

A novel framework for analyzing internet of things datasets for machine learning and deep learning-based intrusion detection systems

Muhammad Arief^{1,2}, Made Gunawan², Agung Septiadi², Mukti Wibowo², Vitria Pragesjvara²,
Kusnanda Supriatna², Anto Satriyo Nugroho², I Gusti Bagus Baskara Nugraha^{1,3},
Suhono Harso Supangkat^{1,3}

¹School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, Indonesia

²Research Center for Artificial Intelligence and Cyber Security, National Research and Innovation Agency, Jakarta, Indonesia

³Smart City and Community Innovation Center, Bandung Institute of Technology, Bandung, Indonesia

Article Info

Article history:

Received Jul 28, 2023

Revised Sep 14, 2023

Accepted Nov 7, 2023

Keywords:

Cyber-attack

Deep learning

Internet of things dataset

Intrusion detection system

Machine learning

ABSTRACT

To generate a machine learning (ML) and deep learning (DL) architecture with good performance, we need a decent dataset for the training and testing phases of the development process. Starting with the knowledge discovery and data mining (KDD) Cup 99 dataset, numerous datasets have been produced since 1998 to be utilized in the ML and DL-based intrusion detection systems (IDS) training and testing process. Because there are so many datasets accessible, it might be challenging for researchers to choose which dataset to employ. Therefore, a framework for evaluating dataset appropriateness with the research to be conducted is becoming increasingly crucial as new datasets are regularly created. Additionally, given the growing popularity of internet of things (IoT) devices and an increasing number of specific datasets for IoT in recent years, it is essential to have a specific framework for IoT datasets. Therefore, this research aims to develop a new framework for evaluating IoT datasets for ML and DL-based IDS. The study's findings include, first, a novel framework for assessing IoT datasets, second, a comparison of this novel framework to other existing frameworks, and third, an analysis of five IoT datasets by using the new framework.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Muhammad Arief

School of Electrical Engineering and Informatics, Bandung Institute of Technology

Bandung, Indonesia

Email: 33221045@std.stei.itb.ac.id

1. INTRODUCTION

The internet of things (IoT) allows various devices enable various devices to have the ability to connect to networks, interact with them, and share data [1], [2]. IoT has a complex architecture, making security implementation challenging. Until now, the complexity of IoT devices has expanded, making IoT systems become more vulnerable to various attacks. One of the primary considerations is having a highly secure IoT device. By exploiting and controlling networks, stealing, modifying, or destroying user data, cyber-attacks target numerous IoT devices. As a result of this, the primary concerns are to safeguard the IoT device's availability, data confidentiality, and integrity.

Using intrusion detection systems (IDS) is one method for defending the IoT system against online threats. IDS [3]–[5] are used to quickly identify and classify attacks on hosts and network infrastructure in real time. IDS are classified into three categories based on how they identify anomalies: anomaly-based, signature-

based, and hybrid IDS. In general, a signature-based approach performs better against known cyber-attacks, whereas an anomaly-based approach performs better against unknown attacks. The drawback of the anomaly-based detection approach is that it has the potential to create a significant number of false positives. The majority of artificial intelligence research on anomaly-based IDS uses machine learning (ML) and deep learning (DL) techniques to create models for detecting intrusion.

The establishment of ML/DL architecture requires a good dataset for the training and testing process so that we can obtain an architecture with high performance. Starting with the knowledge discovery and data mining (KDD) Cup 99 dataset, numerous datasets have been generated since 1998 for use in the ML/DL IDS training and testing process. Currently, there are about 40 datasets available. The availability of so many datasets makes it difficult for researchers to determine which dataset to use.

The development of a framework for evaluating ML/DL IDS datasets has been attempted in numerous publications [6]–[10], however, none has yet established a standard. A framework for evaluating the appropriateness of the study to be carried out is becoming increasingly essential especially for IoT datasets, as new datasets are constantly being created. In recent years, there has been an increase in both the use of IoT devices and the number of IoT-specific datasets, necessitating the development of an IoT-specific framework for analyzing datasets. Therefore, the primary goal of this research is to provide a new framework for analyzing IoT datasets. We believe our framework is much more comprehensive than other frameworks for evaluating datasets. The contributions of this research are as follows: i) Proposed a new framework for analyzing IoT datasets for ML/DL-based IDS; ii) Comparing the new framework to the existing dataset analysis frameworks; iii) Analyzing five IoT datasets from 2019-2022 by using the new framework; iv) The developer of the IoT dataset can utilize this new framework as a guideline for generating their IoT datasets.

The following is the structure of this article: section 2 will describe the research method used in this research. Section 3 will discuss the literature review/related works regarding the existing framework and available IoT dataset. Section 4 discusses the proposed new framework for analyzing IoT datasets, the results, and findings of the study and experiments. Section 5 discusses the conclusions.

2. METHOD

The main objective of this research is to develop a novel framework that can be used by researchers to analyze IoT datasets for ML/DL IDS research in IoT systems. After analyzing the datasets using this novel framework, researchers can determine which datasets they want to utilize in accordance with their research objectives. As previously indicated, the performance of the model being constructed will depend on the dataset used. Additionally, it is also important to understand that creating an ML/DL architecture for IDS is not a magical process in which we create an excellent design without comprehending the details of the dataset.

To achieve this research goal, the methodology we use is as follows, the first step we took was to review articles describing frameworks for selecting IDS datasets, especially those related to network traffic datasets in the last five years. Second, because IoT technology has advanced so quickly in recent years, we also analyze articles that are relevant to the IoT system to comprehend the most recent technological advancements. Third, we examine the network traffic structure of five recent IoT datasets that were created in the last five years.

Then a new framework for analyzing IoT Datasets for ML and DL-based IDS is developed. The aspects of the new framework are then compared with those of the four existing frameworks. As a final step, in order to demonstrate the advantages of utilizing the new framework in analyzing IoT datasets, the new framework is evaluated on five IoT datasets. By looking at the assessment results, researchers may select the best dataset for their research.

3. LITERATURE REVIEW/RELATED WORKS

3.1. Existing dataset analysis frameworks

This section will describe four frameworks that have been created to evaluate network datasets despite the fact that some of them do not specifically address IoT datasets. The order of this evaluation is based on the year that the framework was created. The most recent framework was developed in 2022, while the earliest was developed in 2016.

In 2022, Al-Hawawreh *et al.* [6] introduces a framework as a guide for creating datasets that can be used for evaluating, testing, and tuning solutions to security issues in industrial internet of things (IIoT) systems. The framework proposed in this article consists of some aspects that can be used to examine IIoT datasets, including complete system and network configuration, heterogeneous data sources, complete capture, realistic normal network traffic, diverse attack typers, diverse data collection duration, feature set, recent IIoT application protocols, recent attacks, agnostic features, IIoT traces, fully labeled, metadata, public availability.

This framework is unable to explain the different kinds of attacks that can be found in a dataset, data uniqueness, availability of raw data, as well as benign traffic, and information about the balanced dataset.

In 2020, Kenyon *et al.* [7] proposed a framework for analyzing datasets. This framework provides a 'best-practice' guide in creating datasets and has 9 aspects that are mandatory and 5 aspects that are desirable. Nine mandatory aspects include dataset provenance, domain context, consistent labeling, representative events, sample duration, temporal scope, and geospatial scope. This framework is not specifically intended for IoT datasets, some aspects of the framework do not have clear boundaries, and there are no aspects related to balanced data, data uniqueness, protocol type, and separation of datasets for training and testing required for building ML/DL-based IDS models.

In 2019, Ring *et al.* [8] explains that the selection of the dataset to be used depends on the research needs. In this paper, they evaluate 34 datasets by comparing 15 aspects, so that the dataset selection process can be done more easily. The aspects observed are when the dataset was created, public availability, presence of normal and attack traffic, availability of metadata, dataset format, data anonymity, data volume and duration, traffic type, network type, complete network, predefined splits, and balanced data and labeled data. This research does not determine which dataset is the best or most important because the use of datasets depends on the research being conducted, but the results of this study can be used to help researchers determine which datasets will be used for ML/DL research in cybersecurity by looking at the characteristics considered important. This framework is not specifically intended for IoT datasets, and there is no information regarding the detailed type of attacks, protocol type, availability of raw data, when the dataset last updated, and data uniqueness.

In 2016, [9], [10], consider the need for dynamically generated IDS datasets, which not only reflect network and intrusion composition but also can be modified, further developed, and reproducible. Datasets like these are demanded because of changing behavior and network patterns and growing attacks. Therefore, they proposed an evaluation framework for intrusion detection datasets. The framework has the following aspects, complete network configuration, complete traffic, labeled dataset, complete interaction, complete capture, available protocols, attack diversity, anonymity, heterogeneity, feature sets, and metadata. This framework is not specifically developed for IoT dataset analysis and there are some important aspects such as a balanced dataset, traffic volume, public availability, and data uniqueness that are not included in the framework aspects.

More thorough information on these frameworks can be found in the cited publication. Because all framework creators do not provide identifiers for their frameworks, for ease of usage, we shall henceforth refer to them as Al-Hawawreh's framework, Kenyon's framework, Ring's framework, and Gharib's framework. We are going to use the naming in the following sections.

3.2. IoT datasets

The dataset to be used in this study consists of five IoT datasets produced between 2019 and 2022. The choice to use datasets from the previous five years was made considering that they would cover the latest attacks and IoT devices. The IoT datasets are Edge-IIoTset dataset, X-IIoTID dataset, TON_IoT dataset, Bot-IoT dataset, and Aposemat IoT-23.

The Edge-IIoTset dataset was generated in 2022 by Ferrag *et al.* [11] using a testbed composed of 7 layers: edge layer, cloud computing layer, network function virtualization (NFV) layer, blockchain layer, fog layer, software defined network (SDN) layer, and IoT/IIoT perception layer. During the simulation, data is retrieved, captured, and saved in pcap file format using the Zeek and Wireshark tools. There are 63 features in this dataset, which are separated by the layer protocol. Extrapolated features are categorized into the following: internet protocol (IP), address resolution protocol (ARP), internet control message protocol (ICMP), hypertext transfer protocol (HTTP), transmission control protocol (TCP), user datagram protocol (UDP), domain name system (DNS), message queuing telemetry transport (MQTT), and modbus TCP (MBTCP).

The X-IIoTID dataset was generated by Al-Hawawreh *et al.* [6] in 2022. The X-IIoTID dataset simulates the Brown-IIoTbed testbed, which has been set up in the IoT lab at the University of New South Wales (UNSW) in Canberra and captures data from network traffic in an end-to-end method. The Industrial Internet Reference architectural (IIRA) model served as the foundation for the lab architectural design. Additionally, information on resources from edge gateways is gathered in this dataset. The author utilized zeek and dumpcap to extract essential network data from connection logs and store it in pcapng file format. A total of 68 features from host logs, resources, and other sources have been captured in this dataset.

TON_IoT is an IoT/IIoT dataset generated in 2020 by [12], [13]. In order to link physical and simulation systems with the Industry 4.0/Industrial IoT (IIoT) testbed built at the Research Cyber Range lab of UNSW Canberra. The designed Testbed architecture consists of three layers, namely the Edge/IoT layer, the Fog layer, and the Cloud layer. The ToN-IoT dataset includes network traffic from IoT networks as well as telemetry data from IoT/IIoT services. There are 45 features captured in this dataset, which is divided into the

following seven categories: Connection activity, Statistical activity, DNS activity, SSL activity, HTTP activity, Violation activity, and Data labeling.

The Bot-IoT dataset, which Koroniotis *et al.* [14] presented in 2019, replicates IoT network traffic. In order to detect and identify botnets on IoT dedicated networks, they developed a testbed based on three elements, namely network platforms, simulated IoT services, and extracting features and forensics analytics at the Research Cyber Range lab of UNSW Canberra. This dataset simulates the presence of IoT devices such as thermostats, garage doors, refrigerators, weather monitoring systems, and lights. There are 29 features in this dataset that were captured in the pcap file format.

The Aposemat IoT-23 dataset was generated by Parmisano *et al.* [15] in 2019. The dataset was created by simulation at the Stratosphere Laboratory, CTU University, Czech Republic. This testbed includes three actual IoT devices: an Amazon Echo smart home personal assistant, a Philips HUE smart light-emitting diode (LED) light, and a Somfy smart door lock. The Aposemat IoT-23 dataset is made up of 23 captures (referred to as scenarios) of various IoT network traffic. This dataset contains traffic that was recorded as pcap files.

4. RESULTS AND DISCUSSION

4.1. A novel framework for selecting IoT dataset

To develop a new framework, we started by thoroughly examining the structure of network traffic because this would serve as a fundamental guide for what characteristics of network traffic should be acquired. Subsequently, examine the dataset specifications required for ML/DL research in IDS. Then a study was carried out on several existing dataset analysis frameworks. Four frameworks were found to be pertinent to this study from the study's findings [6]–[10].

In addition to the findings of the aforementioned investigation, in this section, we propose a novel approach for analyzing IoT datasets. Using this new framework, one may select the most appropriate dataset for a certain research project. As shown in Table 1, the proposed new framework comprises 19 aspects that must be investigated and categorized into 3 groups. The 19 aspects of the framework will be thoroughly detailed in the following paragraphs.

Table 1. Aspect of the novel framework for analysing IoT dataset

No.	Group	Aspects
1	Background information	Dataset generation time, dataset metadata, dataset source location, dataset feature description, open and free to the public.
2	Data information	Labeled dataset, privacy and data protection, availability of raw data, updated dataset, benign traffic, type of attacks, balanced dataset, training-testing dataset splits, unique data entry, traffic volume.
3	Network information	Network topology, iot datasources, traffic generation, protocol type.

Dataset generation time is an aspect that indicates when a dataset was produced. During the process of selecting IoT datasets, it is important to take into account the time when the dataset was created. Numerous datasets that have been produced are significantly out of date; some were even created more than 20 years ago. Inevitably the older the dataset, the fewer types of attacks that can be detected. The value for this aspect can be obtained in the timestamp of the dataset. If the timestamp for the dataset is not available, the value can be retrieved in the dataset's documentation. Value: Indicate if it is yes or no and the specific date.

Dataset metadata, this aspect provides an explanation of the information that the dataset has; the more detailed the metadata's contents are, the easier it will be for users to fully understand the dataset. The dublin core metadata initiative's definitions and component metadata [16] can be utilized as a baseline. Value: indicate whether or not there is available metadata in detail.

Dataset source location, this aspect indicates whether a location is provided from where the user can download the dataset or not, this is to ensure that the user can use the original data that has not been modified. It is also important to note that there is evidence to prove that the dataset has not changed, for instance through the use of hashcode. Value: yes or no, provide the URL address for the data source location.

Dataset feature description, this aspect describes the set of network and IoT traffic that is captured, including whether the entirety contains all of the features of the traffic or merely a portion of them. Additionally, this information reveals the format of the network traffic that is represented in the dataset, requiring that each feature of the traffic be accompanied by details regarding the data that may be filled in. At the very least, each feature should include a description, a data format, a data range, etc. This will have an impact on the model that emerges from the dataset training procedure. Value: yes, no, or partial, explain the feature description, including the data type, format, and range.

Open and free to the public, open and free to the public, this aspect specifies whether the dataset is publicly accessible and free to the public [17]. Open datasets are accessible to the general public online and are available in machine-readable formats. The term "free" describes the accessibility of datasets to people, organizations, initiatives, and researchers without charge. Additionally, "private" information should not be included in datasets that are accessible to the general public. Simulated datasets typically do not have "privacy" issues. According to a survey of datasets generated between 2016 and 2020, only 49 (or 79%) of 62 datasets are available to the public [18]. Value: yes or no, give the prerequisite for use if any.

Labeled dataset, datasets can be classified into two categories, namely labeled datasets and unlabeled datasets, labeled datasets are used for supervised learning, whereas unlabeled datasets are used for unsupervised learning. For the purpose of training ML/DL-based network intrusion detection system (NIDS) systems, high-quality labeled datasets are required [19]. Consequently, is essential to understand if the dataset has a label or not in order to specifically match it to research purposes. Value: yes or no, if yes, give the detail of whether the label is bi-class or multiple-class.

Privacy and data protection, this aspect indicates whether there is protection for the privacy and data of the user, for example by anonymizing it so that the dataset cannot be used to identify the user. IoT technology can cause privacy problems due to data collection including data that can lead to identifying personal information via user devices such as ip addresses, mac addresses, browser fingerprints, usernames, passwords, etc. For collaborative research across numerous organizations in the development of ML models, data privacy concerns are becoming more and more significant [20], [21]. Value: yes or no, and if applicable, describe how it was made anonymous.

Availability of raw data, this aspect indicates whether or not raw traffic data is provided [22], [23]. Even though not all users require raw traffic data, this will be beneficial if certain users would like to specifically evaluate raw traffic data in order to discover more about the traffic that occurs. Value: yes or no; if yes, specify the type of raw data.

Updated dataset, this aspect reveals whether the dataset creator regularly updates the dataset. Because the types of attacks continue to evolve and IoT devices also continue to grow in number rapidly, therefore, it is important to take into consideration adding the types of available datasets and the types of attacks, as well as the date and nature of the dataset's latest update. Value: yes or no, provide the date of the most recent update.

Benign traffic, this aspect shows whether normal traffic is also available [24], and if so, whether the current normal traffic adequately represents the entirety of the traffic that commonly happens. It is very important to have normal traffic because IDS is used to monitor a network or system for attacks or policy breaches among normal traffic. Value: yes or no.

Type of attacks, this aspect indicates the different sorts of attacks included in the dataset (including a list of the layers that have been targeted and the different types of attacks based on Open Systems Interconnection (OSI)/IoT layer). Additionally, it demonstrates if the dataset contains a full range of attacks or not. The more comprehensive the sorts of attacks are, the built model will be more effective at spotting prospective attacks. Additionally, in order to demonstrate whether the model created using this dataset can be used to detect the most recent attacks, it is also preferable to identify the most recent types of attacks in the dataset. This is done by highlighting the various attack types that can affect IoT systems [25] as well as the vulnerable system layers [26]. Value: a list of the available attacks.

Balanced dataset, this aspect indicates whether the dataset is balanced or imbalanced, either for biclass or multiclass classification. Dataset considered imbalance as the amount of data in certain categories include significantly less data than others. An imbalanced dataset needs to be processed to make it balanced before being used to generate a good ML/DL model because when using traditional classification methods, this can cause the majority of classifications to tend to categories with much larger amounts of data and ignore categories with the small amount of data, causing the classification accuracy in this category to be low [27]. Value: balanced or imbalanced for either bi-class and multi-class.

Training-testing dataset splits, this aspect indicates whether there is a standard training and testing dataset provided so that every user can compare the results of the models generated by various researchers during the training and testing process. There are several techniques for splitting datasets into training data and testing data [28]. Value: yes, partial or no. If the training dataset and testing dataset are integrated into one dataset, answer with "yes partial", if the training dataset and testing dataset are in separate datasets, answer with "yes" and "no" if there is no training-testing dataset available, also describe the location from where to download the training and testing dataset.

Unique data entry, this information indicates whether or not the dataset contains duplicate data. The model that is produced will depend on how much data is duplicated. Value: yes or no.

Traffic volume, this aspect shows the size of the traffic in the dataset [29], does not display the dataset's size in gigabytes but rather the number of instances that were recorded in the dataset. The ML and DL training and the testing process require large and complete data so that the greater the number of instances, the

more accurate the model will be. Value: provide the number of instances in the dataset, if official information is not available, you may calculate the number of instances using the dataset.

Network topology, this aspect reveals the configuration of the system and network, the context in which the system is used, how the internal network interacts with it, and if it covers a wide network or not [30]. If this dataset is simulated, it is important to describe how the testbed configuration was created because this will allow us to determine whether or not it closely resembles the real network architecture. Value: yes or no, provide the network topology if known.

IoT datasources, this aspect identifies the IoT devices that contribute to the dataset or from which IoT devices the traffic is generated [31]. The dataset will be more valuable if it incorporates more IoT data sources. Value: provide a list of the IoT devices from which the traffic is collected.

Traffic generation, this aspect demonstrates the creation of traffic [32]. If it is derived from real traffic, it should be provided the duration that it will take to generate the dataset. If it was produced by simulation, it should be recognized whether or not it resembles real traffic. If there is a combination of both real and simulated traffic, it can be classified as simulation, and the time needed is calculated by summing the data from all of the sub-simulations. Value: real or simulated traffic, provide information about the time period during which the data was created.

Protocol type, this aspect shows which protocols are comprised in the dataset, including whether an IoT protocol is present. Because IoT requires "light-weight" communication to minimize the additional overhead incurred in internet connection, it employs a different protocol than the protocol used for network systems. In this instance, IoT utilizes protocols like MQTT, constrained application protocol (CoAP), extensible messaging and presence protocol (XMPP), RESR, and web socket to make communication systems simpler and faster. Consequences of this include security risks to the underlying IoT protocol [33]. Value: provide the list of all available protocol types.

4.2. Comparing the aspects of the novel framework to existing frameworks

The next step is to map the aspects of the four frameworks that were covered in section 3 by comparing their definitions with the newly established framework. This stage is challenging because it involves a thorough comprehension of how each existing framework and the new framework define their respective aspects. According to the study's findings, the following conditions exist: i) Aspects have the same name and definition; ii) Aspects have different names but the definition is similar; and iii) Aspects have a broad definition that includes several aspects in the new framework.

Table 2 contains the analysis results from the comparison of aspects and definitions in the four frameworks with the new framework. The aspects that are compatible with the specification of this new framework are listed in the Table 2. It is apparent that the degree of similarity between the new framework and the existing framework varies; there are 9 aspects that are similar when compared to Al-Hawawreh's Framework, whereas there are 9 aspects that are similar when compared to Kenyon's Framework, 12 aspects are similar when compared to Ring's Framework, and 10 aspects are similar when compared to Gharib's Framework. The aforementioned circumstances have led us to conclude that our newly proposed framework will function better and be able to assist researchers in determining which IoT dataset is best appropriate for their research. In the following section, we will demonstrate how this innovative framework might be applied to five IoT datasets.

4.3. IoT dataset analysis by using the new framework

This section will explain how the five IoT datasets covered in section 3 (Edge-IIoTSet, X-IIoTID, TON-IoT, Bot-IoT, and Aposemat IoT-23) are evaluated using the new framework. Table 3 contains the analysis findings. Due to the limitation of space, we will only cover three of the aspects that require extra explanation in the next paragraph, namely: dataset source location, type of attacks, and IoT data sources. Other aspects can be understood easily by reviewing the Table 3.

Regarding the dataset source location aspect, considering that the definition we have given for this aspect is the official website location owned by the dataset creator, it follows that 2 IoT datasets do not have an official website location from which users can download their datasets, however, both can be downloaded from other websites, the two datasets are Edge-IIoTSet and X-IIoTID dataset. The other 3 datasets have official website locations, namely TON-IoT, Bot-IoT, and Aposemat IoT-23 dataset. The location for downloading the dataset can be seen in Table 4.

Regarding the IoT datasources aspect, each dataset uses various numbers of IoT data sources, and the IoT device types also vary. For this element, we first examine the dataset to check if it contains any information regarding IoT devices; if not, we then examine the creator's documentation. The results of the dataset analysis demonstrate that only two datasets (Edge-IIoTSet and TON-IoT) explicitly indicate the IoT devices that were used; the other three datasets (X-IIoTID, Bot-IoT, and Aposemat IoT-23) do not directly indicate the type of IoT devices so it must be referred from the documentation. Table 5 shows the list of IoT devices in each dataset.

Table 2. Comparison of the aspects of the novel framework to the existing framework

No.	Aspects of the novel framework	Al-Hawawreh's Framework	Kenyon's Framework	Ring's Framework	Gharib's Framework
1	Dataset generation time	-	Data provenance	Year of creation	-
2	Dataset metadata	Metadata	Useful metadata	Metadata	Metadata
3	Dataset source location	-	-	-	-
4	Dataset feature description	Feature set/IIoT Traces	-	Format	Feature set/complete capture
5	Open and free to the public	public availability	Ethical context	Public availability	-
6	Labeled dataset	labeled dataset	Consistent labeling	Labeled	Labeled dataset
7	Privacy and data protection	agnostic-features	De-identification context	Anonymity	Anonymity
8	Availability of raw data	-	Origin data	-	-
9	Updated dataset	-	-	-	-
10	Benign traffic	-	Representative events	Normal user behavior	Complete traffic
11	Type of attacks	-	-	-	Attack diversity
12	Balanced dataset	-	-	Balanced	-
13	Training-testing dataset splits	-	-	Predefined splits	-
14	Unique data entry	-	-	-	-
15	Traffic volume	-	-	Count	-
16	Network topology	Complete network and system configuration	Complete network and system configuration	Type of network/complete network	Complete traffic/complete network configuration/complete interaction
17	Iot datasources	Heterogeneous data sources	-	-	Heterogeneity
18	Traffic generation	Realistic network traffic/diverse data duration	Calibration details/sample duration/temporal scope	Kind of traffic/duration	Complete traffic
19	Protocol type	Iiot connectivity protocols	-	-	Available protocols

Table 3. Analysis of five IoT datasets in the proposed framework

No.	Aspects	Edge-IIoTSet	X-IIoTID	TON-IoT	Bot-IoT	Aposemat IoT-23
1	Dataset generation time	yes [11]	yes (timestamp)	yes (timestamp)	yes (timestamp)	yes (timestamp)
2	Dataset metadata	yes	yes	yes	yes	yes
3	Dataset source location	no	no	yes	yes	yes
4	Dataset feature description	yes, partial (description, format).	yes, partial (description only).	yes, partial (description, format)	yes, partial (description)	yes, partial (format)
5	Open and free to the public	yes	yes	yes	yes	yes
6	Labeled dataset	yes	yes	yes	yes	yes
7	Privacy and data protection	yes	yes	yes	yes	yes
8	Availability of raw data	no	yes	yes	yes	yes
9	Updated dataset	no	no	no	no	no
10	Benign traffic	yes	yes	yes	yes	yes
11	Type of attacks	5 categories, 14 types	9 categories, 18 types	9 types	3 categories, 6 types	15 types
12	Balanced dataset	balanced (bi-class) imbalanced (multi-class)	imbalanced (bi-class) imbalanced (multi-class)	imbalanced (bi-class) imbalanced (multi-class)	imbalanced (bi-class and multi-class)	imbalanced (bi-class and multi-class)
13	Training-testing dataset splits	no	no	yes (partial)	yes	no
14	Unique data entry	yes	yes	yes	yes	yes
15	Traffic volume	20,939,622	820,680	31,504,615	73,370,443	325,309,945
16	Network topology	yes	yes	yes	yes	no
17	Iot datasources	10 IoT devices	5 IoT devices	7 IoT devices	5 IoT devices	3 IoT devices
18	Traffic generation	simulation	simulation	simulation	simulation	simulation
19	Protocol type	arp, icmp, http, tcp, udp, mqtt	tcp, udp, icmp	tcp, udp	udp, tcp, arp, ipv6-icmp, icmp, igmp, rarp	udp, tcp, icmp

Table 4. Dataset source location

No.	Dataset	Dataset source location	Download
1	Edge-IIoTSet	No	https://www.kaggle.com/datasets/mohamedamineferrag/edgeiiotset-cyber-security-dataset-of-iiot https://iee-dataport.org/documents/edge-iiotset-new-comprehensive-realistic-cyber-security-dataset-iiot-and-iiot-applications#files
2	X-IIoTID	No	https://www.kaggle.com/datasets/munaalhawawreh/xiiotid-iiot-intrusion-dataset https://iee-dataport.org/documents/x-iiotid-connectivity-and-device-agnostic-intrusion-dataset-industrial-internet-things
3	TON-IoT	Yes	https://research.unsw.edu.au/projects/toniot-datasets
4	Bot-IoT	Yes	https://research.unsw.edu.au/projects/bot-iiot-dataset
5	Aposemat IoT-23	Yes	https://www.stratosphereips.org/blog/2020/1/22/aposemat-iiot-23-a-labeled-dataset-with-malicious-and-benign-iiot-network-traffic

Table 5. List of IoT devices in the dataset

Edge-IIoTSet	X-IIoTID	TON-IoT	Bot-IoT	Aposemat IoT-23
Distance, flame sensor, heart rate, IR receiver, modbus, pH Value, Soil moisture, sound sensor, temperature, and humidity and water level	sensors, actuators, various mobile and IT devices, access media, APIs. Does not specify the type of device.	Fridge, global positioning system (GPS) tracker, motion light, garage door, modbus, thermostat and weather	Simulated IoT services (weather station, smart fridge, Motion activated lights, remotely activated garage door, smart thermostat)	Philips HUE smart LED lamp, Amazon Echo home intelligent personal assistant dan somfy smart door lock

One of the most crucial factors to take into account when selecting an IoT dataset is the type of attacks, as it demonstrates attacks that can be identified. Based on the analysis of the dataset that has been done as shown in Table 6, it can be seen that each IoT dataset has various types of attacks. By looking at the results of dataset analysis using the new framework in this section, researchers can compare which IoT datasets are the most suitable for use in their research.

Table 6. List of attacks in the iot dataset

No.	Dataset	Attack category	Attack type
1	Edge-IIoT	Denial of service (DoS)/distributed denial of service (DDoS), Information gathering, Man in the middle, Injection, Malware	TCP SYN FloodDDoS, user datagram protocol (UDP) flood DDoS, HTTP flood DDoS, internet control message protocol (ICMP) flood DDoS, Port Scanning, OS Fingerprinting, Vulnerability scanning, Man in the middle, Man in the middle, Cross-site Scripting (XSS), SQL Injection, Uploading attack, Backdoor, Password cracking, Ransomware
2	X-IIOTD	Reconnaissance, Weaponization, Exploitation, Lateral Movement, Command & Control, Exfiltration, Tampering, Crypto Ransomware, RDoS	Generic Scanning, Scanning vulnerability, Discovering resources, Fuzzing, Brute-force, Dictionary, Insider malicious, Reverse shell, man-in-the-middle (MitM), Modbus-register reading, MQTT-cloud broker subscription, TCP Relay, Command & Control, Exfiltration, False data injection, Fake notification, Crypto Ransomware, RDoS
3	ToN-IoT	-	Backdoor, DDoS, DoS, Injection, MitM, Password, Ransomware, Scanning, XSS
4	BoT-IoT	Information Gathering, Denial of Service, Information Theft	Port Scanning, OS Fingerprinting, DDoS TCP, DDoS UDP, DDoS HTTP, DoS TCP, DoS UDP, DoS HTTP, Keylogging, Data theft
5	Aposemat IoT-23	-	C&C, C&C-FileDownload, C&C-HeartBeat, C&C-HeartBeat-Attack, C&C-HeartBeat-FileDownload, C&C-Mirai, C&C-PartOfAHorizontalPortScan, C&C-Torii, DDoS, FileDownload, Okiru, Okiru-Attack, PartOfAHorizontalPortScan, PartOfAHorizontalPortScan-Attack

5. CONCLUSION

This research initiative aims to propose a novel framework for assessing IoT datasets for ML/DL-based IDS, in order to assist researchers, select the IoT dataset that best matches their requirements. The steps taken begin with comparing the aspects of the four existing frameworks for analysis of datasets, analysis of the five IoT datasets produced between 2019 and 2022, development of a new framework specifically for the IoT dataset, and analysis of the five IoT datasets using the new framework. The proposed new framework comprises 19 aspects that must be investigated and categorized into 3 groups. It has been shown that our proposed framework is much more comprehensive than other frameworks for evaluating IoT datasets. The result of the

IoT datasets analysis shows that the new framework is able to support researchers in determining which IoT dataset is most appropriate for their research on ML/DL-based IDS. Another advantage of this new framework is that the creator of an IoT dataset can use this new framework as a reference while creating their IoT dataset.




REFERENCES

- [1] A. Khanna and S. Kaur, "Internet of Things (IoT), Applications and Challenges: A Comprehensive Review," *Wireless Personal Communications*, vol. 114, no. 2, pp. 1687–1762, Sep. 2020, doi: 10.1007/s11277-020-07446-4.
- [2] I. Idriissi, M. Boukabous, M. Azizi, O. Moussaoui, and H. El Fadili, "Toward a deep learning-based intrusion detection system for IoT against botnet attacks," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 110–120, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp110-120.
- [3] M. Arief and S. H. Supangkat, "Comparison of CNN and DNN Performance on Intrusion Detection System," in *2022 International Conference on ICT for Smart Society (ICISS)*, Aug. 2022, pp. 1–7, doi: 10.1109/ICISS55894.2022.9915157.
- [4] Y. Ayachi, Y. Mellah, M. Saber, N. Rahmoun, I. Kerrakchou, and T. Bouchentouf, "A survey and analysis of intrusion detection models based on information security and object technology-cloud intrusion dataset," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 4, pp. 1607–1614, Dec. 2022, doi: 10.11591/ijai.v11.i4.pp1607-1614.
- [5] L. Ashiku and C. Dagli, "Network Intrusion Detection System using Deep Learning," *Procedia Computer Science*, vol. 185, pp. 239–247, 2021, doi: 10.1016/j.procs.2021.05.025.
- [6] 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, Mar. 2022, doi: 10.1109/JIOT.2021.3102056.
- [7] A. Kenyon, L. Deka, and D. Elizondo, "Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets," *Computers & Security*, vol. 99, Dec. 2020, doi: 10.1016/j.cose.2020.102022.
- [8] 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, Sep. 2019, doi: 10.1016/j.cose.2019.06.005.
- [9] I. Sharafaldin, A. Gharib, A. H. Lashkari, and A. A. Ghorbani, "Towards a Reliable Intrusion Detection Benchmark Dataset," *Software Networking*, vol. 2017, no. 1, pp. 177–200, 2017, doi: 10.13052/jsn2445-9739.2017.009.
- [10] A. Gharib, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "An Evaluation Framework for Intrusion Detection Dataset," in *2016 International Conference on Information Science and Security (ICISS)*, Dec. 2016, pp. 1–6, doi: 10.1109/ICISSEC.2016.7885840.
- [11] 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. 40281–40306, 2022, doi: 10.1109/ACCESS.2022.3165809.
- [12] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems," *IEEE Access*, vol. 8, pp. 165130–165150, 2020, doi: 10.1109/ACCESS.2020.3022862.
- [13] T. M. Booi, I. Chiscop, E. Meeuwissen, N. Moustafa, and F. T. H. den Hartog, "ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 485–496, Jan. 2022, doi: 10.1109/JIOT.2021.3085194.
- [14] 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, Nov. 2019, doi: 10.1016/j.future.2019.05.041.
- [15] A. Parmisano, S. Garcia, and M. J. Erquiaga, "IoT-23: A labeled dataset with malicious and benign IoT network traffic," *Zenodo*, vol. 31, 2020.
- [16] D. U. Board, "DCMI Metadata Terms," *DublinCore*. 2020. Accessed: Sep. 08, 2023. [Online]. Available: <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>
- [17] M. Newcomer *et al.*, "Open and Free Datasets for Hydrology Research: Insights, Challenges and Opportunities," *IAHS-AISH Scientific Assembly 2022*, 2022, doi: 10.5194/iahs2022-310.
- [18] Amarudin, R. Ferdiana, and Widyawan, "A Systematic Literature Review of Intrusion Detection System for Network Security: Research Trends, Datasets and Methods," in *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, Nov. 2020, pp. 1–6, doi: 10.1109/ICICoS51170.2020.9299068.
- [19] R. Ishibashi, K. Miyamoto, C. Han, T. Ban, T. Takahashi, and J. Takeuchi, "Generating Labeled Training Datasets Towards Unified Network Intrusion Detection Systems," *IEEE Access*, vol. 10, pp. 53972–53986, 2022, doi: 10.1109/ACCESS.2022.3176098.
- [20] Q. Li *et al.*, "A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3347–3366, Apr. 2023, doi: 10.1109/TKDE.2021.3124599.
- [21] M. Kuznetsov, E. Novikova, I. Kotenko, and E. Doynikova, "Privacy Policies of IoT Devices: Collection and Analysis," *Sensors*, vol. 22, no. 5, Feb. 2022, doi: 10.3390/s22051838.
- [22] Z. Wu, H. Zhang, P. Wang, and Z. Sun, "RTIDS: A Robust Transformer-Based Approach for Intrusion Detection System," *IEEE Access*, vol. 10, pp. 64375–64387, 2022, doi: 10.1109/ACCESS.2022.3182333.
- [23] Y. A. Farrukh, I. Khan, S. Wali, D. Bierbrauer, J. A. Pavlik, and N. D. Bastian, "Payload-Byte: A Tool for Extracting and Labeling Packet Capture Files of Modern Network Intrusion Detection Datasets," in *2022 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, Dec. 2022, pp. 58–67, doi: 10.1109/BDCAT56447.2022.00015.
- [24] R. Lohiya and A. Thakkar, "Application Domains, Evaluation Data Sets, and Research Challenges of IoT: A Systematic Review," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8774–8798, Jun. 2021, doi: 10.1109/JIOT.2020.3048439.
- [25] S. M. Tahsien, H. Karimipour, and P. Spachos, "Machine learning based solutions for security of Internet of Things (IoT): A survey," *Journal of Network and Computer Applications*, vol. 161, Jul. 2020, doi: 10.1016/j.jnca.2020.102630.
- [26] B. Kaur *et al.*, "Internet of Things (IoT) security dataset evolution: Challenges and future directions," *Internet of Things*, vol. 22, Jul. 2023, doi: 10.1016/j.iot.2023.100780.
- [27] Q. Li, C. Zhao, X. He, K. Chen, and R. Wang, "The Impact of Partial Balance of Imbalanced Dataset on Classification Performance," *Electronics*, vol. 11, no. 9, Apr. 2022, doi: 10.3390/electronics11091322.
- [28] Y. Xu and R. Goodacre, "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning," *Journal of Analysis and Testing*, vol.




- 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2.
- [29] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, “Deep Learning Approach for Intelligent Intrusion Detection System,” *IEEE Access*, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.
- [30] X. S.de-Camara, J. L. Flores, C. Arellano, A. Urbieto, and U. Zurutuza, “Gotham Testbed: A Reproducible IoT Testbed for Security Experiments and Dataset Generation,” *IEEE Transactions on Dependable and Secure Computing*, pp. 1–18, 2023, doi: 10.1109/TDSC.2023.3247166.
- [31] R. Al-amri, R. K. Murugesan, M. Man, A. F. Abdulateef, M. A. Al-Sharafi, and A. A. Alkahtani, “A Review of Machine Learning and Deep Learning Techniques for Anomaly Detection in IoT Data,” *Applied Sciences*, vol. 11, no. 12, Jun. 2021, doi: 10.3390/app11125320.
- [32] 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, Apr. 2021, doi: 10.1109/JIOT.2020.3032093.
- [33] C. Patel and N. Doshi, “A Novel MQTT Security framework In Generic IoT Model,” *Procedia Computer Science*, vol. 171, pp. 1399–1408, 2020, doi: 10.1016/j.procs.2020.04.150.

BIOGRAPHIES OF AUTHORS






Muhammad Arief    holds Erasmus Mundus Master of Science in Network and e-Business Centered Computing (EM MSc. NeBCC) degree from University of Reading, UK, Universidad Carlos III de Madrid, Spain, Aristotle University of Thessaloniki, Greece, in 2007. He also received his Master of Science in Electrical Engineering degree from Delft University of Technology, Holland, in 1994. He is currently pursuing his study into PhD in Bandung Institute of Technology, Indonesia. He is also a senior researcher in National Research and Innovation Agency, Indonesia. His research interests are in cyber security, intrusion detection system, machine learning, deep learning. He can be contacted at email: 33221045@std.stei.itb.ac.id and muha032@brin.go.id.






Made Gunawan    holds Master of Engineering degree in Electrical Engineering from Auckland University, New Zealand, in 2000. He also received a Master of Science in Electrical Engineering degree from Delft University of Technology, Holland, in 1991. He is also a researcher in National Research and Innovation Agency, Indonesia. His research interests are in data science, machine learning, and deep learning. He can be contacted at email: made001@brin.go.id.






Agung Septiadi    received the M.Sc. degree in Electrical Engineering from the Korea Advanced Institute of Science and Technology, Korea, in 2018. and bachelor degree in Electrical Engineering from Bandung Institute of Technology, Indonesia, in 2008. He is senior researcher in National Research and Innovation Agency, Indonesia. His research interests include intrusion detection system, cyber security, machine learning. His email is agun021@brin.go.id.






Mukti Wibowo    received the bachelor's degree in informatics engineering from the Islamic State University Syarif Hidayatullah Jakarta, Indonesia, in 2019. He is first expert engineer at the National Research and Innovation Agency, Indonesia. His research interests are data science, machine learning, and deep learning. He can be contacted via email: mukt003@brin.go.id.






Vitria Pragesjvara    received bachelor degree in Electrical Engineering from The Hague University of Applied Sciences, Holland, in 1992. She is also a researcher in National Research and Innovation Agency, Indonesia. Her research interests are in data science, machine learning, and deep learning. She can be contacted at email: vitr001@brin.go.id.






Kusnanda Supriatna    received a bachelor degree in Computer Science from Universite of Nantes, France in 1993, He is currently as data analyst in National Research and Innovation Agency, Indonesia. His research interests are in data science, machine learning and deep learning. He can be contacted at kusun001@brin.go.id.






Anto Satriyo Nugroho    received the B.Eng, M.Eng, and Dr.Eng degrees in Electrical and Computer Engineering from Nagoya Institute of Technology, Japan, in 1995, 2000, and 2003, respectively. He is currently the Head of the Research Center for Artificial Intelligence and Cyber Security of the National Research and Innovation Agency, Indonesia. From 2017-2022 he also served as the President of the Indonesian Association for Pattern Recognition and has become a Governing Board Member of IAPR, representing Indonesia. From 2003-2007, he served as a Visiting Professor at Chukyo University, Japan. His research interests include pattern recognition, image processing, and biometrics. He can be contacted at email: anto006@brin.go.id.



I Gusti Bagus Baskara Nugraha    holds a Ph.D. degree from The University Of Electro Communications, Japan, in 2006, Master of Science degree from Bandung Institute of Technology, Indonesia in 2001 and Bachelor of Science degree from Bandung Institute of Technology, Indonesia in 1999. He is currently with the School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia. He is also with Smart City and Community Innovation Centre. His research interests are in information systems and networks. He can be contacted at email: baskara@stei.itb.ac.id.



Suhono Harso Supangkat    holds a Doctor degree from The University of Tokyo, Japan in 1998, Master of Science degree from Meisei University, Japan in 1994 and bachelor degree in 1986 from Bandung Institute of Technology, Indonesia. He is currently a professor in the School of Electrical Engineering and Informatics at Bandung Institute of Technology (ITB), Bandung, Indonesia. He formerly served as Head of the Smart City and Community Innovation Centre. His research interests include smart X concepts, smart city governance and the Internet of Things. He can be contacted at email: suhono@stei.itb.ac.id.