Mesopotamian Journal of Big Data
Vol. **(2025)**, 2025, **pp**. 50–60
DOI: https://doi.org/10.58496/MJBD/2025/004 ; ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

MESOPOTAMIAN
JOURNAL OF BIG DATA

Research Article

Overview of the CICIoT2023 Dataset for Internet of Things Intrusion Detection Systems

Wisam Ali Hussein Salman [1,*], Chan Huah Yong [2,]

[1] *Ministry of Education, Directorate General of Education in Holy Karbala Province, Karbala, Iraq.*

[2] *Department of Computer Science, UNIVERSITI SAINS MALAYSIA, Pulau Penang, Malaysia.*

ARTICLE INFO

Article History

Received 13 Feb 2025
Revised 12 Apr 2025
Accepted 14 Apr 2025
Published 10 Jun 2025

Keywords

Dataset

Machine Learning

Deep Learning

IDS

IPS

ABSTRACT

The rapid expansion of the use of the Internet of Things (IoT) has encouraged many attackers to exploit the vulnerabilities in these networks to violate data privacy or disrupt service; they are easy targets due to the diversity of devices within the network, which has led to the loss of unified security standards. intrusion detection system (IDS) play a pivotal role in securing IoT networks by monitoring inbound and outbound traffic to these networks and issuing a security alarm when there is an attack; moreover, they respond directly to these security threats to prevent them from harming the network and violating data privacy. To design an IDS capable of performing work with high efficiency, an appropriate dataset must be chosen to train and evaluate the designed model. This dataset works as a fundamental task in the success of these systems because it plays a major role in training the system, feature engineering, evaluating the performance of the model, and other tasks. This paper focused on one of the modern datasets used in training and evaluating IDS models, that is, the CICIOT2023 dataset. The CICIOT2023 dataset is distinguished from other datasets, such as CICIDS2017, UNSW-NB15, and KDD1999. It focuses on the IoT environment, unlike other datasets that focus on data traffic in traditional networks, and it uses a variety of devices and protocols; moreover, it contains modern and complex attacks and a balance between the data of those attacks and normal traffic. This paper discusses the structure of the dataset, the kinds of attacks it contains, the applications and fields in which it is used, the strengths that distinguish it from other datasets, its role in developing cybersecurity research, the most important studies that have been written and dealt with this dataset, and finally, the future visions for developing the dataset.

1. INTRODUCTION

A dataset is an organized sequence of data arranged within a specific structure (in the form of tables or something such as that); each column in that table represents a specific feature within the scope of the group's work, and each row represents an individual data sample containing the values associated with each feature [1]. The major function is to train and assess artificial intelligence systems (machine and deep learning systems) utilized in different domains of life, such as intrusion detection and prevention systems [2]. The dataset is divided according to its organization as follows:

- Structured dataset: Organized as tables with columns and rows, each column is a feature, and every row provides a sample, for example, the CICIDS2017, NSL-KDD1999, CICIOT2023 and other datasets [3].
- Unstructured dataset: refers to data without a specific format, such as text, photos, and videos. For example, the PCAP dataset and text log files [4].

There is another classification of the dataset:

- Supervised dataset: This type of dataset contains a label that is used to train machine learning models.
- Unsupervised dataset: This type does not contain labels and uses unsupervised learning with it [5].

The selection of a good dataset contributes significantly to the development of robust and effective intelligent IDS systems [6]. There are several criteria for measuring the quality of a dataset, such as the following:

- Size: This size should be large enough to cover different types of scenarios.
- Realism: It must represent real data with which the model works.
- Variety: It should contain various types of data covering as much real-world data as possible according to which the model works.

*Corresponding author. Email: wissamali77@gmail.com

- Balance: It should contain equal proportions of the required types of data; for example, it should contain balanced proportions of the types of attacks [7].

There are many datasets used in the cybersecurity field, such as CICIDS2017, NSL-KDD1999, UNSW-NB15, and CICIOT2023. In this study, we provide a comprehensive view of the CICIOT2023 dataset [8].

Most of the previous datasets were a major challenge for the design of IDS because of their lack of realism and comprehensiveness [9]. The CICIOT2023 dataset has emerged to address this gap by providing a rich dataset designed specifically to simulate the IoT environment and capture the traffic that travels in it, whether normal or malicious traffic [10].

The Canadian Institute for Cybersecurity (CIC) created the CICIOT2023 dataset, which is one of the most important datasets used in conducting cybersecurity research, as it is characterized by comprehensiveness because it contains integrated data on realistic IoT network traffic by capturing all normal traffic, as well as the various attacks that IoT networks are exposed to in the real world [11].

Compared with previous datasets, CICIOT2023 contains modern attacks, and their numbers are larger than those of the previous datasets (which contain 34 attacks). Moreover, the data are collected from a set of real IoT devices, such as sensors, smart cameras, smart doorbells, and smart plugs, in addition to the diversity of protocols used, which makes it more qualified to work in IoT environments because of its reliability and the diversity of scenarios within it. In addition, most of the studies discussed in this paper proved that their use produced more accurate intrusion detection systems that can be used in distributed and hybrid environments [11].

1.1. CICIOT2023 Data Collection

A CICIOT2023 dataset was created by the Canadian Cybersecurity Institute to provide a real and comprehensive dataset for the IoT environment, collected through a real physical simulation environment containing various devices such as sensors, smart cameras, etc.; multiple protocols such as DNS, HTTP, HTTPS, and MQTT; various modern attacks; and a real operating platform (servers, Wi-Fi, access point, firewall, and Wireshark). Moreover, the attacks they contain are real attacks generated using professional tools such as Nmap, Hping3, and Metasploit, and all of these generated a rich dataset containing more than 14 million records, 46 features and 34 attacks [12].

This helps make it completely reliable in building robust intrusion detection systems, especially in training and evaluating IDSs [13]. As described below:

- The CICIOT2023 dataset contains (47) features that represent various types of information of the network traffic, some of which assist in the speed of data training and performance of intrusion detection tasks, whereas others do not help; moreover, it can cause noise that helps in the breakdown of the training process and the occurrence of so-called overfitting Table (1), which illustrates the features enclosed in the dataset [13].

TABLE I. ILLUSTRATES THE FEATURES INCLUDED IN THE DATASET.

Drate	**Duration**	**Protocol type**	**Header_Lenght**	**Flow_duration**
Fin_count	Cwr_flag_number	Ece_flag_number	Ece_flag_number	Psh_flag_number
SSH	DNS	HTTPS	HTTP	RST_COUNT
Syn_flag_number	Fin_flag_number	Rst_flag_number	DHCP	IRC
LLC	Max	IPv	ICMP	ARP
Radius	IAT	Tot size	Std	AVG
label	weight	Variance	Covariance	
UDP	Srate	TCP	Rate	
Min	Syn_count	Tot sum	Ack_count	
Magnitue	SMTP	Number	telnet	

- The CICIOT2023 dataset contains (34) different attacks representative of real attacks that the IoT network was exposed to while collecting data in the real world, and it also represents most of the common attacks used by attackers [13]. Table (2) lists the types of attacks and the number of records of each type for a sample of data that includes (1,000,000) restrictions.

TABLE II. ILLUSTRATES THE TYPES OF ATTACKS IN THE CICIOT2023 DATASET

Class Type	**No. of Records**	**Class Type**	**No. of Records**	**Class Type**	**No. of Records**
DDoS-ICMP_Flood	153893	DoS-SYN_Flood	43132	DNS_Spoofing	4691
DDoS-UDP_Flood	115631	BenignTraffic	23320	DoS-HTTP_Flood	4299

DDoS-TCP_Flood	96640	Mirai-greeth_flood	21100	Recon-HostDiscovery	3475
DDoS-PSHACK_Flood	88089	Mirai-udpplain	19222	Recon-OSScan	3192
DDoS-SYN_Flood	87382	Mirai-greip_flood	16165	DictionaryBruteForce	1809
DDoS-RSTFINFlood	85479	DDoS-ICMP_Fragmentation	9744	Recon-PortScan	1775
DDoS-SynonymousIP_Flood	76916	MITM-ArpSpoofing	6664	VulnerabilityScan	959
DoS-UDP_Flood	71192	DDoS-ACK_Fragmentation	6148	DDoS-HTTP_Flood	598
DoS-TCP_Flood	57053	DDoS-UDP_Fragmentation	6137	DDoS-SlowLoris	466
SqlInjection	143	XSS	106	Backdoor_Malware	74
BrowserHijacking	134	CommandInjection	103	Recon-PingSweep	39
Uploading_Attack	22				

Figure (1) illustrates the types of attacks in the CICIOT2023 dataset.

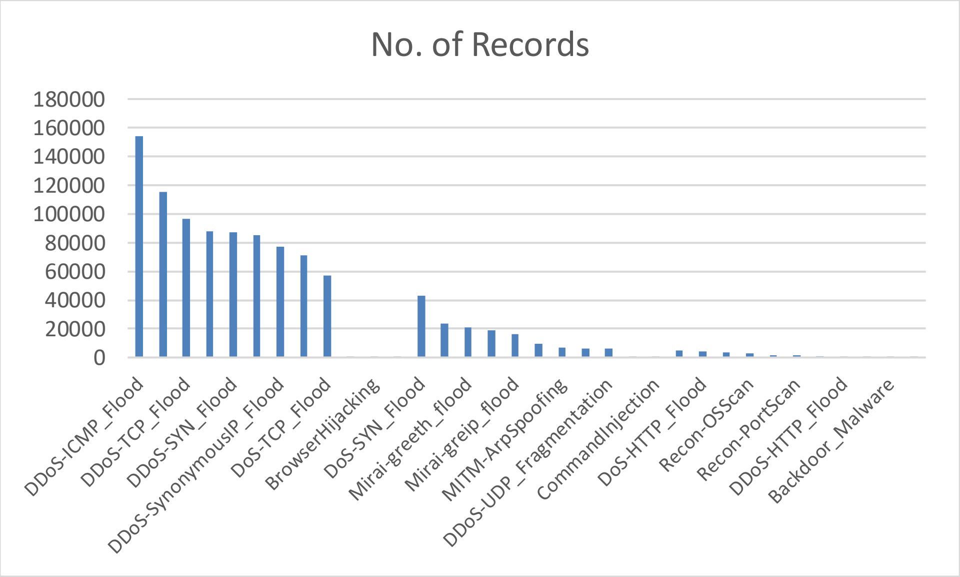

Fig. 1. illustrates the types of attacks in the CICIOT2023 dataset.

The CICIOT2023 dataset is a multiclass labelling type; at the end of each entry, it is tagged as either normal traffic or a specific type of malicious motion, which helps with both binary and multiclass classification problems, and it is also mainly used with supervised machine learning [13].

1.2. CICIOT2023 Dataset Advantages

1.2.1. Realistic IoT Traffic Simulation

A comprehensive dataset simulating various real-world IoT environments, representing an integrated smart home network with various devices (smart cameras, smart home appliances, sensors, etc.) helps build highly accurate IDSs, reducing both false positive rate (FPR) and false negative rate (FNR) [15].

1.2.2. Wide Variety of Attack Types

It contains a variety of modern attacks and approaches, including (34) different attacks, such as distributed denial-of-service attacks, brute force, botnet activity, man-in-the-middle (MITM) attacks, and port scanning. This helps build robust security systems that can handle multiple scenarios and detect new and hybrid attacks [16].

1.2.3. Comprehensive feature set

It contains a wide range of (47) features, such as traffic data, time, and network data, which provide an accurate representation of network behavior and help support more powerful machine learning algorithms and build accurate hybrid models [17].

1.2.4. Labelled Data for Supervised Learning

The CICIOT2023 dataset contains a classification that indicates network traffic (either normal or suspicious traffic) along with the attack type. This helps accurately distinguish between attacks and normal activities, accurately measures model performance, supports multiclass models, and finally helps analyses the different behaviors of each type [17].

1.2.5. Supports both binary and multiclass classification

The CICIoT2023 dataset can adapt to a wide range of detection methods, ranging from widespread anomaly detection to target attack type classification. This approach can handle binary classification tasks (normal vs. attack) and multiclass classification tasks (distinguishing between attack types), which helps adapt to diverse kinds of detection approaches and the possibility of combining different models, in addition to supporting different algorithms in multiple disciplines [18].

1.2.6. Enables time series analysis

To analyse time series data, models such as recurrent neural network (RNN) and long short term memory (LSTM) can be trained on the temporal attributes of a dataset. Time series analysis is critical for capturing time patterns in IoT traffic, which enables the detection of complex and evolving threats such as DDoS and MITM attacks [19].

1.2.7. Applicability to real-time intrusion detection

Real-time detection of hackers is extremely important for detecting and controlling threats before they cause significant damage to IoT networks, so using a CICIoT2023 dataset is very important because of its ability to detect hackers in real time by providing traffic flow data that resemble real-world situations [20].

1.2.8. Large and Diverse Dataset

The CICIoT2023 dataset has a large amount of normal and unwanted traffic collected across multiple attack scenarios, resulting from daily traffic in IoT networks, which helps to improve model generalizability and avoid overfitting, both of which are critical in building a robust IDS [21].

1.2.9. Class Imbalance Handling

The dataset has a natural category imbalance (attacks bypassing normal traffic) that reflects real-world scenarios in which attacks are rare. This allows academics to create solutions to imbalance problems for IDS categories, such as the synthetic minority oversampling technique (SMOTE) [22].

1.2.10. Open Source and Easily Accessible

The CICIoT2023 dataset is freely accessible to researchers and developers around the world. The dataset's open-source design promotes collaboration and innovation, enabling individuals and organizations to compare systems utilizing a common dataset [23].

1.2.11. Encourages Feature Engineering Research

The comprehensive feature set of CICIoT2023 allows researchers to use feature selection and engineering to increase detection performance. Feature engineering is critical for reducing model complexity and increasing detection accuracy, especially in resource-limited IoT devices [23].

1.2.12. Relevant for Smart Home and Industrial IoT Security

Smart home simulation datasets can be used not only in consumer IoT contexts but also for industrial IoT security. This application enables academics to concentrate on both home automation and corporate IoT devices, broadening dataset results [24].

1.3. CICIoT2023 Dataset Limitations.

1.3.1. Class Imbalance

One of the most critical limitations of the CICIoT2023 dataset is the imbalance between the classes in the sample due to the superiority of regular traffic over attack traffic, which leads to bias in systems designed using Machine learning (ML) algorithms, leading to failure to detect rare types of attacks. This limitation can be mitigated by using one of the following methods: data resampling, the use of techniques that consider imbalanced data (random forest (RF), support vector machine (SVM), or XGBoost), or the modification of evaluation criteria such as F1 score, recall, or precision [25].

1.3.2. Limited IoT Device Diversity

The CICIoT2023 dataset is focused on a virtual smart home environment that includes a specified collection of IoT devices (such as smart lighting devices, recording devices, and smart home assistants such as google home and Amazon Alexa, Personal Computers and local servers, smart plugs, and, finally, routers and switches). While this arrangement is useful, it does not represent the whole range of IoT devices available in other settings, such as industrial (IoT) or health IoT. This limited diversity may limit the applicability of artificial intelligence systems trained on this dataset to diverse IoT contexts, where devices may use various methods of communication and vulnerabilities. This problem can be alleviated by combining with another dataset or using a transfer learning strategy [26].

1.3.3. Lack of encrypted traffic

The statistics exclude network traffic that is encrypted, which is becoming increasingly frequent in IoT environments. Modern IoT devices frequently use encryption to secure connections, and the lack of encrypted communication in the dataset limits their capacity to model real-world settings. Machine learning models created using CICIoT2023 may suffer when deployed in situations with encrypted internet traffic since they have not been trained on secured communication patterns. This limitation can be solved by using technologies such as MQTT or Wireshark to generate synthetic encrypted traffic and enrich the dataset, using technology such as metadata-based features to encrypt the dataset [21].

1.3.4. Simplified attack scenarios

The attack simulations in CICIoT2023, however diverse, are frequently simplified and may fail to replicate the complexities of real-world intrusions on IoT devices. In real-world circumstances, attackers frequently combine approaches, adapt their strategies, and conceal their actions, which are not always captured in static datasets. Machine learning models built on CICIoT2023 may perform well in the dataset's-controlled environment, but they may not be effective in real-world IoT systems with greater complexity and multiphase attacks. This limitation can be mitigated by utilizing reinforcement learning or by using advanced simulation techniques such as Metasploit and CALDERA [27].

1.3.5. Limited Temporal Coverage

The dataset contains network traffic statistics collected during a brief period in computer simulation. Real-world IoT devices emit traffic continually over lengthy periods of time, allowing attack patterns to grow or long-term abnormalities to emerge. Models developed on CICIoT2023 may struggle to perform in long-running IoT networks when attacker dynamics or normal activity patterns change over time. This limitation can be mitigated by combining a CICIOT2023 dataset with a long-term dataset, as well as by using advanced time series techniques such as the LSTM and RF algorithms [15].

1.3.6. Absence of Physical Layer Data

The CICIoT2023 dataset primarily collects network layer data, including IP addresses, network ports, protocols, and packet sizes. However, it lacks data from the physical and data-link layers, which are critical for identifying certain forms of IoT-specific assaults, such as jamming or attacks on protocols for wireless communication, such as ZigBee or LoRa. The lack of physical layer data limits dataset utility for constructing intrusion detection systems that target a small number of attacks on wireless communication protocols. This limitation can be mitigated by using specialized sensors or devices for the physical layer, such as "USRP (Universal Software Radio Peripheral)" and "RTL-SDR dongle". Moreover, a dataset can be combined with another dataset covering the physical layers, such as the "Contiki-NG" dataset or the "Wireless IoT Attack" dataset [28].

1.3.7. Limited real-time evaluation

The dataset is valuable for constructing intrusion detection systems; however, it does not provide real-time evaluation. Many academics concentrate on offline detection, which may underestimate the difficulties of installing IDS in real-time IoT contexts with stringent latency and performance constraints. Machine learning models built on CICIoT2023 may perform well in offline tests, but they fail to fulfil real-time requirements in production scenarios where latency, the use of resources, and quick response are crucial. This limitation can be mitigated by converting a CICIOT2023 dataset to a streaming-supported format or using lightweight models suitable for real-time execution [29].

1.3.8. Synthetic Nature of the Dataset

Like many other cybersecurity datasets, CICIoT2023 was developed in a monitored simulated setting. While this allows for precise recording of attack circumstances, it may not completely mimic the randomness and unpredictable character of real-world IoT network traffic; however, IDS systems trained on synthetic data may fail to generalize to real-world environments where noise, benign abnormalities, and other unknown events affect network traffic [30].

1.4. CICIoT2023 Dataset Applications

1.4.1. Machine Learning for IoT Intrusion Detection

The dataset provides a solid foundation for the application of Artificial Intelligent approaches to detect odd data flows and classify various types of attacks. Supervised and unsupervised techniques can be used to differentiate between lawful and harmful conduct [31]. Popular methods include the following:

- Deep Learning (DL) algorithms, such as the LSTM algorithm and RNN technology, may be effective in detecting similarities in network data [28].
- Traditional ML algorithms, including RF, SVM, and KNN, are commonly applied to datasets such as CICIoT2023 [28].

1.4.2. Real-Time IDS models

The dataset may be utilized to create real-time IDSs that track network flow from IoT devices and detect potential threats. These models can be used in industrial IoT applications, smart homes, etc. [33].

2. RELATED WORK

The CICIoT2023 dataset was generated via the Canadian Cybersecurity Institute in 2023. It is intended to replicate the smart home environment and gadgets, collect both regular and harmful network traffic, and then train and evaluate the efficiency of intrusion detection systems meant to defend the IoT environment. It competes with other datasets, including CICIDS2017, NSL-KDD 1999, UNSW-NB 15, TON_IoT 2020, and UN. Several researchers have utilized the latest dataset. This section describes the most important work.

2.1 ML methods for IoT intrusion detection systems

(Neto et al., 2023) Academics have used common methods such as SVMs and RFs to detect brute force and robotic attacks. They obtained accurate rates by carefully selecting traffic features based on flow statistics and protocol specifications. The paper emphasized the importance of engineering aspects in improving the efficiency of standard ML algorithms when dealing with IoT traffic [33].

2.2 Applying DL Approaches for IoT Intrusion Detection

Traditional ML approaches are unable to detect sophisticated threats in IoT environments; therefore, academics have turned to DL strategies to develop enhanced threat detection models capable of extracting information from network data automatically and immediately.

(Akinul Islam Jony & Arjun Kumar, 2024) proposed the use of the LSTM algorithm and the CICIoT2023 dataset to detect malicious code in network traffic data, which changes over time. Their LSTM-based model outperformed traditional ML techniques, with a high detection rate for Distributed Denial of Serves (DDOS) and MITM attacks. LSTM's temporal characteristics make it excellent for assessing time-based aspects in IoT data flows, which are frequently ignored in static systems [34].

(Gueriani et al., 2024) The CICIoT2023 dataset is suggested for usage with a hybrid DL model for IDSs that combines a Convolution Neural Network (CNN) and LSTM. By integrating CNNs for physical property extraction and LSTM for historical data analysis, their model improved the accuracy and speed of zero-day attack detection, making it appropriate for real-time IDS deployment in IoT environments [16].

2.3 Hybrid methods and ensemble learning

Many researchers have investigated mass learning techniques to improve the effectiveness of systems for detection.

(Xiao et al., 2023) The CICIoT2023 dataset, which includes packaged technologies such as XGBoost and LightGBM workbooks, was utilized for detecting a wide range of threats, including DDoS and port scanning. Bulk technology has excelled in individual techniques in terms of false positives and computing efficiency. This is crucial for resource-constrained IoT devices [35].

(Alsaedi et al., 2020) employed a hybrid technique to determine traffic in the IoT, combining ML models with traditional DL. Radio frequencies are chosen to select the key feature, and then automatic encoders are used to detect abnormalities. This technique allows the model to effectively liquidate unprocessed traffic while focusing on the most arduous task of finding new attacks [36].

2.4 Intrusion Detection with Real-Time and Deployment

The feasibility of real-time detection was investigated via datasets such as CICIoT2023.
(Thamraj Nareudra & Dr. Amol V. Zade, 2023) focused on lowering the detection latency while maintaining high accuracy. They developed a lightweight IDS for low-energy IoT devices that make use of recurrent neural networks. The model was validated via CICIoT2023 traffic data in a simulated smart home setting and achieved real-time detection without adversely influencing system performance [37].

2.5 Comparison with other IoT datasets

One of the primary topics of cybersecurity research is comparing datasets utilized with other datasets to measure performance to create efficient and rapid IDSs that detect abnormal traffic.
(Gad et al., 2022) One of the main fields of research in the domain of cybersecurity is to compare the dataset used with other data groups to determine the performance of designing effective and rapid infiltration systems in discovering a movement that has been used by the TOOT2020 data collection to assess the effectiveness of many ML techniques to detect threats in the IoT. The study stressed the benefit of data groups such as CICIOT2023, which not only picks up traffic but also reflects different kinds of attacks in houses of questionable traffic [38].
(Koroniotis et al., 2018) The Bot-IoT dataset was utilized to test various IDS models in IoT networks, leading to the conclusion that datasets such as CICIoT2023 are necessary for better simulating various kinds of IoT devices and network traffic behaviors encountered in practice [21].
The following table (number 3) shows the comparison between the CICIOT2023 dataset and other datasets used in the same field.

TABLE III. SHOWS THE COMPARISON BETWEEN THE CICIOT2023 DATASET AND OTHER DATASETS

Features	CICIDS - 2017 Dataset	UNSW-NB 15 Dataset	ISOT-CID Dataset	HIKARI-2021 Dataset	CICIOT2023 Dataset
Year	2017	2016	2018	2021	2023
Size	51.1 G.B.	100 G.B.	2.5 T.B.	11.770 G.B.	27 G.B.
Features Number	78	49	19	86	47
Records Number	2,827,876	2,540,044	24,519,987	400,000	5.062.868
Number of Attack categories	7	9	4	4	34
Network	Traditional	Traditional	Traditional	IoT	IoT
Feature extraction tools	CICFlowmeter 4.0	Argus, Bro-IDS	Bro-IDS tool	Zeek-IDS, python tools	CICFlowmeter 4.0

2.6 Feature engineering and imbalanced classes

One of the most difficult challenges in using this set of data to train and test IDSs designed to secure the IoT environment is developing reliable and robust IDSs, as studies have had to employ a variety of methods, such as excessive samples, insufficient samples, and sensitive learning. To address this issue, researchers must carefully select solutions based on their application type and the unique challenges of the data group. The following are the major methods utilized to address the issue of imbalance in the CICIOT2023 data group.
(waskle subbash et al., 2020) proposed addressing this issue in their analysis of the CICIoT2023 dataset by employing the SMOTE technique to balance the dataset and improve the performance of their IDS. Their findings indicate that rectifying class imbalances is crucial in constructing the most reliable intrusion detection models, particularly in IoT environments where certain forms of attacks are not exploited effectively [39].
(Baniak & Kingsmith, 2018) stressed the importance of feature engineering in improving the efficiency of ML model discovery on IoT data. Their research on extracting features from network traffic characteristics on the CICIoT2023 dataset demonstrated that selecting relevant features can lead to enhanced classification results with lower computational costs, making them perfect for deployment on low-resource IoT devices [40].
(Shanmugam et al., 2024) used the oversampling technique from smaller classes by creating synthetic samples from that class to balance the dataset via the SMOTE technique. Owing to the increased number of attacks, the trained models were able to better differentiate between normal traffic and attack traffic, which led to a significant enhancement in the detection performance of unusual types of attacks [21].
(He & Garcia, 2009)We suggest randomly removing a set of predominant traffic samples (either normal or attack) to balance the dataset, resulting in faster model training times and improved feature detection, although there is a risk of losing important data due to deletion [41].

(Chawla et al., 2002) proposed the hybrid sampling approach, which involves oversampling the minority class and under sampling the majority class to balance the classes without unnecessarily expanding the dataset size. The results lead to the building of model training with high IDS capabilities [27].
(Chandola et al., 2009)showed that since attacks in IoT environments are relatively rare compared with normal traffic, anomaly detection techniques treat the majority class as "normal" and flag deviations as potential intrusions. This method circumvents the need for strict balancing of classes. Anomaly detection models, such as autoencoders or isolation forests, can be trained on normal traffic from CICIOT2023, and any significant deviation from this behavior is flagged as an attack. Thus, anomaly detection can be particularly effective in detecting zero-day or novel attacks that may not be well represented in the dataset [15].
(Elkan, n.d.) proposed that assault categories incur greater erroneous categorization costs. When ML methods are trained via CICIoT2023, the technology pushes the model to focus on minimizing mistakes in underrepresented categories. This method improves attack detection without altering the dataset and can be easily integrated into systems such as random forests or SVM [42].
(Alsaedi et al., 2020) The use of existing data to generate extra data samples for smaller groups has been proposed. For network traffic data, this can involve strategies such as changing feature values or creating offensive traffic with simulators. This can improve the strength of models for variations in assault patterns and allow for better generalization of invisible attack types [36].
(Xiao et al., 2023) To enhance the discovery of the minority category, group methods such as reinforcement (for example, XGBoost and AdaBoost) and packing (for example, random forest) are used to integrate the outputs of many works. Since the group methods are designed to improve the performance of the model frequently on difficult samples, they can be used on CICIOT2023 to discover the types of attacks that are not more active. The group models are generally more resistant to imbalance between categories and can improve performance without modifying the dataset. Compared with typical automatic learning techniques, the results revealed that the attack categories were discovered more accurately [35].
(Saito & Rehmsmeier, 2015)highlighted the importance of using suitable evaluation tools for unbalanced data groupings, particularly when spotting irregularities or hazardous programs. Traditional accuracy assessments may not be appropriate for unbalanced datasets because the dominating group may control them. Instead, metrics such as accuracy, remembering, F1 degree, and space fall under the accuracy and remembrance curve (AUC-PR) and are better suited to evaluate trained models on unbalanced data. These scales enable a better grasp of the model on minority groups while avoiding the consequences of tragedy [43].

3. CONCLUSION

With the use of several datasets to build infiltration detection systems to secure IoT networks, such as CICIDS2017, NSL-KDD1999, USW-NB15, and TON_IOT2020, the Ciciot2023 data group emerged as a valuable resource to compete with the available data groups, as this group is characterized by its ability to utilize automatic learning strategies such as RF, SVM and the ability to use it with deep learning methods such as RNN, CNN, LSTM, and others. It can also be combined with hybrid models to create effective protection systems. CICIOT2023 collection data is valuable resources for IoT security researchers and practitioners. It accurately simulates the movement of the IoT, including a variety of assaults that are appropriate for today's cybersecurity concerns. The gathering of detailed information and classifications enables the building and testing of ML models to detect intrusion. However, issues such as group imbalance and the ability to generalize data groups must be addressed to establish a robust and successful penetration discovery. Most of the previous datasets focused on traditional computer networks, unlike the CICIOT2023 dataset, which was specifically designed to represent a smart home with all its devices, which caused it to contain real traffic. In addition, it contains many attacks type (34) attacks, whereas the previous attack contains only (7) attacks, which helps with the ability to address different scenarios. Moreover, it contains high-quality feature (47) features, and finally, it contains label traffic, which helps in classifying each record in it as normal traffic or an attack.

4. FUTURE WORKS

The limitations of the CICIOT2023 dataset should be noted when it is utilized for research or development. While it has many benefits, overcoming these constraints in future datasets or studies would enhance the accuracy and practicality of IDSs for IoT. Future directions for scientific research to develop the CICIOT2023 dataset could follow five main paths:

4.1. Improving the representation of the dataset

The CICIOT2023 dataset should capture a new variety of IoT devices, in addition to the devices that have been used previously, in addition to new types of attacks and encrypted traffic that are becoming increasingly common in the IoT environment.

4.2. Using advanced ML strategies

Researchers can use sophisticated ML methods, such as mixed learning and collaborative learning, to improve attack detection accuracy.

4.3. The ability to adapt to novel attacks

The process of constantly updating dataset models, such as CICIOT2023, with new attacks to address new and evolving threats represents the most prominent challenge for IDS systems used to maintain the security of the IoT environment.

4.4. Use of new devices

One of the essential recommendations of the CICIOT2023 dataset is to add new devices to the network, such as healthcare devices (heart rate monitors, blood glucose meters) and Industrial Internet of Things devices such as pressure sensors, temperature sensors, humidity sensors, smart traffic lights, and building management systems, as well as devices that use modern protocols such as Z-Wave, NB-IoT, and Bluetooth low energy.

4.5. Using federated learning and transfer learning techniques

Using federated learning techniques with the CICIOT203 dataset helps train models without the data leaving the machines, which helps increase privacy. We also recommend using transfer learning techniques, which involve transferring knowledge from a task with specific data to a new task with different but related data. This helps mitigate dataset imbalances and reduce training costs.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

None.

Acknowledgment

None.

References

[1] A. A. H. I. K. and T. A.-Q. , "Using Data Anonymization in big data analytics security and privacy," *Mesopotsmisn journal of Big Data,* vol. 2024, no. 1, pp. 118-127, 2024.

[2] K. I. F. E. F. &. S. A. "Machine learning methods for cyber security intrusion detection: Datasets and comparative study," *Computer Networks,* vol. 188, 2021.

[3] A. S. E. K. L. K. G. I. L.-D. K. E. &. G. P. "Dataset search: a survey," *The VLDB Journal,* vol. 29, p. 251–272, 2020.

[4] Y. Z. L. X. L. T. W. D. W. J. Z. Y. &. H. H. "A systematic literature review of methods and datasets for anomaly-based network intrusion detection," *Computers and Security,* vol. 116, 2022.

[5] Q. T. &. L. Y. J. "Fast LDP-MST: An Efficient Density-Peak-Based Clustering Method for Large-Size Datasets," *IEEE Transactions on Knowledge and Data Engineering,* vol. 35, p. 4767–4780, 2023.

[6] K. B. D. S. S. F. N. E. C. P. X. P. I. S. L. P. R. S. Ghorbani, A. A., "Internet of Things (IoT) security dataset evolution: Challenges and future directions," *In Internet of Things (Netherlands),* vol. 22, 2023.

[7] M. Z. K. Y. R. B. N. M. S. A. &. F. C. F. M. "Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset," *IEEE Access,* vol. 9, p. 22351–22370, 2021.

[8] B. S. A. O. H. I. O. O. A. A. O. O. A. M. C. &. A. M. O. "Survey dataset on the types, prevalence and causes of deviant behavior among secondary school adolescents in some selected schools in Benin City," *Edo State,* vol. 20, p. 101–107, 2018.

[9] G. H. Q. &. A.-Y. W. L. "Two-step data clustering for improved intrusion detection system using CICIoT2023 dataset," *E-Prime - Advances in Electrical Engineering, Electronics and Energy,* vol. 9, 2024.

[10] E. A. A. M. A. M. H. M. E. A. D. "Intrusion Detection System for loT Using CICloT2023 Dataset," *NILES 2024 - 6th Novel Intelligent and Leading Emerging Sciences Conference, Proceedings,* p. 512–516, 2024.

[11] K. A. G. R. A. &. R. V. "Evaluation of Different Machine Learning Classifiers on New IoT Dataset CICIoT2023," *2024 International Conference on Intelligent Systems for Cybersecurity,* 2024.

[12] T. N. &. R. K. "Development of Intrusion Detection Models for IoT Networks Utilizing CICIoT2023 Dataset," *Proceedings of the 3rd 2023 International Conference on Smart Cities, Automation and Intelligent Computing Systems, ICON-SONICS 2023,* p. 66–72, 2023.

[13] T. S. M. W. Y. Q. &. W. Y. C. "Multi-Class Intrusion Detection Based on Transformer for IoT Networks Using CIC-IoT-2023 Dataset," *Future Internet,* vol. 16, 2024.

[14] C. F. &. R. G. "ToN-IOT Set: Classification and Prediction for DDoS Attacks using AdaBoost and RUSBoost," *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2023,* p. 2842–2847, 2023.

[15] G. A. K. H. &. M. A. C. "Enhancing IoT Security with CNN and LSTM-Based Intrusion Detection Systems. http://arxiv.org/abs/2405.18624," *IEEE Transactions on Knowledge and Data Engineering,* 2024.

[16] A. I. J. &. A. K. "A long short-term memory based approach for detecting cyber attacks in IoT using CIC-IoT2023 dataset," 2024.

[17] S. I. L. A. H. &. G. A. A. "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy,* p. 108–116, 2018.

[18] S. A. E. E. I. I. A. S. A. "DCNN: a novel binary and multi-class network intrusion detection model via deep convolutional neural network," *Eurasip Journal on Information Security,* vol. 1, 2024.

[19] D. A. A. A. E. A. Z. &. K. F. "Intrusion Detection System for Internet of Things Based on Temporal Convolution Neural Network and Efficient Feature Engineering," *Wireless Communications and Mobile Computing,* 2020.

[20] S. H. "Designing an Adaptive Effective Intrusion Detection System for Smart Home IoT A Device-Specific Approach with SDN Deployment," *International Journal of Advanced Computer Science and Applications ,* vol. 15, no. 1, 2024.

[21] K. N. M. N. S. E. T. B. "Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset," 2018.

[22] N. K. R. M. S. O. V. P. T. A. C. &. D. A. K. "IIDS: Design of Intelligent Intrusion Detection System for Internet-of-Things Applications," 2023.

[23] C. F. &. R. G. "ToN-IOT Set: Classification and Prediction for DDoS Attacks using AdaBoost and RUSBoost," *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2023,* p. 2842–2847, 2023.

[24] G. E. &. J. A. "Intrusion Detection in Internet of Things Systems: A Review on Design Approaches Leveraging Multi-Access Edge Computing," *Machine Learning, and Datasets. In Sensors,* vol. 22, no. 10, 2022.

[25] S. V. R.-F. R. &. H. E. "Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches," *Electronics,* vol. 14, no. 1, 2024.

[26] D. W. T. M. M. T. &. Z. S. "A Novel Realistic Dataset for Intrusion Detection in IoT based on Machine Learning," *2021 International Symposium on Networks, Computers and Communications, ISNCC 2021,* 2021.

[27] K. A. &. C. "EFFICIENT ATTACK DETECTION IN IOT DEVICES USING FEATURE ENGINEERING-LESS MACHINE LEARNING," 2023.

[28] E.-H. E. B. and A. H. K. , "Optimizing Big Data Analytics for Reliability and Resilience_ A Survey of Techniques and Applications," *Mesopotamain Journal of Big Data,* vol. 2023, no. 1, p. 118124, 2023.

[29] D. S. C. P. N. E. F. R. C. M. R. S. S. Ghorbani, A., "CICIoMT2024: Attack Vectors in Healthcare devices-A Multi-Protocol Dataset for Assessing IoMT Device Security," 2024.

[30] B.-S. F. L. T.-M. V. A. M.-C. H. I. &. A.-D. J. "Performance Evaluation of Deep Learning Models for Classifying Cybersecurity Attacks in IoT Networks," *Informatics,* vol. 11, no. 2, 2024.

[31] D. R. S. L. Z. W. D. L. D. Z. &. X. B. "Towards universal and transferable adversarial attacks against network traffic classification. Computer Networks, 254. https://doi.org/10.1016/j.comnet.2024.110790," *The Foundations of Cost-Sensitive Learning,* 2024.

[32] Y. S. &. D. M. "IoT-Based Intrusion Detection System Using New Hybrid Deep Learning Algorithm," 2024.

[33] M. A. A. R. A. H. and S. O. , "A Framework for Automated Big Data Analytics in Cybersecurity Threat Detection," *Mesopotamian Journal of Big Data,* vol. 2024, no. 1, pp. 175-184, 2024.

[34] N. E. C. P. D. S. F. R. Z. A. L. R. &. G. A. A. "CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment," *Sensors,* vol. 13, p. 23, 2023.

[35] X. P. X. M. C. N. Q. B. Z. S. &. W. H. "Adaptive Hybrid Framework for Multiscale Void Inspection of Chip Resistor Solder Joints," *IEEE Transactions on Instrumentation and Measurement,* vol. 72, 2023.

[36] A. A. M. N. T. Z. M. A. &. A. N. A. "TON-IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access,* vol. 8, p. 165130–165150, 2020.

[37] T. N. &. D. A. V. Z. "Hybrid CNN+LSTM Deep Learning Model for Intrusions Detection Over IoT Environment," 2023.

[38] G. A. R. H. M. N. A. A. B. T. M. "A Distributed Intrusion Detection System using Machine Learning for IoT based on ToN-IoT Dataset," *International Journal of Advanced Computer Science and Applications,* vol. 13, no. 6, 2022.

[39] w. s. L. P. &. U. S. "Intrusion Detection System Using PCA with Random Forest Approach," *Institute of Electrical and Electronics Engineers (IEEE),* 2020.

[40] B. G. M. &. K. K. G. "Sedimentological and stratigraphic characterization of Cretaceous upper McMurray deposits in the southern Athabasca oil sands," *Alberta, Canada. AAPG Bulletin,* vol. 102, no. 2, p. 309–332, 2018.

[41] H. H. &. G. E. A. "Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering," vol. 21, no. 9, p. 1263–1284, 2009.

[42] E. "The Foundations of Cost-Sensitive Learning," 2023.

[43] S. T. &. R. M. "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced dataset," *PLoS ONE,* vol. 10, no. 3, 2015.