NTU Journal of Engineering and Technology (2025) 4 (2) : 1-10
DOI: https://doi.org/10.56286/ntujet.v4i2

IRAQI Academic Scientific Journals

P-ISSN: 2788-9971 E-ISSN: 2788-998X

NTU Journal of Engineering and Technology

Available online at: https://journals.ntu.edu.iq/index.php/NTU-JET/index

A Review of Existence Intrusions Detection-Based Machine Learning Datasets of Future Generation Networks

Nora Rashid Najem[1], Razan Abdulhammed[2]
[1]Department of Computer computer science, Alnoor University ,41012, Iraq.
[2]Technical Engineering College for Computer and AI., Northern Technical University, Mosul- Iraq.
nora.rashid@alnoor.edu.iq, rabdulhammed@ntu.edu.iq

Article Informations

Received: 01-10- 2022,
Revised: 12-12- 2022,
Accepted: 18-01-2023,
Published online: 23-06-2025

Corresponding author:
Name: Razan Abdulhammed
Affiliation: Northern Technical University, Iraq
Email: rabdulhammed@ntu.edu.iq

Key Words:
IDS,
machine learning future,
generation networks,
cyber-attacks,
challenges of dataset.

ABSTRACT

Innovative technologies of future generation networks such as Cyber-Physical System (CPS), Mobile Ad Hoc Network (MANET), Vehicular Ad-Hoc Network (VANET), Internet of Things (IOT), and Wireless network commonly known as Wi-Fi have emerged, which require a distinguished understanding of the main challenges and constraints that face the design and implementation of an Intrusion Detection Systems (IDS) for such type of networks. Moreover, a dramatic increase in the rate of cyber-attacks has increased, and new cases of intrusions, bugs, novel attacking tactics, and vulnerabilities are evolving daily. Intrusion Detection Systems (IDS) are one of the solutions against these attacks. Thus, IDS needs to improve its performance in terms of its ability to detect new attacks and respond to threats. Getting suitable datasets for evaluating various research designs in IDS design domains is a significant challenge . The machine learning (ML) design approach can quickly identify trends and patterns of intrusions, bugs, tactics, and cyber vulnerabilities with minimum human intervention. This paper reviews datasets for the research community. Furthermore, it explores the challenges of Dataset for intrusion detection based on Machine learning. It glances through a period of 6 years of intrusion detection datasets, explores what is currently applicable, outlines criteria for selecting the best Dataset, and explores future directions for creating relevant datasets.

1. Introduction

A dataset collects correlating information with a mutual relationship or connection. One piece of information affects or depends on another part of the information belonging to the same Dataset. A common approach to represent a dataset is the tabular paten. In this approach, every column of a table represents a particular variable in the Dataset, and each row (set of variables) represents a specific record of the Dataset[1]. Machine learning algorithms are used to extract and discover the correlated information from a given dataset by treating each pace of data as a single unit running on a single or multiple computer/processors by ML[1]. Generally, and from this review perspective, in research and as shown in Figure 1: datasets fall into several categories such as baseline data, simulations, Synthetic, traffic generation, and live network/ Realistic[2]. A Baseline dataset is a data set frequently used to evaluate other data obtained later. Simulated data is a form of data in which the owner uses a simulation program such as Ns3, Riverbed OPNET, QualNet, GNS3, and so on to generate normal and malicious behavior that follows the set of predetermined rules and objectives of the dataset owner[3]. In traffic generation categories, the Dataset is obtained with a network traffic generator tool or network traffic simulator to mimic actual network traffic. A live network dataset is obtained from an entire network consisting of devices, users, servers, and attackers. In network security domains, getting suitable datasets for evaluating and designing an intrusion detection system based on machine learning approaches is a significant challenge for researchers, developers, and dat a donors [4]. These challenges are highlighted in Table 1[5]. The article presents a simple and brief explanation of each challenge highlighted in Table 1. Please refer to reference[5] for readers interested in in-depth details. The imbalance ratio (I.R.) is the ratio of the number of instances in the majority class or negative class to the number of instances in the minority class or positive class [6]. For this review, the positive class is the attack class, whereas the negative class is the normal class. the reviewed studies such as the applied algorithms, the size and type of the used Dataset for training and testing, classification output classes as binary or multi-class, and imbalance ratio Table 2: is the summary of research articles included in the review process for this article. Information extracted and summarized in this table are datasets, published year, as well as ML approach, adopted these datasets to design an IDS.[1],[7].

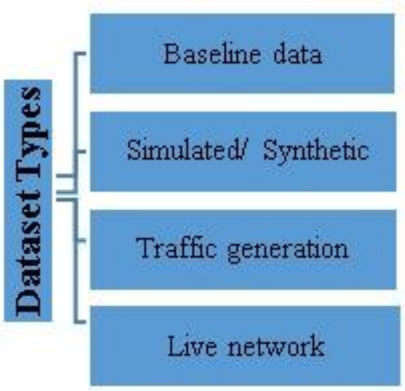

Fig. 1. The dataset types in Machine Learning.

This paragraph summarize the finding shown in the table 2.The reader can notice from the table 2 that ML approaches are widely used due to their ability to detect any threats. vulnerabilities accurately, quickly and quakily [1],[6]. Furthermore, these studies relied solely on labeled datasets. The review found that developers prefer labeled Dataset over unlabeled Dataset. Also, Decision Tree, Neural Network, SVM, and K mean ML approaches are the most broadly used between (2016-2021) period. The paper presents a pie graph showing the distribution of the applied ML approaches in Figures 2 (a and b). It shows the frequency (Figure 2 a) and relative frequency (Figure 2 b) of values in the data. Frequency is the number of times that value appeared in the data, and relative frequency is the percentage of the total. Per the selected articles from 2016 to 2021, the review found that 31% of the latest published research utilized the Decision Tree Algorithm. Conversely, Regression approaches were adopted in only 1% of published research. For provide more details related to Dataset and approaches, Figure 3: illustrates how the researchers have adopted the datasets. In summary, developers utilized each of CICIDS2017, N-BaIoT, and IoTID20, which are seven times, making 12% of the total Dataset's usage. Next, developers used the BoT_IoT and CSE-CICIDS2018 Dataset six times. At the same time, developers used IoT-23, DS2OS, and CICDDoS2019 Dataset five times. Authors adopt The ToN-IoT and MQTT-IOT-IDS2020 three times. The designer used CIRA-CIC-DoHBrw-2020 four times. Also, the creator of IDS used both AWID and MedBIoT once.

Table 1. List of recent research in IDS from 2016 to 2021

Challenges	Meaning
Approval Forms of the Dataset	Some dataset donors require approval to obtain access to the Dataset, and this approval process is frequently delayed.
Privacy of the Dataset	Realistic data are not allowed to be shared among users due to security policies, sensitivities of actual data, lack of trust, and risk of disclosing digital information.
Labeling of the Dataset	some available datasets are manually labeled datasets, while some are packet traces without identifiers, which influences the validity of the datasets.
Availability of the Dataset	Indicate the data is available to developers and researchers when and where they need it.
Objectives of the Dataset	It Refers to rules and goals, choice of attack type, target protocol environment, and categories of Dataset.
Scope of the Dataset	it specifies a set of report data, and most publicly available datasets become outdated and unsuitable for making strong scientific claims because of variability in network segments.
Documentations of the Dataset	Information related to Dataset such as attack type, Operating system, number of machines, features, dataset collection environment
Scenario of dataset collection gather	Most data donors do not publish the intruders' success level in the datasets. Thus, a high level of expertise is often required to understand these categories of attacks present in the same Dataset.
Imbalance Ratio of the Dataset	A factor that estimates the ratio of the number of normal class instances to the number of attack class instances in a dataset

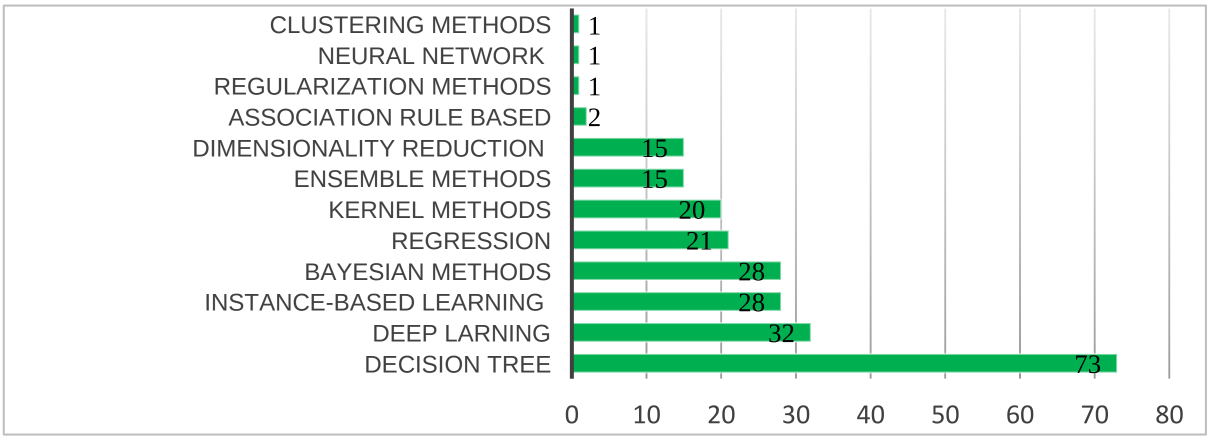

Fig. 2a. Methods used in IDS research from 2016-2021.

Fig. 2b. Methods used in IDS research from 2016-2021.

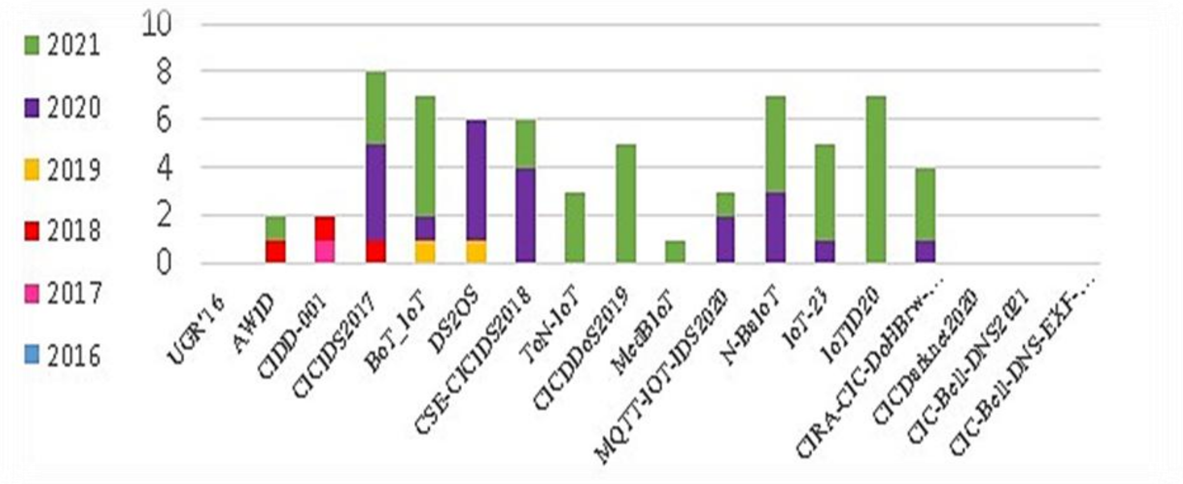

Table 2. List of recent research in IDS from 2016 to 2021

Rfe	Year	Dataset	Method	Rfe	Year	Dataset	Method
[8]	2021	IoT-23	NB, SVM, LR, DT, RF	[37]	2020	DS2OS	KNN, LDA, DT, RF, LR, SVM, ANN, AdaBoost
[9]	2021	CICDDoS2019	Stacked Auto Encoder-CNN	[38]	2021	IoTID20	MLP, ET, k-NN, DT, RF, Bagging, Ada boost, GBM, XGBoost
[10]	2021	MQTT-IoT-IDS2020	DNN, NB, RF, kNN, DT, LSTM, GRUs	[39]	2020	CSE-CIC-IDS2018	KNN, RF, Gradient Boosting, Adaboost, DT, LDA
[11]	2021	ToN-IoT, BoT-IoT	NB, KNN, SVM	[40]	2021	CIDDS-001	RF, MLP, LSTM
[12]	2020	AWID	Bagging, Extra Trees, XGBoost, RF, NB	[41]	2020	N BaIoT	NB, KNN, LR, DT, RF, CNN, RNN, LSTM
[13]	2021	CICDDoS2019	DT, KNN, NB	[42]	2020	BoT-IoT	RF, CNN, MLP
[14]	2021	IoT-23	RF, SVM, KNN	[43]	2021	N BaIoT	LR, ANN
[15]	2021	IoTID20	LSTM, SLFN, NB, KNN	[44]	2021	IoTID20	CNN, LSTM, CNN-LSTM
[16]	2020	DS2OS	SVM, DT, ANN, RaNN	[45]	2021	ToN-IoT	LR, NB, D.T., R.F., ADABOOST, KNN, SVM, XGBOOST
[17]	2020	CICIDS2017	Fast kNN classifier (FkNN)	[46]	2021	BoT-IoT	deep autoencoder
[18]	2021	CICDDoS2019	AE-MLP	[47]	2021	Bot-IoT IoTID20	LR, SVM, DT., ANN
[19]	2021	CIRA-CIC-DoHBrw2020	RF, DT, GNB, KNN, LR,SVC, QDA, SGD	[48]	2021	IoTID20	SVC, XGBoost, Random Forest
[20]	2020	DS2OS	LR, SVM, DT, RF, ANN	[49]	2020	DS2OS	LR, ANN
[21]	2020	CSE-CIC-IDS2018	KNN, NB, Adaboost- DT, SVM, RF, MLP	[50]	2020	MQTT-IoT-IDS2020	LR, k-NN, DT, RF, SVM (RBF Kernel), NB, SVM (Linear Kernel)
[22]	2020	CSE-CIC-IDS2018	DT+RF+NB+SVM+KNN	[51]	2020	MQTT-IoT-IDS2020	Neural network, RF, NB, DT, Gradient boost, MLP
[23]	2021	CICIDS2017	ANN, DT, k-NN, NB, RF, SVM, CNN, EM, k-means, SOM	[52]	2020	CSE-CIC-IDS2018	DNN
[24]	2020	N BaIoT	Naïve Bayes, CART	[53]	2020	CICIDS2017	KNN, RF, AdaBoost, LR, NB LDA, QDA, MLP
[25]	2021	N BaIoT	LSTM-RNN, CNN, RNN, BiLSTM-CNN	[54]	2021	ToN-IoT	GBM, RF, NB, DNN
[26]	2021	N-BaIoT, MedBIoT, IoT-23	LR, KNN, NB, DT, RF, MLP, LSTM	[55]	2021	BoT-IoT	KNN, RF, LR
[27]	2021	IoTID20, Bot-IoT	GNB+LR+KNN+DT+Ensemble	[56]	2021	AWID	KNN, RF, LR
[28]	2020	DS2OS	DNN	[57]	2021	CSE-CIC-IDS2018	KNN, RF, LR
[29]	2021	CICIDS2017	XGB, R.F., D.T., GBM	[58]	2020	CICIDS2017	RF, BN, RT, NB, J48
[30]	2021	CICDDoS2019	RF, Light Gradient Boosting, CatBoost, CNN	[59]	2021	BoT_IoT	RF, NB, J48, REPTREE, BNET, ONER
[31]	2020	N BaIoT	CNN, RNN-LSTM	[60]	2019	DS2OS	LR, SVM, DT, RF, ANN
[32]	2020	CIRA-CIC-DoHBrw2020	DT, KNN, RF, Extra Tree, GB, LGBM, Kernel SVM, LR, SVM, ANN	[61]	2019	BoT-IoT	FNN
[33]	2021	CIRA-CIC-DoHBrw2020	NB, LR, RF, KNN, GBM	[62]	2017	CIDDS-001	DNN
[34]	2021	CSE-CIC-IDS2018	DT, RF, CatBoost, LGB, XGB, NB, LR	[63]	2018	AWID	Autoencoder
[35]	2020	CICIDS2017	RF, IBK, JRip, MLP, NB, OneR	[64]	2018	CIDDS-001	LSTM
[36]	2021	CICDDoS2019	CNN LSTM, BLSTM, SLSTM GRT	[65]	2018	CICIDS2017	KNN, RF, MLP, ID3, NB, QDA, ADA

Finally: no usage of the CICDarknet2020, CIC-Bell-DNS2021and CIC-Bell-DNS-EXF-2021 dataset was noticed in this period (2016-2021).

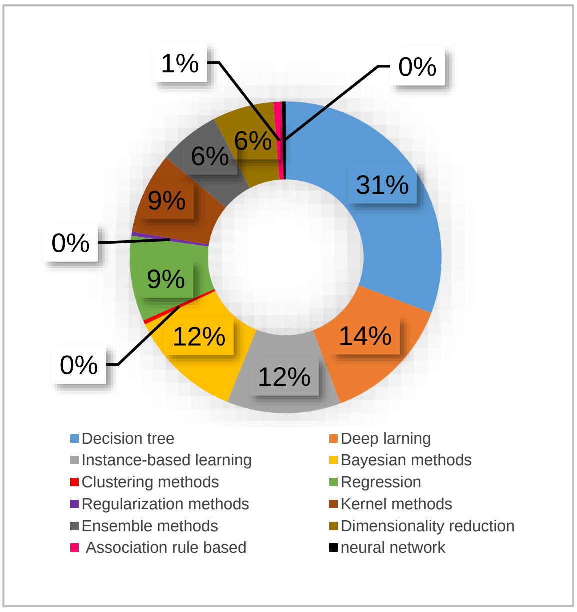

Fig. 3. Year-wise distribution of adopted datasets by the researchers 2016-2021.

3. Aproaches based on machine learning

In the ML field, and as shown in figure 4, Dataset used to build models is available in two forms: One Single Dataset File (1-D-F) (Single file) and Two Dataset Files (2-D-F) (Two Files). In 1- D-F, some of the developers divide the original subset into training, validation, and test subsets. In contrast, others split the initial subset into train and test subsets. In the case of (2-D-F) (Two Files), The developers can either break the training subset into train and validation subset or use the train as it is without further splitting it into train and validation subset. This review recommends using three subsets for neural network model types. At the same time, the two subsets are peripheral to other kinds of models. Nonetheless, developers may use the three subsets for models not based on neural network approaches. In Two Dataset Form (2-D-F), the original Dataset comes into two subsets of data, the training subset, and the test subset, so there is no need to partition the Dataset unless they wish to use part of the Dataset for validation purposes. All these subsets must randomly sample a larger body of the data and should be made uniform and understandable for a machine learning algorithm. In general, the results size of the training subset must be the largest among the three subsets. Moreover, developers recommended using a 70%: 20%:10%ratio for training, test, and validation subsets Was add, This is consider as the rule of thumb in machine learning based designs. The majority of developers follow this rule This is consider as the rule of thumb in machine learning based designs.The majority of developers follow this rule. This review and IDS-based ML designs recommend utilizing three Dataset Files (3-D-F) to partition the Dataset by either developers or donors. In (3-D-F), the developer can use Two subsets of data to train the model: the original subset and the balanced subset. The initial subset represents the original Dataset without any processes applied by a developer to change it. In contrast, the balanced subset represents the original Dataset after repairing the imbalance issue in the original Dataset through data balancing techniques. The third subset is the test subset and should be used to test the model [1],[6]. Figure 5 illustrates the three forms: One Dataset Form, Two Dataset Form, and Three Dataset Form (suggested by this review). In general, machine learning design developers follow procedures explained in figure 3 to create the final accepted model that meets the expectations or needs of the design requirements[1],[6]. Moreover, it is vital to building the models using a representative dataset to design an efficient machine learning-based IDS. A representative dataset is a particular subset of the original Dataset. It stimulates the training subset, which has two main characteristics: It is significantly smaller in size compared to the original Dataset, and it captures the most information from the original Dataset compared to other subsets of the same size. Existing intrusion detection datasets have several flaws. For example, old, they lack current attack trends or contain obsolete network traffic patterns. Second, many datasets are not publicly available and repeatable due to copyright and privacy concerns. Furthermore, datasets are usually tailored to specific scenarios (DDoS attacks in backbone networks) and frequently do not include attack labels A distributed denial-of-service(DDoS) attacks is a malicious effort to interrupt a server, service, or network's regular traffic by flooding it withInternet traffic[1],[6].

Some of the researchers use publicly available datasets to design IDS. Other researchers use simulation software to simulate and record network traffic of both normal and attack behavior. While the rest maychange well-known datasets to add synthetically manufactured attacks well-known datasets such that to evaluate and compare different IDSs.For decades, IDS developers have used various approaches to build IDS. One such approach is machine learning (ML) .Attacks are becoming more numerous and sophisticated. intrusion detection systems (IDSs) are robust. However, evaluating the performance, detection accuracy, and false positive rate of intrusion detection system is critical for their development. In addition to latency for some networking environments such as industrial cypher physical systems. The design of an ML-based IDS should consider these.

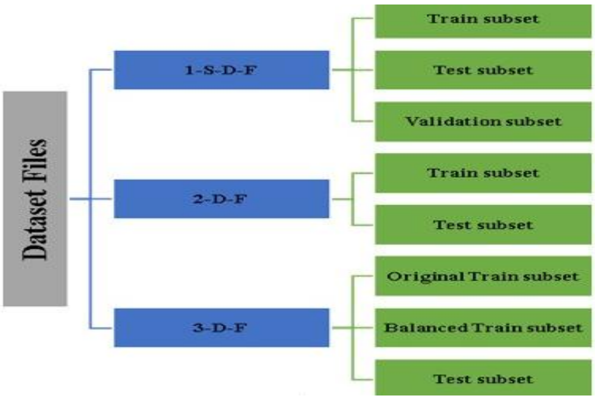

Fig. 4. The dataset in Machine Learning.

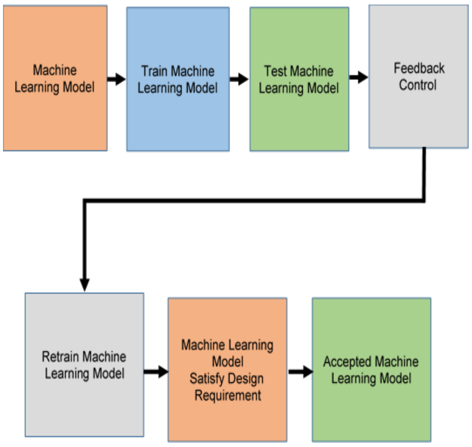
Machine Learning Model
Train Machine Learning Model
Test Machine Learning Model
Feedback Control
Retrain Machine Learning Model
Machine Learning Model Satisfy Design Requirement
Accepted Machine Learning Model

Fig. 5. Most commonly procedures followed by machine learning design developers.

4. Dataset Selection Criteria

In this section[67], the paper demonstrates the selection process criteria that the review follows to either select or exclude a particular dataset in the under review article and the review process. The paper present some essential measures to ensure review research of interest only. Firstly, the article must be published in 2016 and up to 2021 to ensure get only the most recent research so that our study is relevant and not outdated. Secondly, the article was published in a scientific journal or conference to ensure the validity of the content, which has been peer-reviewed and approved. Thirdly, the paper must use ML for IDS since this study's objective, so this paper must work within the scope of our research in this subsection, this paper present datasets used for designing IDSs by developers, and include them in this review since they pass the selection criteria process, as shown in figure 6.

Table 3: Summarizes 18 existing datasets, their release year, and Dataset's donor through a period that spans from 2016 into 2021.These datasets namely: AWID[68], CIDD-001[69], UGR'[70], CICIDS2017[65], CSE-CICIDS2018[65], CICDDoS2019 [71], N-BaIoT[72], IoT-23[73],BoT_IoT[74], IoTID20[75], ToN-IoT[76], CIRA-CIC-DoHBrw2020[77], S2OS[78], MedBIoT[79], CICDarknet2020 [80], CIC-Bell-DNS2021[81], CIC-Bell-DNS-EXF-2021[82] and MQTT-IoT-IDS2020[83]. As we can see in Table 2 that the Canadian Institute for Cybersecurity is the largest donor of the Dataset, with five datasets. Furthermore, more donors are taking their roles to provide datasets for researchers and helping to reduce the problem of lack of Dataset through providing publicly free intrusions detection datasets. In addition, it is good to notice that Universities are the leading in this donation in this perio.Conversely,federalandmilitary organiz ations have been leading in this field for the previous period (from 1999 to 2010) Figure 7 illustrates the Year-wise distribution of the datasets on the subject area of intrusion detection-based machine learning. As heightened in Figure 7, In 2020, there were five publicly available datasets,and an increase in the number of sophisticated IOTs IDS Dataset is notable starting from 2019.

5. Relevant Work and Further Reading

This section takes care of a review related to intrusion detection system Dataset in IoTs and compares specific previous reviews with our review as shown in Table 4.

6. Description of Datasets of IoT and TICS Environments

This section first displays the review's included datasets and provides a simple description of each of them. Then, in table 5: the number of features, class type (binary or multi), and the environment (IoT, TICS, WiFi). The paper describes CICIDS2017, N-BaIoT, and IoTID20 primarily.

CIC-IDS-2017[65],was created within an emulated environment over five days and contains network traffic in packet-based and bidirectional flow-based formats. The authors extracted more than 80 attributes.

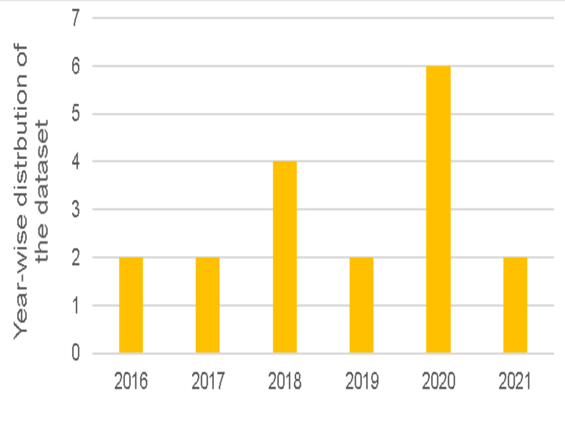

Fig. 7. Year-wise distribution of the Dataset on the subject area IDS

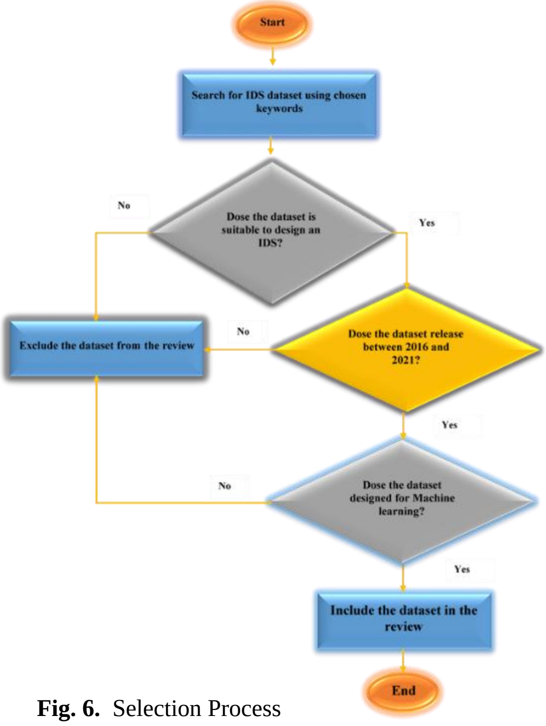
Start
Search for IDS dataset using chosen keywords
Dose the dataset is suitable to design an IDS?
No
Yes
Exclude the dataset from the review
No
Dose the dataset release between 2016 and 2021?
Yes
No
Dose the dataset designed for Machine learning?
Yes
Include the dataset in the review
End

Fig. 6. Selection Process Criteria Flowchart

Dataset is the University of New Brunswick (Canadian Institute for Cybersecurity). The Imbalance ratio of this Dataset is 471453:2359087, and the Dataset is of live network categorized. The CIDS 2017 was a set of CSV and PCAP files.

Next is N-BaIoT [72], which was collected from a simulated IoT environment to capture several benign and botnet events. Some IoT devices are six linked with WiFi, several access points, and a wire connected to a switch and router. Network traffic was collected from a small-scale network using the Wireshark that drops many packets in high-bandwidth networks. The main constraints of this Dataset are that it does not include telemetry data of IoT sensors to determine the efficiency of new federated security solutions and does not include data traces of operating systems. This Dataset is simulated categories. The Imbalance ratio of this Dataset is 6506674:555932. The N-BaIoT was a set of CSV files as well as PCAP files.

The second position utilized dataset group Bot-IoT [74] created at the University of New South Wales Canberra for evaluating intrusion detection and network forensics systems. It has various botnet and malware events and large-scale raw packets collected from different virtual machines. It has several IoT systems for benign operations with various data features. Nonetheless, the Dataset does not have hacking vectors against IoT systems and does not include audit traces of operating systems. This Dataset is traffic-generated categories. The Bot-IoT was a set of CSV files as well as PCAP files.

CSE-CIC-IDS2018 [65], the Dataset was captured for ten days. The donors extracted more than 78 attributes of 83 features. Some features include flow duration, minimum/maximum packet size, and destination port. The full dataset has 13,484,708 benign flows and 2,748,235 attack flows, that is, 16,232,943 flows in total. The data set is publicly available1 and contains brute force, DoS, web, port scan, Infiltration, Botnet, and DDoS attacks. The donor of this Dataset is the University of New Brunswick (Canadian Institute for Cybersecurity). The Imbalance ratio of this Dataset is 2748235:13484708, and the Dataset is of live network categorized. This Dataset has similar properties to CICIDS 2017 that make it suitable to compare both. The CSE-CIC-IDS2018 was a set of CSV files as well as PCAP files.

The Avast AIC laboratory created IoT-23 [73]. The Dataset contains 20 malware captures from various IoT devices and three captures for benign anomalies. A partnership with the Czech Technical University in Prague helped to collect IoT-23. Furthermore, it is suitable for malware detection, and our review recommends reference engineering methods for this Dataset. Also, it is compatible with its environment. The IoT-23 was a set of CSV file well as PCAP files.

CIDDS-001[69], The CIDDS-001 Dataset incorporates four weeks of unidirectional flow network traffic. As a unique feature, the Dataset encompasses an external server in the cloud computing environment. The CIDDS-001 data set is encompases SSH brute force, DoS and port scan attacks, and several attacks captured from the wild. Moreover, the donor of Dataset updated it in two versions. This Dataset is a benefit to building a specified IDS for brute force, port scanning attack and traffic analysis.

DS2OS[78], by Pahl and Aubet is very helpful in evaluating the effectiveness of ML as well as DL-based algorithms not only in smart cities but also in intelligent factory architectures. It is open-source and introduced concerning new generation IoT security environments. The Dataset contains a total of 357952 samples with 10017 anomalous and 347935 benign values. It comprises 13 features and seven types of attacks: denial of service, malicious behavior, wrong setup, spying, scan, and data type probing attacks. The Imbalance ratio of this Dataset is 1431:49705, and it is of traffic generation categories. The DS2OS was a set of CSV files as well as PCAP files.

CICDDoS2019 [71] Dataset has only DDOS attack instances and encompasses 87 features. Twelve DDoS attacks on training day and seven attacks on testing day.The training sample contained a malicious profile of MSSQL, SNMP, NTP, UDP, DNS, LDAP, NetBIOS, SSDP, UDP-Lag, WebDDoS, SYN, and TFTP. Thus, the types of DDoS attacks in the Training sample are slightly different and less in the number of the types of DDoS attacks in the Testing sample.

Table 3. Existence of publicly available Dataset from 2016 to 2021.

Issued Year	Dataset	Donor
2016	AWID	University of AEGEAN
2016	UGR'16	University of De Granda
2017	CIDD-001	Coburg University of Applied Sciences
2017	CICIDS2017	Canadian Institute for Cybersecurity
2018	BoT_IoT	University of New South Wales Canberra
2018	DS2OS	Francois Xavier Aubet
2018	CSE-CICIDS2018	Canadian Institute for Cybersecurity
2018	N-BaIoT	Singapore University of Technology and Design and Ben-Gurion University of the
2019	ToN-IoT	University of New South Wales Canberra
2019	CICDDoS2019	Canadian Institute for Cybersecurity
2020	MedBIoT	Tallinn University of Technology; Estonia
2020	MQTT-IOT-IDS2020	IEEE Data Port
2020	IoT-23	Stratosphere Laboratory CTU University
2020	IoTID20	King Faisal University
2020	CIRA-CIC-DoHBrw-	Canadian Institute for Cybersecurity
2020	CICDarknet2020	Canadian Institute for Cybersecurity
2021	CIC-Bell-DNS2021	Canadian Institute for Cybersecurity
2021	CIC-Bell-DNS-EXF-	Canadian Institute for Cybersecurity

Table 4. Comparison of this Review and similar Review.

Ref	Purpose of the review	The years the papers were published	summarize popular benchmark datasets	IOT Environment
[84]	Machine Learning Techniques	2015-2020	no	no
[85]	Machine Learning Techniques	2015-2018	no	no
[86]	techniques, datasets, and methods used on IDS	2016-2020	no	no
[87]	intrusion detection techniques and Datasets	__________	no	no
[88]	ML and D.M. Techniques used for IDS	__________	no	no
Our Review	Review for Datasets of future generation Networks	2016- 2021	yes	yes

Therefore, the testing sample can serve two purposes. The first is a sample to evaluate the proposed model, and the second has a zero-day attack as it was missing from the training sample. The authors executed the PortScan attack on testing day. Thus, the PortScan attack is unknown when evaluating the proposed model by developers who adopt CICDDoS2019. In comparison, the Testing sample contained a malicious profile of PortScan, NetBIOS, LDAP, MSSQL, UDP, UDPLag, and SYN. The Imbalance ratio of this Dataset is 50006249:56863, and the Dataset is of live network categorize. This Dataset has only DDos attacks type, and thus it makes an excellent choice to develop an IDS for IoTs since DOS and DDoS attacks. The most dangerous attacks in the IoTs environment now and in the future are DoS and DDoS. The CICDDoS2019 was a set of CSV files as well as PCAP files. CIC-DoHBrw-2020 [77] dataset captures benign and malicious DoH traffic along with CIRA- non-DoH traffic. To obtain the representative Dataset, the authors generated The HTTPS (benign DoH and non-DoH) and DoH traffic. Benign DoH and non-DoH are achieved by accessing the top 10,000 Alexa websites and using browsers with DNS tunneling that support DoH protocol for the browsers. The Dataset has a total of 34 features and four classes. The number of samples in this Dataset is around 1.4 million. During the data collection phase, donors ignored too-small packets for dimensionality reduction. This Dataset is suitable for studying and developing a firewall against intrusions of domain name system servers. this Dataset is of real network traffic categorized. However, the concern is that the level of success of the intruders in the datasets is some who are not sharply clean. Thus, a more advanced ML approach is required to extract effective patterns and features of anomalies behavior. The CIC-DoHBrw-2020 was a set of CSV files as well as PCAP files.

TON_IoT [76] addressed the main limitations of existing datasets. The designer uses a novel orchestrated architecture to demonstrate edge, fog, and cloud layers' interconnections. The donors dynamically deployed these interactions using SDN, NVF, and service orchestration technologies. The datasets have four heterogeneous data sources and concurrent collections of legitimate and attack events traces of operating systems related to IoT and IIoT and network systems service. TON_IoT has 796,380 benign flows and 21,542,641 attack flows and 22,339,021 flows in total. The Imbalance ratio of this Dataset is 796380:21542641. TON_IoT fits the requirements needed to design an Intrusion detection system for cloud computing, IoTs environments, and Fog computing. The TON_IoT was a set of CSV files as well as PCAP files.

MQTT-IoT-IDS2020 [83] simulates a network that utilizes MQTT protocol in an MQTT network architecture. The network comprises sensors, a broker, a simulated camera, and an attacker. At the same time, the behavior includes Benign, UDP scan, Sparta SSH, and MQTT brute-force attack. MQTT-IoT-IDS2020 is of simulated type categories. The MQTT-IoT-IDS2020 was a set of CSV files as well as PCAP files. Donors of AWID [68] create a small 802.11 network environment of 11 clients to capture WLAN traffic in packet-based format. The AWID [6] has Thirty-seven million packets and 156 attributes during one hour. The AWID includes 16 attacks. AWID is split into a training and a test subset. This Dataset's Imbalance ratio for AWID-Train is 1633190:162385 and for AWID-Test is 530785:44858, a live network category. Moreover,

The data set includes many attack types like brute force, Po

The author updated Dataset in version two (AWID2) and version 3 (AWID3) of the AWID project. All three versions were a set of CSV files.

The Tallinn University of Technology donated a medium-sized IoT botnet dataset represented by MedBIoT [79]. The testbed is a combination of actual and simulated IoT devices. The Dataset consists of three significant botnets: Mirai, Bash Lite, and Torii. The Imbalance ratio of this Dataset is 4782:994828, which is of the synthetic category.

Donors of CIC Bell DNS 2021 collected Real-time DNS-related data in [83]. CIC Bell DNS 2021 adds an extra advantage for ML developers to mine patterns of intrusions. It helps flag a request as benign, spam, phishing, or malware, and it has 32 features formed from DNS-statistical and linguistic components. CIC Bell DNS 2021 contained 400,000 benign instances and 13,011 malicious instances. The Imbalance ratio of this Dataset is 13011:400000. Also, and it is of the real live network category. The CIC Bell DNS 2021 was a set of CSV and PCAP files.

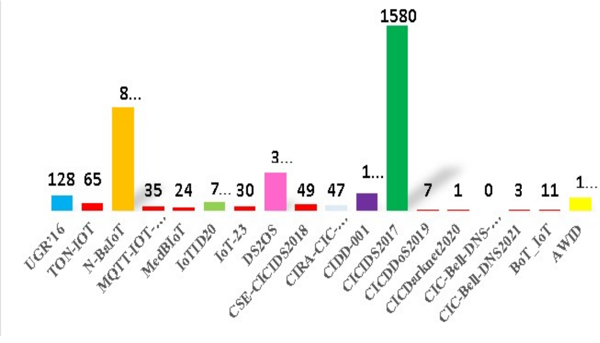

Fig. 8. Frequency distribution of datasets adopted by the researchers 2016-2021

7. Description of Datasets of Other Environments

1.WSN:

A. WSN-DS [89], For the WSN environment, the WSN-DS, a specialized dataset for WSN, was constructed to classify four types of DoS attacks. The considered attacks are Black hole, Gray hole, Flooding, and Scheduling attacks. The data were collected using NS-2 and processed to produce 23 features, in addition to including normal behavior, it was also able to collect 374661 records containing the signatures of DoS attacks. In addition, the authors provide mathematical validation of the created Dataset to ensure its correctness. The constructed Dataset is called WSN-DS. It is in a simulated category.

B. GPRS: is a dataset specific to the wireless environment(IEEE802.11)[90].The authors generated two WEP/WPA and WPA2 topologies. It has 9600 instances and 15 features from WEP/WPA. The proportion of normal and malicious classes is 62.5 and 37.5, respectively. Furthermore, the authors obtained 7500 samples and 16 features from the WPA2 topology. GPRS comprises normal class (60) and malicious class (40). It is in the simulated category.

C. LWSNDR[91], which is a labeled dataset. The data consists of humidity and temperature measurements collected from the sensor nodes in two versions, single-hop and multi-hop single, for 6 hours at intervals of 5 seconds. LWSNDR cannot accommodate IoT environments regarding security threats related to IoT.

2. SWaT:

The SWaT[92], dataset was systematically generated from the Secure Water Treatment Testbed to address this need. The authors create the testbed for 11 days of continuous operation seven days under normal scenarios and four days under attack scenarios. All network traffic was collected equally from the sensor and actuator during the data collection process. The dataset contains all the network traffic captured during this time. The Dataset also consists of all the values obtained from all the 51 sensors and actuators available in SWaT. The authors labeled all the data acquired during this process according to normal and abnormal behaviors. Attack Scenarios.The attack model considers the intent space of a CPS as an attack model.

SWaT has thirty-six attacks. This Dataset aims to assist various researchers in designing solutions, testing, evaluating, and comparison purposes for CPSs research. It is in the simulated category.

3. V2X Security Threats:

This V2X[93], contains normal and attack messages. The attack messages include DoS and Fabrication attacks. The authors of V2X divide the class of DoS attacks in terms of the frequency of the attack. Conversely, fabrication attacks are divided into speed, acceleration and heading. This Dataset is unlikely to be suitable for ML-based approaches.

4. LITNET-2020[94], dataset was benchmark generate from the real-world academic network consists of senders, collectors and this dataset contains real-world network traffic data and annotated attack captured over 10 months. It has 85 feature attributes and 12 attack types. The dataset can be useful to cybersecurity researchers and could be used as a modern benchmark network intrusion dataset.

VeReMi [95], Finally, the authors used the OMNeT++ simulation program to produce VeReMi. It is a simulated category consisting of message logs of onboard units and labeled ground truth. The VeReMi includes malicious messages intended to trigger incorrect application behavior. Moreover, the initial Dataset contains several simple attacks.

8. Data Set Characteristics comparison

In this section, the review explores our list of datasets. To be able to compare different datasets side by side and to help researchers find appropriate datasets for their specific evaluation scenario, it is necessary to define common Therefore,

Table 5. Description of Datasets

Dataset	No. Of Features	Binary classes	Multi-class	Environment
AWID	154	Yes	Yes (15)	WiFi
BoT_IoT	86	Yes	Yes (11)	IOT
UGR'16	47	Yes	Yes (9)	TICS+NIDS
CIC-Bell-DNS2021	32	Yes	Yes (4)	TICS+NIDS
CIC-Bell-DNS-EXF-2021	30	Yes	Yes (13)	TICS+NIDS
CICDarknet2020	85	Yes	Yes (9)	TICS+NIDS
CICDDoS2019	88	Yes	Yes (13)	TICS+NIDS
CICIDS2017	80	Yes	Yes (15)	TICS+NIDS
CIDD-001	15	Yes	Yes (5)	TICS+NIDS
CIRA-CIC-DoHBrw-2020	34	Yes	Yes (4)	TICS+NIDS
CSE-CICIDS2018	80	Yes	Yes (10)	TICS+NIDS
DS2OS	13	Yes	Yes (7)	IOT
IoT-23	86	Yes	Yes (10)	IOT
IoTID20	86	Yes	Yes (5)	IOT
MedBIoT	100	Yes	Yes (5)	IOT
MQTT-IOT-IDS2020	598	Yes	Yes (7)	IOT
N-BaIoT	115	Yes	Yes (11)	IOT
TON-IOT	83	Yes	Yes (10)	IOT

Table 6 : explore particular datasets properties used in the literature to assess intrusion detection datasets seven properties reflect general information about the Dataset, the year of creation, availability, presence of [normal and malicious network traffic], Balanced, Labeled and Updated Process. properties as an evaluation basis[96].

1) year of creation: Since network traffic is subject to concept drift and new attack scenarios show every day, the year of an intrusion detection data set plays an important role. The year in which the author's captured Dataset is more relevant for timeliness than the year of its publication.

2) Public Availability: datasets should be available to researchers for comparing different intrusion detection methods. Table 8 includes three other characteristics for public availability datasets: yes (means that the Dataset is available for researchers), no (means that the Dataset is not available for researchers), and on request (means that access will be given after sending a message to the authors).

3) Normal User Behavior: This property indicates the availability of normal user behavior within a dataset and takes the values yes or no. The value yes demonstrates that there is normal user behavior within the dataset. In general, academics primarily determine the quality of an IDS by its attack detection rate and false alarm rate. Therefore, normal user behavior is indispensable to evaluate an IDS. Nonetheless, the absence of normal user behavior does not make a dataset unusable. Instead, it indicates that it has to be merged with other datasets or real-world network traffic,such as merging overlaying and orsalting[97],[98].

4) Attack Traffic: IDS datasets should include variant attack scenarios. This property indicates the presence of malicious network traffic within a dataset and has the value yes if the data set contains at least one attack. Table 6 provides additional information about the specific attack types.

5) Balanced: For an accurate and high-accuracy model, our review recommended that either authors or donors balance the datasets concerning their class labels. Consequently, datasets should contain an equal or almost equal number of samples from the normal and attack classes. One of the potential issues in the field of machine learning is Imbalanced. Developers can approach this problem by correctly analyzing the data. A few approaches that help tackle the issue at the data point level are under-sampling, oversampling, and feature selection. He and Garcia[99]provide an overview of learning from imbalanced data.

6) Labeled: Labeled data sets are necessary for training supervised methods and evaluating supervised and unsupervised intrusion detection methods. This property denotes if data sets are labeled or not. Our review set this property to yes if there are at least two classes, normal and attack. Possible values in this property are: yes, yes with B.G. (yes with background), yes (IDS), indirect, and no. Yes, with background means that there is a third-class background. Packet flows or data points that belong to the class background could be normal or attacked.

7) Update process efficiently: can be upgraded Dataset to a new version as new attacks appear or the attacker uses new tactics.

9. Review Conclusion and findings

In this section, we will summarize popular benchmark datasets used for designing intrusion detection systems for IoT by developers. Machine learning research widely depends on datasets, whether these are publicly available or restricted access. Table 2, introduces standard datasets developers utilize to design intrusion detection systems for IoT. Moreover, previous surveys of IoT have not dealt with available datasets in the academic field that can be used to assess, design, and evaluate IDS for IoT.

Since donors of most of the Dataset gathered their Dataset over a TICST field lacks datasets appropriate for research in the IoT computing field. Thus, the survey sees a need for datasets gathered and collected based on real IoT computing environments. This context includes datasets that incorporate protocols such as MQTT, XMPP, LoRWAN, Bluetooth, Wi-Max, Zigbee, and NFC.

However, designers and developers can utilize datasets such as AWID and GRPS for an IoT that uses WiFi as the essential communication protocol across connected things. Nonetheless, the Dataset is collected based on datalink layer protocols and host-based audit material, which makes it suitable for host-based intrusion detection systems in the data link layer. The same rules are applied to GPRS and MQTT-IOT-IDS2020 Dataset.

Choosing a specific dataset to assess the IDS performance is a process that depends on both the IDS problems and the targeted security requirements. Furthermore, datasets are collected based on Hadoop, Bigdata, and AWS schemes.

Table 6. Characteristics of Datasets

SDataset	Issued Year	Publicly available	Normal Traffic	attack Traffic	Imbalance Ratio	Labeled	Updated Process
AWID	2016	Yes	Yes	Yes	No	Yes	Yes
UGR’16	2016	Yes	Yes	Yes	No	Yes	Yes
BoT_IoT	2018	Yes	Yes	Yes	No	Yes	Yes
CIC-Bell-DNS2021	2021	Yes	Yes	Yes	No	Yes	Not Yet
CIC-Bell-DNS-EXF-2021	2021	Yes	Yes	Yes	No	Yes	Not Yet
CICDarknet2020	2010	Yes	Yes	Yes	No	Yes	Yes
CIC-DDoS2019	2019	Yes	Yes	Yes	No	Yes	Not Yet
CICIDS2017	2017	Yes	Yes	Yes	No	Yes	Not Yet
CIDD-001	2017	Yes	Yes	Yes	No	Yes	Yes
CIRA-CIC-DoHBrw-2020	2020	Yes	Yes	Yes	No	Yes	Yes
CSE-CIC-IDS2018	2018	Yes	Yes	Yes	No	Yes	Yes
DS2OS	2018	Yes	Yes	Yes	No	Yes	Yes
IoT-23	2020	Yes	Yes	Yes	No	Yes	Yes
IoTID20	2020	Yes	Yes	Yes	No	Yes	Not Yet
MedBIoT	2020	Yes	Yes	Yes	No	Yes	Not Yet
MQTT-IOT-IDS2020	2020	Yes	Yes	Yes	No	Yes	Yes
N-BaIoT	2020	Yes	Yes	Yes	No	Yes	Not Yet
TON-IOT	2020	Yes	Yes	Yes	No	Yes	Yes

Developers and Academics can use IoT simulators such as Simple IoTSimulator, SUMO, QualNet, and Contiki Cooja for collecting data. The simulation scenario can simulate the IoT network that uses MQTT, CoAP, MQTT-Broker, COAP, or LoRWAN to collect the data related to the protocols and the network traffic of malicious and normal data. Indeed, the potential absence of efficient intrusion benchmarks or datasets relevant to IoT communication protocols and standards is a critical issue in academic research. This survey recommends choosing a dataset close to or identical to real-time network traffic.

Table 4 presents the most common intrusion detection system datasets widely used in academic research. Our review found that 40% of reported IDS samples used the NSL-KDD Dataset, even though this Dataset is less suitable for the IoT environment. Moreover, 12% of

reported IDS samples used the CICIDS2017, N-BaIoT, and IoTID20. RPL dataset and IoTs MQTT are more suitable for the IoT environment. Nonetheless, these two datasets were not collected by highly reputable cybersecurity centers such as CICIDS2017, N-BaIoT, and IoTID20.Figure 8 illustrates the Percentage of the Dataset Utilization in the state-of-the-art IDS for IoT.

Other environments datasets are scarce resources, and one of the findings of this review is that there is an urgent need for datasets that developers can use to design an IDS for these environments. We recommend a call for donations to the university so that a more publicly available dataset for these environments will be available soon. Moreover, most Datasets compatible with VANET, MANET, and CPS are either not publicly available or not from an authenticated source or donner. In this review, it include only publicly available Dataset from a certified source in the academic research field. Future developers must deal with the challenge of finding acceptable datasets for intrusion detection in IOTs and VANET, MANET, and CPS. The donors must design and create efficient and accepted realistic, albeit synthetic, datasets for intrusion detection. To develop an IDS-based machine learning for intrusion detection, the Dataset must include audit logs and raw network data.

To illustrate, CVS and PCAP files. Moreover, it must include a variety of present-day attacks.

Also, the Dataset must represent realistic and diverse normal traffic. In addition, one of the crucial aspects of the Dataset is that it should be carefully labeled and ensure privacy. Finally, the Dataset must be accepted by the academic research community. To test the designated, I.D.s, developers must use several datasets to check for results compatibility.

Acknowledgments

The project presented in this article is supported by Dr. Razan Abdulhammed. The authors would like to acknowledge the infrastructure support provided by the computer networking lab, Dept. of computer engineering technology, NTU.

Competing Interests

The authors declare that there are no competing interest.

References

[1] R. Abdulhammed, “Intrusion Detection: Embedded Software Machine Learning and

Hardware Rules Based Co-Designs," Ph.D. dissertation, Univ. of Bridgeport, 2019.

[2] S. Xu, M. Marwah, M. Arlitt, and N. Ramakrishnan, "Stan: Synthetic network traffic generation with generative neural models," in *Proc. Int. Workshop Deployable Mach. Learn. Security Defense*, 2021, pp. 3–29.

[3] A. Balyk, M. Karpinski, A. Naglik, G. Shangytbayeva, and I. Romanets, "Using graphic network simulator 3 for DDoS attacks simulation," *Int. J. Comput.*, vol. 16, no. 4, pp. 219–225, 2017.

[4] P. Mishra, E. S. Pilli, V. Varadharajan, and U. Tupakula, "NvCloudIDS: A security architecture to detect intrusions at network and virtualization layer in cloud environment," in *Proc. 2016 Int. Conf. Adv. Comput., Commun. Inform. (ICACCI)*, 2016, pp. 56–62.

[5] J. O. Nehinbe, "A critical evaluation of datasets for investigating IDSs and IPSs researches," in *Proc. 2011 IEEE 10th Int. Conf. Cybern. Intell. Syst. (CIS)*, 2011, pp. 92–97.

[6] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, pp. 1–30, 2018.

[7] R. Abdulhammed, M. Faezipour, A. Abuzneid, and A. AbuMallouh, "Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic," *IEEE Sens. Lett.*, vol. 3, no. 1, pp. 1–4, 2018.

[8] L. Gotsev, M. Dimitrova, B. Jekov, E. Kovatcheva, and E. Shoikova, "A cybersecurity data science demonstrator: Machine learning in IoT network security," unpublished.

[9] H. Tekleselassie, "A deep learning approach for DDoS attack detection using supervised learning," in *MATEC Web Conf.*, vol. 348, p. 01012, 2021.

[10] M. A. Khan *et al.*, "A deep learning-based intrusion detection system for MQTT enabled IoT," *Sensors*, vol. 21, no. 21, p. 7016, 2021.

[11] I. S. Thaseen *et al.*, "A Hadoop based framework integrating machine learning classifiers for anomaly detection in the Internet of Things," *Electronics*, vol. 10, no. 16, p. 1955, 2021.

[12] A. A. Reyes, F. D. Vaca, G. A. Castro Aguayo, Q. Niyaz, and V. Devabhaktuni, "A machine learning based two-stage Wi-Fi network intrusion detection system," *Electronics*, vol. 9, no. 10, p. 1689, 2020.

[13] W. T. W. A. of R. in Sci. and Eng., "A machine learning-based intrusion detection of DDoS attack on IoT devices," *Int. J. Adv. Trends Comput. Sci. Eng.*, Jan. 2021. [Online]. Available: https://www.academia.edu/50813262/

[14] S. Strecker, R. Dave, N. Siddiqui, and N. Seliya, "A modern analysis of aging machine learning based IoT cybersecurity methods," *arXiv preprint arXiv:2110.07832*, 2021.

[15] R. Qaddoura *et al.*, "A multi-layer classification approach for intrusion detection in IoT networks based on deep learning," *Sensors*, vol. 21, no. 9, p. 2987, 2021.

[16] S. Latif *et al.*, "A novel attack detection scheme for the industrial internet of things using a lightweight random neural network," *IEEE Access*, vol. 8, pp. 89337–89350, 2020.

[17] K. V. Krishna, K. Swathi, and B. B. Rao, "A novel framework for NIDS through fast kNN classifier on CICIDS2017 dataset," *Int. J. Recent Technol. Eng. (IJRTE)*, vol. 8, no. 5, 2020.

[18] Y. Wei *et al.*, "AE-MLP: A hybrid deep learning approach for DDoS detection and classification," *IEEE Access*, vol. 9, pp. 146810–146821, 2021.

[19] M. T. Jafar *et al.*, "Analysis and investigation of malicious DNS queries using CIRA-CIC-DoHBrw-2020 dataset," *Manch. J. Artif. Intell. Appl. Sci.*, vol. 2, pp. 65–70, 2021.

[20] A. Huč, J. Šalej, and M. Trebar, "Analysis of machine learning algorithms for anomaly detection on edge devices," *Sensors*, vol. 21, no. 14, p. 4946, 2021.

[21] V. Kanimozhi and T. P. Jacob, "Artificial intelligence outflanks all other machine learning classifiers in network intrusion detection system on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing," *ICT Express*, vol. 7, no. 3, pp. 366–370, 2021.

[22] K. S. Huancayo Ramos, M. A. Sotelo Monge, and J. Maestre Vidal, "Benchmark-based reference model for evaluating botnet detection tools driven by traffic-flow analytics," *Sensors*, vol. 20, no. 16, p. 4501, 2020.

[23] Z. K. Maseer *et al.*, "Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset," *IEEE Access*, vol. 9, pp. 22351–22370, 2021.

[24] C. S. Htwe, Y. M. Thant, and M. M. S. Thwin, "Botnets attack detection using machine learning approach for IoT environment," in *J. Phys.: Conf. Ser.*, vol. 1646, no. 1, p. 012101, 2020.

[25] A. A. Hezam *et al.*, "Combining deep learning models for enhancing the detection of botnet attacks in multiple sensors Internet of Things networks," *JOIV Int. J. Inform. Vis.*, vol. 5, no. 4, pp. 380–387, 2021.

[26] R. Gandhi and Y. Li, "Comparing machine learning and deep learning for IoT botnet detection," in *Proc. 2021 IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, 2021, pp. 234–239.

[27] A. Farah, "Cross dataset evaluation for IoT network intrusion detection," Ph.D. dissertation, Univ. of Wisconsin-Milwaukee, 2020.

[28] D. K. Reddy *et al.*, "Deep neural network based anomaly detection in Internet of Things network traffic tracking for the applications of future smart cities," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 7, p. e4121, 2021.

[29] Y. Banadaki, J. Brook, and S. Sharifi, "Design of intrusion detection systems on the internet of things infrastructure using machine learning algorithms," in *Proc. SPIE, NDE 4.0 Smart Struct. Ind., Smart Cities, Commun., Energy*, vol. 11594, p. 115940J, 2021.

[30] M. Alkasassbeh *et al.*, "Detecting distributed denial of service attacks using data mining techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 1, pp. 436–445, 2016.

[31] M. Alenezi, "Detecting network intrusions using hybrid learning approach," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 5, pp. 4540–4548, 2021.

[32] S. K. Vigneshwar and S. Selvakumar, "Detection of HTTP-DoS attacks using HODACS," in *Proc. 3rd Int. Conf. Comput. Commun. Inf. Technol. (CCICT)*, 2015, pp. 1–6.

[33] A. Ometov *et al.*, "DDoS attack dataset for machine learning applications," *Data*, vol. 6, no. 11, p. 117, 2021.

[34] M. N. Aman, B. Sikdar, and K. A. N. D. Parra, "Enhanced lightweight authentication and key agreement for smart grid communications," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 200–210, 2019.

[35] S. Srivastava *et al.*, "Evaluating the effectiveness of ML classifiers for anomaly-based intrusion detection using CICIDS2017 dataset," in *Proc. 2021 Int. Conf. Adv. Comput. Commun. Control Netw. (ICACCCN)*, 2021, pp. 1–7.

[36] H. N. Azizan, A. H. Basori, and A. H. Shah, "Feature selection and classification of the Bot-IoT dataset for the Internet of Things network traffic using machine learning approach," in *J. Phys.: Conf. Ser.*, vol. 1529, no. 3, p. 032011, 2020.

[37] S. Siraj, D. B. C. Wong, and R. J. Linda, "Improving intrusion detection in cloud environment using deep learning," *Int. J. Cloud Comput.*, vol. 9, no. 2/3, pp. 248–266, 2020.

[38] H. Shahriar, R. A. Shaikh, M. A. Qureshi, and H. Gao, "Intelligent botnet detection using machine learning for internet of things network: Dataset and techniques," *IEEE Access*, vol. 9, pp. 122068–122086, 2021.

[39] M. A. A. Alkadi *et al.*, "Intrusion detection system using machine and deep learning techniques: A review," *IEEE Access*, vol. 9, pp. 20635–20675, 2021.

[40] M. Aborokbah *et al.*, "IoT-based DDoS attack detection using machine learning: An empirical analysis," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 1, pp. 435–443, 2021.

[41] A. A. Fouda, "IoT devices' network security using a hybrid intrusion detection system," *Int. J. Comput. Netw. Inf. Secur.*, vol. 13, no. 3, pp. 24–32, 2021.

[42] S. Tangade *et al.*, "IoT intrusion detection using ensemble deep learning model," in *Proc. 2021 Int. Conf. Artif. Intell. Data Eng. (AIDE)*, 2021, pp. 135–139.

[43] M. Jabbar, A. A. Farhan, and R. Jalil, "Machine learning for IoT intrusion detection system: A survey," in *Proc. 2021 1st Int. Conf. Comput. Inf. Eng. (ICCIE)*, 2021, pp. 40–45.

[44] R. A. Shaikh *et al.*, "Machine learning-based intelligent botnet detection system: An empirical study," *Sensors*, vol. 21, no. 16, p. 5522, 2021.

[45] G. D. M. Garcia and M. J. Montesino, "Network intrusion detection using machine learning in IoT networks," in *Proc. 2021 IEEE World AI IoT Congr. (AIIoT)*, 2021, pp. 0374–0378.

[46] S. S. Kumar, "Network intrusion detection in IoT using machine learning algorithms," *Mater. Today: Proc.*, 2021. [Online]. Available: https://doi.org/10.1016/j.matpr.2021.11.054

[47] M. Anwar and S. M. M. Rahman, "Performance analysis of machine learning algorithms for smart intrusion detection system," in *Proc. 2021 IEEE 11th Annu. Comput. Commun. Workshop Conf. (CCWC)*, 2021, pp. 0202–0206.

[48] M. T. Jafar *et al.*, “Performance evaluation of supervised machine learning techniques for DNS over HTTPS traffic classification,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 126–133, 2021.

[49] D. Doshi *et al.*, “Preserving security and privacy in IoT using machine learning,” *ICT Express*, vol. 7, no. 1, pp. 106–112, 2021.

[50] K. T. Ali and B. S. Ali, “Secure architecture for IoT communication using AI-based IDS,” in *Proc. 2021 2nd Int. Conf. Comput. Inf. Sci. (ICCIS)*, 2021, pp. 1–5.

[51] N. I. Ali, M. K. Jabbar, and B. S. Ali, “Security of Internet of Things using machine learning and deep learning approaches: Survey,” in *Proc. 2021 1st Int. Conf. Comput. Inf. Eng. (ICCIE)*, 2021, pp. 36–39.

[52] I. S. B. Sardjono and R. R. A. Hidayat, “The implementation of machine learning in network intrusion detection system,” *J. Phys.: Conf. Ser.*, vol. 1811, no. 1, p. 012035, 2021.

[53] R. K. Sharma, “Two-stage network intrusion detection system using feature selection and ensemble learning,” *J. King Saud Univ.-Comput. Inf. Sci.*, 2021. [Online]. Available: https://doi.org/10.1016/j.jksuci.2021.07.007

[54] N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets," *Future Generation Computer Systems*, vol. 72, May 2021.

[55] S. S. Gopalan, "Towards Effective Detection of Botnet Attacks using BoT-IoT Dataset," M.S. thesis, [Institution not provided], p. 76, [Year not provided].

[56] D. L. R. Wilson, *Towards Effective Wireless Intrusion Detection Using AWID Dataset*, M.S. thesis, Rochester Institute of Technology, 2021.

[57] D. Ravikumar, *Towards Enhancement of Machine Learning Techniques Using CSE-CIC-IDS2018 Cybersecurity Dataset*, M.S. thesis, Rochester Institute of Technology, 2021.

[58] D. Stiawan, M. Y. B. Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 dataset feature analysis with information gain for anomaly detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020.

[59] D. Stiawan, "Important Features of CICIDS-2017 Dataset For Anomaly Detection in High Dimension and Imbalanced Class Dataset," [Online]. Available: [Publisher not specified], 2021.

[60] 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, 2019.

[61] M. Ge et al., "Deep learning-based intrusion detection for IoT networks," in *Proc. 2019 IEEE 24th Pacific Rim Int. Symp. Dependable Computing (PRDC)*, 2019, pp. 256–25609.

[62] B. A. Tama and K.-H. Rhee, "Attack classification analysis of IoT network via deep learning approach," *Research Briefs on Information and Communication Technology Evolution (ReBICTE)*, vol. 3, pp. 1–9, 2017.

[63] S. Wang, B. Li, M. Yang, and Z. Yan, "Intrusion detection for WiFi network: A deep learning approach," in *Proc. Int. Wireless Internet Conf.*, 2018, pp. 95–104.

[64] L. Nicholas et al., "Study of long short-term memory in flow-based network intrusion detection system," *J. Intell. Fuzzy Syst.*, vol. 35, no. 6, pp. 5947–5957, 2018.

[65] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. ICISSP*, vol. 1, pp. 108–116, 2018.

[66] N.-A. Stoian, *Machine Learning for Anomaly Detection in IoT Networks: Malware Analysis on the IoT-23 Data Set*, B.S. thesis, University of Twente, 2020.

[67] M. Lavallée, P.-N. Robillard, and R. Mirsalari, "Performing systematic literature reviews with novices: An iterative approach," *IEEE Trans. Educ.*, vol. 57, no. 3, pp. 175–181, 2013.

[68] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, "Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset," *IEEE Commun. Surv. Tutor.*, vol. 18, no. 1, pp. 184–208, 2015.

[69] M. Ring et al., "Flow-based benchmark data sets for intrusion detection," in *Proc. 16th European Conf. Cyber Warfare and Security*, ACPI, 2017, pp. 361–369.

[70] [70] G. Maciá-Fernández et al., "UGR ‘16: A new dataset for the evaluation of cyclostationarity-based network IDSs," *Comput. Secur.*, vol. 73, pp. 411–424, 2018.

[71] I. Sharafaldin et al., "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *Proc. 2019 Int. Carnahan Conf. Security Technology (ICCST)*, 2019, pp. 1–8.

[72] Y. Meidan et al., "N-BaIoT—Network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Comput.*, vol. 17, no. 3, pp. 12–22, 2018.

[73] A. Parmisano, S. Garcia, and M. J. Erquiaga, "A labeled dataset with malicious and benign IoT network traffic," *Stratosphere Lab, Czech Republic*, 2020.

[74] N. Koroniotis et al., "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset," *Future Gener. Comput. Syst.*, vol. 100, pp. 779–796, 2019.

[75] I. Ullah and Q. H. Mahmoud, "A scheme for generating a dataset for anomalous activity detection in IoT networks," in *Proc. Canadian Conf. Artificial Intelligence*, 2020, pp. 508–520.

[76] A. Alsaedi et al., "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.

[77] M. MontazeriShatoori et al., "Detection of DoH tunnels using time-series classification of encrypted traffic," in *Proc. 2020 IEEE DASC/PiCom/CBDCom/CyberSciTech*, 2020, pp. 63–70.

[78] M.-O. Pahl and F.-X. Aubet, "All eyes on you: Distributed multi-dimensional IoT microservice anomaly detection," in *Proc. 2018 Int. Conf. Network and Service Management (CNSM)*, 2018, pp. 72–80.

[79] A. Guerra-Manzanares et al., "MedBIoT: Generation of an IoT Botnet Dataset in a Medium-sized IoT Network," in *Proc. ICISSP*, 2020, pp. 207–218.

[80] A. H. Lashkari, G. Kaur, and A. Rahali, "DIDarknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning," in *Proc. 2020 Int. Conf. Commun. Network Security*, pp. 1–13.

[81] S. Mahdavifar, N. Maleki, A. H. Lashkari, M. Broda, and A. H. Razavi, "Classifying malicious domains using DNS traffic analysis," in Proc. IEEE Intl Conf. Dependable, Autonomic and Secure Comput., Pervasive Intell. Comput., Cloud Big Data Comput., Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech), 2021, pp. 60–67.

[82] S. Mahdavifar et al., "Lightweight hybrid detection of data exfiltration using DNS based on machine learning," in Proc. 11th Int. Conf. Commun. Netw. Security (CNS), 2021, pp. 80–86.

[83] H. Hindy, C. Tachtatzis, R. Atkinson, E. Bayne, and X. Bellekens, "MQTT-IoT-IDS2020: MQTT Internet of Things intrusion detection dataset," IEEE Dataport, 2020.

[84] U. S. Musa, M. Chhabra, A. Ali, and M. Kaur, "Intrusion detection system using machine learning techniques: A review," in Proc. Int. Conf. Smart Electron. Commun. (ICOSEC), 2020, pp. 149–155.

[85] S. H. Kok, A. Abdullah, N. Z. Jhanjhi, and M. Supramaniam, "A review of intrusion detection system using machine learning approach," Int. J. Eng. Res. Technol., vol. 12, no. 1, pp. 8–15, 2019.

[86] R. Ferdiana, "A systematic literature review of intrusion detection system for network security: Research trends, datasets and methods," in Proc. 4th Int. Conf. Informatics Comput. Sci. (ICICoS), 2020, pp. 1–6.

[87] P. Sadhana, P. Priyanka, and A. Chetan, "A review of intrusion detection datasets and techniques," Smart Moves J. IJOSCIENCE, vol. 6, no. 3, pp. 14–22, 2020.

[88] A. Thakkar and R. Lohiya, "A review of the advancement in intrusion detection datasets," Procedia Comput. Sci., vol. 167, pp. 636–645, 2020.

[89] I. Almomani, B. Al-Kasasbeh, and M. Al-Akhras, "WSN-DS: A dataset for intrusion detection systems in wireless sensor networks," J. Sens., vol. 2016, 2016.

[90] D. W. Vilela, T. F. Ed'Wilson, A. A. Shinoda, N. V. de Souza Araújo, R. De Oliveira, and V. E. Nascimento, "A dataset for evaluating intrusion detection systems in IEEE 802.11 wireless networks," in Proc. IEEE Colombian Conf. Commun. Comput. (COLCOM), 2014, pp. 1–5.

[91] S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, "Labelled data collection for anomaly detection in wireless sensor networks," in Proc. 6th Int. Conf. Intell. Sensors, Sensor Netw. Inf. Process., 2010, pp. 269–274.

[92] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A dataset to support research in the design of secure water treatment systems," in Int. Conf. Critical Inf. Infrastructures Security, 2016, pp. 88–99.

[93] F. Gonçalves et al., "Synthesizing datasets with security threats for vehicular ad-hoc networks," in Proc. IEEE Global Commun. Conf. (GLOBECOM), 2020, pp. 1–6.

[94] R. Damasevicius et al., "LITNET-2020: An annotated real-world network flow dataset for network intrusion detection," Electronics, vol. 9, no. 5, p. 800, 2020.

[95] R. W. Heijden, T. Lukaseder, and F. Kargl, "Veremi: A dataset for comparable evaluation of misbehavior detection in VANETs," in Int. Conf. Security Privacy Commun. Syst., 2018, pp. 318–337.

[96] I. Sharafaldin, A. Gharib, A. H. Lashkari, and A. A. Ghorbani, “Towards a reliable intrusion detection benchmark dataset,” Softw. Netw., vol. 2018, no. 1, pp. 177–200, 2018.

[97] A. J. Aviv and A. Haeberlen, “Challenges in experimenting with botnet detection systems,” in Proc. 4th Workshop Cyber Security Experimentation Test (CSET), 2011.

[98] Z. B. Celik, J. Raghuram, G. Kesidis, and D. J. Miller, “Salting public traces with attack traffic to test flow classifiers,” in Proc. 4th Workshop Cyber Security Experimentation Test (CSET), 2011.

[99] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009.