1–19, 2022.

[280] F. Carrino, A. Janka, O. Abou Khaled, and E. Mugellini, "LoRaLoc: Machine learning-based fingerprinting for outdoor geolocation using LoRa," in *2019 6th Swiss Conference on Data Science (SDS)*. IEEE, 2019, pp. 82–86.

[281] J. Purohit, X. Wang, S. Mao, X. Sun, and C. Yang, "Fingerprinting-based indoor and outdoor localization with LoRa and deep learning," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[282] M. Cermák, T. Jirsík, P. Velan, J. Komárková, S. Spacek, M. Drasar, and T. Plesnik, "Towards provable network traffic measurement and analysis via semi-labeled trace datasets," in *Network Traffic Measurement and Analysis Conference, TMA 2018, Vienna, Austria, June 26-29, 2018*. IEEE, 2018, pp. 1–8. [Online]. Available: https://doi.org/10.23919/TMA.2018.8506498

[283] I. Zavalyshyn, "Building Private-by-Design IoT Systems," Ph.D. dissertation, NOVA University, Portugal 2021.

[284] P. M. Chanal and M. S. Kakkasageri, "Security and privacy in IOT: a survey," *Wireless Personal Communications*, vol. 115, pp. 1667–1693, 2020.

[285] C. Perera, R. Ranjan, L. Wang, S. U. Khan, and A. Y. Zomaya, "Big data privacy in the internet of things era," *IT Professional*, vol. 17, no. 3, pp. 32–39, 2015.

[286] R. Pang, M. Allman, V. Paxson, and J. Lee, "The devil and packet trace anonymization," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 29–38, 2006.

[287] M. Mohammady, L. Wang, Y. Hong, H. Louafi, M. Pourzandi, and M. Debbabi, "Preserving both privacy and utility in network trace anonymization," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 459–474.

[288] M. Burkhart, D. Schatzmann, B. Trammell, E. Boschi, and B. Plattner, "The role of network trace anonymization under attack," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 5–11, 2010.

[289] A. Botta, A. Dainotti, and A. Pescapè, "A tool for the generation of realistic network workload for emerging networking scenarios," *Computer Networks*, vol. 56, no. 15, pp. 3531–3547, 2012.

[290] "D-ITG, Distributed Internet Traffic Generator". Accessed: Feb. 15, 2022. [Online]. Available: http://traffic.comics.unina.it/software/ITG

[291] J. Sommers, H. Kim, and P. Barford, "Harpoon: a flow-level traffic generator for router and network tests," *ACM SIGMETRICS Performance Evaluation Review*, vol. 32, no. 1, pp. 392–392, 2004.

[292] J. Sommers. (2005, 7 Dec.) "Harpoon: A Flow-level Traffic Generator". Accessed: Feb. 15, 2022. [Online]. Available: https://jsommers.github.io/harpoon

[293] xorrsprsp. (2017, 28 Nov.) "fudp - threaded udp flooder". Accessed: Feb. 15, 2022. [Online]. Available: https://github.com/xorrsprsp/fudp

[294] (2019) "Trex". Cisco. Accessed: Sept. 26, 2022. [Online]. Available: https://trex-tgn.cisco.com

[295] "IoT-Flock - An Open-Source Framework For IoT Traffic Generation". Accessed: June 24, 2022. [Online]. Available: https://github.com/ThingzDefense/IoT-Flock

[296] W.-H. Hsu, Q. Li, X.-H. Han, and C.-W. Huang, "A hybrid IoT traffic generator for mobile network performance assessment," in *2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2017, pp. 441–445.

[297] S. Hui, H. Wang, Z. Wang, X. Yang, Z. Liu, D. Jin, and Y. Li, "Knowledge Enhanced GAN for IoT Traffic Generation," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3336–3346.

[298] H. Nguyen-An, T. Silverston, T. Yamazaki, and T. Miyoshi, "Generating IoT Traffic in Smart Home Environment," in *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)*, 2020, pp. 1–2.

[299] A. Lavric, A. I. Petrariu, E. Coca, and V. Popa, "LoRa Traffic Generator Based on Software Defined Radio Technology for LoRa Modulation Orthogonality Analysis: Empirical and Experimental Evaluation," *Sensors*, vol. 20, no. 15, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/15/4123

[300] M. Gucciardo, I. Tinnirello, and D. Garlisi, "A cell-level traffic generator for LoRa networks," in *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 2017, pp. 474–476.

[301] J. Franco, A. Aris, B. Canberk, and A. S. Uluagac, "A Survey of Honeypots and Honeynets for Internet of Things, Industrial Internet of Things, and Cyber-Physical Systems," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2351–2383, 2021.