Analysis of IoT Security Datasets

Erdal ÖZDOĞAN
Uludağ University

Onur CERAN
Gazi University

To Cite This Chapter

Özdoğan, E., & Ceran, O. (2024). Analysis of IoT Security Datasets. In M. H. Calp & R. Butuner (Eds.), *Current Studies in Data Science and Analytics* (pp. 124–143). ISRES Publishing.

Introduction

The Internet of Things (IoT) is a revolutionary technology characterized by vast data volumes across various domains. IoT systems continuously generate data through different sensors and devices, creating large data sets. This situation presents various challenges in data analysis and management. These large data volumes' processing, storage, and analysis are crucial for enhancing efficiency and obtaining meaningful insights. However, it is essential to consider that alongside the opportunities presented by IoT, various data security concerns also arise (Maiwada et al., 2024).

The security of IoT systems fundamentally relies on the protection of the data and devices present within these systems (Alrayes et al., 2024). Various data security concerns can pose threats to user privacy, data integrity, and the system itself. Among these concerns, attacks specifically designed for IoT networks hold considerable significance. Such attacks can target vulnerabilities in devices, leading to data breaches, unauthorized access, and service disruptions. Therefore, preventing these threats and mitigating their impacts plays a vital role in ensuring the security of IoT systems.

Effective solutions are required for the prevention and detection of attacks. In this context, Intrusion Detection Systems (IDS) can be utilized to secure IoT environments (Aggarwal & Sharma, 2015). IDS systems monitor network traffic, possessing the capability to detect unusual behaviors and take measures against these behaviors. The implementation of such a protection mechanism in IoT systems is crucial for enhancing the security of devices and data(Liang et al., 2024). IDSs offer a highly effective method for identifying different types of attacks and developing strategies against them.

Advanced IDS systems, when supported by machine learning-based approaches, possess strong protective potential (Bansal & Singhrova, 2023), (Saadouni et al., 2024). Machine learning algorithms can quickly identify anomalous behaviors due to their ability to analyze large data sets. The use of IoT data sets enhances the effectiveness of machine learning models. When enriched with information obtained from various sources, such as inter-device communication and user interactions, it becomes possible to predict and prevent attacks in advance.

Machine learning-based IDS systems both increase efficiency and adopt a more proactive approach against potential security threats. The utilization of such systems is not limited to the detection of security threats; they also contribute to mitigating the impacts of attacks and continuously improving the security of the system. Furthermore, the applications of these systems enhance the secure management of data, thereby raising the overall

security level of IoT systems. In conclusion, IoT systems are characterized by large data volumes and various data security concerns, presenting both opportunities and threats. Attacks unique to IoT threaten the security of these systems, while IDS systems provide a critical solution for protection against these threats. The strong protective potential of machine learning-based IDS can be further strengthened using IoT data sets. Thus, the security of IoT systems can be enhanced, and data management can be carried out more effectively.

The objective of the Research and Hypothesis

This chapter aims to explore the datasets commonly referenced in IDS systems used for IoT security in detail. Given the complexities of the networks formed by numerous interconnected devices, effective data analysis and attack detection are critical for ensuring their security. However, the effectiveness of various datasets cannot be fully assessed without conducting comparative analyses. In this context, the key concepts highlighted in this chapter will guide the examination of the datasets, aiming to identify the most suitable datasets for IoT IDS systems and contribute to research in this field.

Hypothesis 1: Implementing machine learning-based IDS will improve the security of IoT systems by enhancing the detection and prevention of various cyber threats.

Hypothesis 2: A comprehensive introduction of the datasets used in the field of IoT security will facilitate cybersecurity researchers in making more informed decisions regarding attack detection and prevention, thereby increasing the effectiveness of their research activities.

Contribution

This study will comprehensively examine and analyze commonly referenced data sets used in IDS systems for IoT security. IoT systems are characterized by complex network structures formed by the convergence of numerous devices. Effective data analysis and attack detection are critically important for ensuring the security of these structures. However, the effectiveness of various data sets cannot be fully evaluated unless a comparative analysis is conducted. In this context, the chapter will explore data sets based on key concepts that stand out. This aims to identify the most suitable data sets for IoT IDS systems and contribute to research in this field.
In this scope, the contributions of this book chapter can be summarized as follows:

- Multifaceted Analysis of Data Sets: Critical data sets for IoT security are analyzed through various dimensions and labeling types, determining in which scenarios these data sets can be utilized more effectively.
- Identification of Most Effective Features: The pre-processing needs of different data sets have been examined, and the most effective features that positively contribute to model performance have been revealed. This facilitates the development of more optimized data processing methods for IoT security.
- Diversity of Attacks: The diversity of attacks presents in the data sets used for IoT security has been investigated. This analysis will aid security researchers conducting research on specific attack types in making more informed data set preferences.

The remainder of the book chapter is organized as follows: Chapter 2 addresses current studies conducted in the field. Chapter 3 examines the role of data sets in IoT security. Chapter 4 investigates commonly used data sets and compares them within the framework of key concepts. The evaluation and conclusion section summarizes the insights obtained from the study.

Related Work

In recent years, numerous academic studies and research related to IoT-IDS, particularly

those based on machine learning, have been conducted and continue to be undertaken. The increase in the use of IoT, along with the rapid updates of machine learning models, has necessitated a high volume of academic work in this field.

In a study conducted in 2024 (Ozdogan, 2024), the effects of data preprocessing and feature selection on the simplification of machine learning algorithm selection in IoT IDS systems were analyzed in detail. In the study, which utilized multiple datasets, the performance of Machine Learning algorithms was compared from various aspects, including the preprocessing process of the datasets, the process of being balanced, and whether they were generated in a synthetic or natural environment. In the work of, the authors presented a comparative analysis of the IoT datasets used for model training, identifying key features that assist in evaluating their suitability in specific scenarios. Hota and Shrivas utilized different feature selection techniques on the NSL-KDD dataset to contribute to making Intrusion Detection Systems more efficient and effective (Hota & Shrivas, 2014). In another study conducted on the same dataset, Vibhute et al. used the NSL-KDD dataset to develop a network intrusion detection system. The study proposed and implemented a community learning-supported random forest algorithm to select the most suitable features (Vibhute et al., 2024). In his study using the NSL-KDD and UNSW-NB15 datasets, Türk conducted binary and multi-class intrusion detection experiments and achieved high-performance results (Türk, 2023). In another study using the NSL-KDD dataset, a model was created by reducing the number of features to ten using a feature selection method. To enhance prediction performance, imbalanced data was corrected using the SMOTE method (Thana-Aksaneekorn et al., 2024). In another study, the authors proposed a new IDS based on Artificial Neural Networks using the NSL-KDD dataset. The developed model was compared in terms of performance with several classifiers and achieved high success (Alrayes et al., 2024). In another study (Zoghi & Serpen, 2024), the impact of class imbalance and data overlap issues in the UNSW-NB15 dataset on the performance of data-driven models were examined. To improve classifier performance, a scalable overlap visualization method capable of detecting these issues was proposed, and its accuracy was tested with various classifiers.

A recent study using the Bot-IoT dataset proposes a new approach that combines deep learning and three-tiered algorithms to quickly and accurately detect attacks in IoT networks. Evaluations have shown that this method significantly improves detection performance compared to existing methods (Alosaimi & Almutairi, 2023).

In a study utilizing the IoTID20 and BoT-IoT datasets, a hybrid method combining PCA and the Bat Optimization Algorithm (BAT) has been proposed for dimensionality reduction in the datasets (Karamollaoğlu et al., 2024). In the study that achieved high performance, detailed analyses were conducted to determine the effects of dimensionality reduction and data balancing models on classification performance.

In another study that evaluated the performance comparisons of machine learning models, a comparative analysis of various algorithms on the Bot-IoT dataset was presented (Mishra et al., 2023). In another comprehensive study, the aim was to categorize and analyze existing datasets to create future datasets, thereby enhancing the effectiveness of intrusion detection systems and accurately reflecting network threats (Zafar Iqbal Khan et al., 2024).

Upon reviewing the studies conducted in recent years on various datasets, it can be seen that machine learning and deep learning techniques have been applied to detect cyber-attacks in IoT networks. The studies particularly focus on topics such as data preprocessing, feature selection, class imbalance, and dimensionality reduction, aiming to improve model performance. Additionally, comparative analyses of various algorithms are conducted to identify the most effective methods. Therefore, examining IoT security datasets from a data analysis perspective will provide an opportunity to test the effectiveness of new model approaches and algorithms. Furthermore, analyzing different datasets is important for understanding different types of attacks and anomalies,

leading to more comprehensive and generalizable results.

The Role of Datasets in IoT Security

In the context of IoT security, datasets play a critical role in developing effective IDS and enhancing overall system resilience against various cyber threats. The vast array of devices connected to the Internet of Things generates a large amount of data that can be leveraged for security analysis and threat detection. Data sets provide the raw data necessary to understand events and activities occurring within IoT networks. These data sets are used for training machine learning and deep learning models, directly influencing the effectiveness of these models. By utilizing diverse and high-quality data sets, researchers can develop more accurate and reliable models that enhance the security and functionality of IoT systems. The quality and representativeness of the data sets play a crucial role in enabling these models to generalize well to real-world scenarios, ultimately improving the detection and response to various threats and anomalies in IoT environments (Kaur et al., 2023).

A well-structured dataset that encompasses diverse attack scenarios allows researchers and practitioners to develop and evaluate models that can accurately identify malicious activities. For instance, datasets containing labeled instances of both benign and malicious traffic enable supervised learning techniques, which are essential for developing robust detection algorithms.

The effectiveness of an IDS is heavily reliant on the authenticity of the dataset used for training. Datasets that mimic real-world conditions, including variations in traffic patterns, device types, and attack methodologies, enhance the model's ability to generalize and perform effectively in live environments. The current methods used for labeling existing IoT datasets are based on generating synthetic network data, which overlooks the essential aspects needed to differentiate between normal and malicious behaviors (Guerra et al., 2022). Thus, the inclusion of realistic scenarios in datasets is paramount for preparing systems to confront actual threats.

Many datasets face the challenge of class imbalance, where certain attack types are underrepresented compared to benign instances (Qing et al., 2024). This imbalance can lead to biased models that fail to detect less frequent but critical threats. Effective preprocessing techniques, such as oversampling or under-sampling, can help mitigate these issues, ensuring that models are trained on a balanced representation of all classes.

Datasets allow for comparative studies that help identify the strengths and weaknesses of various IDS approaches. By evaluating multiple algorithms against the same dataset, researchers can determine which methods yield the highest accuracy and efficiency in detecting specific types of attacks (Ozdogan, 2024). This comparative analysis contributes to the continuous improvement of detection techniques and enhances the overall security posture of IoT systems.

The establishment of standardized datasets fosters a common ground for research in IoT security. These datasets serve as benchmarks for assessing the performance of IDS algorithms, facilitating a clearer understanding of advancements in the field (Neto et al., 2023). By utilizing widely recognized datasets, researchers can share their findings more effectively, driving innovation and collaboration in the cybersecurity community.

In summary, datasets are indispensable in the realm of IoT security, providing the necessary foundation for developing, testing, and refining intrusion detection methodologies. The careful selection and use of datasets impact the effectiveness of security measures, underscoring their pivotal role in safeguarding IoT environments.

Data Analysis of Datasets Used in IoT Security

The datasets used in IoT security have evolved alongside technological advancements.

The first generation of datasets typically featured simple structures with a limited number of devices and types of attacks. Over time, these datasets have become more complex, creating extensive data pools that encompass a greater variety of device types, various attack vectors, and different use case scenarios.

The primary datasets used for research and development in the IoT field generally focus on network traffic analysis, anomaly detection, classification of attack types, and the security of IoT devices. These datasets vary according to different network environments, types of attacks, and feature engineering requirements.

NSL-KDD

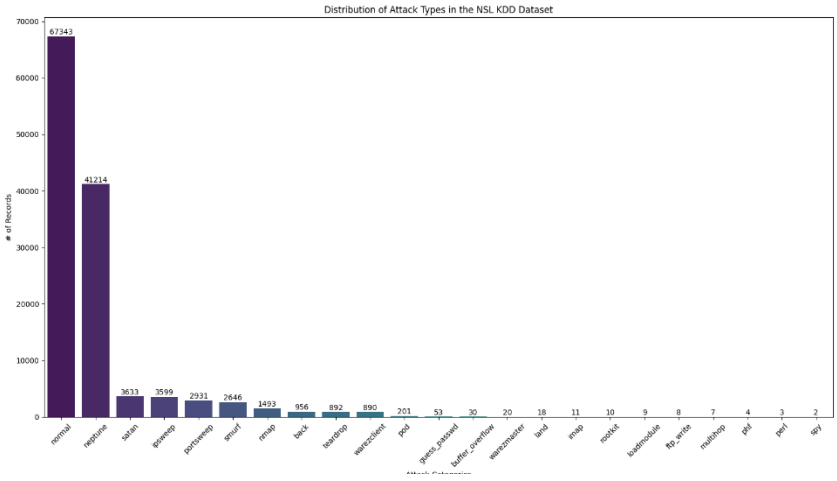
The NSL-KDD dataset is a widely used benchmark for network-based intrusion detection systems (Alosaimi & Almutairi, 2023). Developed in 2009, it addresses several issues found in the original KDD'99 dataset. One of its key improvements is the removal of duplicate and redundant records in both the training and testing subsets. This helps prevent classifiers from being biased towards more frequently occurring examples. Like the original KDD'99 dataset, the NSL-KDD dataset was created in a controlled laboratory environment. It includes records generated from a simulated environment that features various types of attacks as well as normal network traffic.

The NSL-KDD dataset consists of 41 features, each representing a different aspect of network behavior. It is divided into two main subsets Train and Test. The training dataset contains 125,973 and the testing dataset includes 22,544 records.

This dataset features a diverse range of attack types. Figure 1 illustrates the distribution of traffic types in the training dataset.

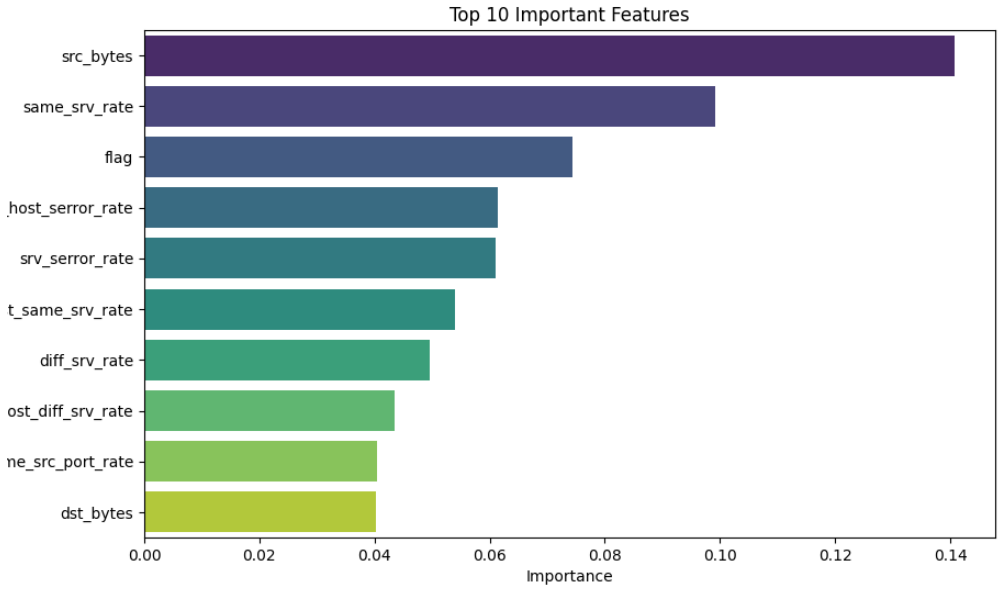
Figure 1

Distribution of Traffic Types in the NSL KDD Training Dataset



In the traffic distribution table, which is imbalanced due to the overwhelming presence of normal traffic, it is evident that the most common attack is the Neptune attack. The most effective features for attack classification are illustrated in Figure 2.

Figure 2
The Top 10 Important Features for Attack Classification of NSL KDD Dataset.



In the dataset, the most effective feature for classification is src_bytes, which indicates the amount of source bytes.

The NSL-KDD dataset provides an effective benchmark for comparing various intrusion detection methods.

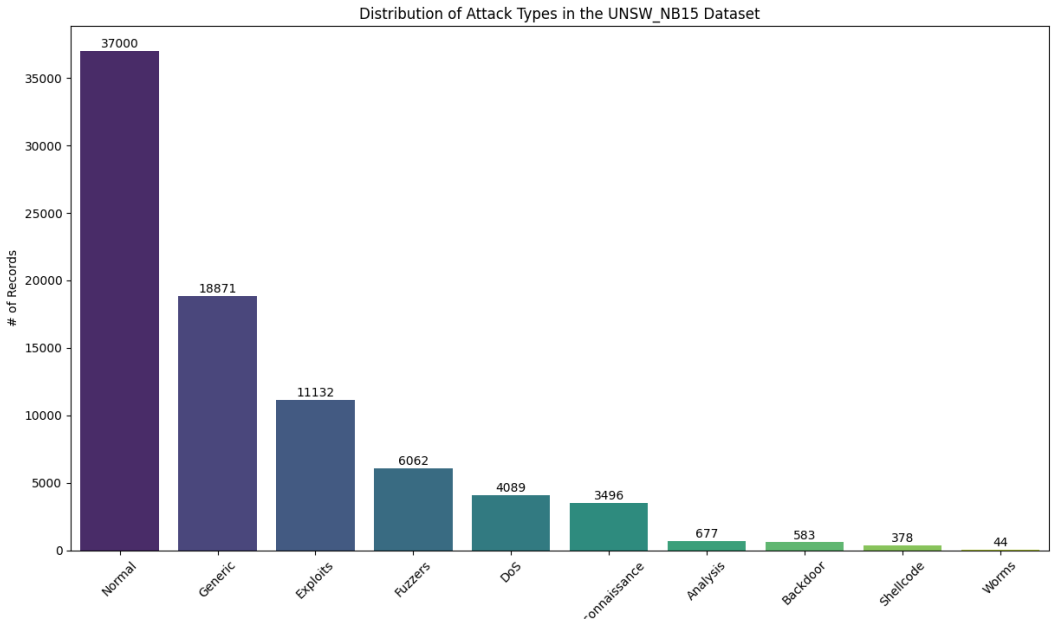
UNSW-NB15

The UNSW-NB15 dataset serves as a critical resource for the evaluation of network intrusion detection systems. This dataset was generated using the IXIA PerfectStorm tool and encompasses a comprehensive array of modern normal activities alongside various malicious attack behaviors (Moustafa & Slay, 2015, 2016).

Comprising a total of 2,540,044 records, the dataset is structured to facilitate division into training and testing subsets, with additional subdivisions available for specific research applications. The UNSW-NB15 dataset was meticulously produced in a controlled laboratory environment, specifically collected from a simulated network setup at the Cyber Range Lab located in Canberra, Australia. The network traffic within this dataset is intentionally designed to reflect contemporary attack vectors. In this unbalanced dataset, traffic is categorized as either normal or attack, with the latter further delineated into multiple attack types.

An analysis of the training subset, which consists of 82,332 records, reveals the distribution of attack types, as illustrated in Figure 3.

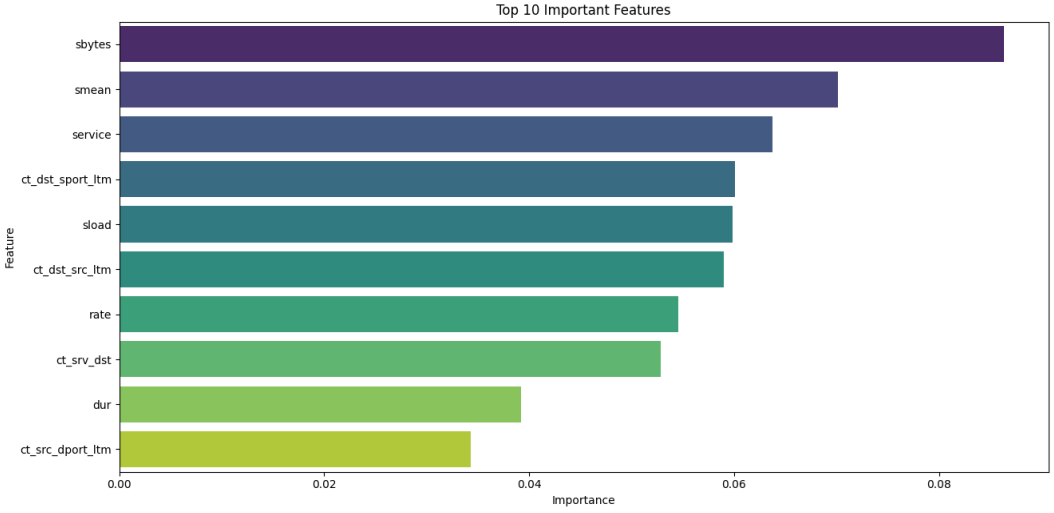
Figure 3
Distribution of Traffic Types in the UNSW-NB15 Dataset



As illustrated in Figure 3, the UNSW-NB15 dataset exhibits significant class imbalance. The distribution of different attack types and normal traffic records is not uniform, leading to discrepancies in the representation of classes within the dataset.

The classification of the traffic data using the Random Forest algorithm identifies the ten most important features contributing to this process, as depicted in Figure 4.

Figure 4
The Best 10 Features for Classification of UNSW-NB15 Dataset



The Random Forest model generates an importance score for each feature based on metrics such as total Gini decrease or entropy gain. This score indicates the model's effectiveness in utilizing that feature for classification tasks. Accordingly, the most important feature in the classification of attacks is the sbytes feature, which represents the number of source bytes. The remaining values and their impact ratios are presented in Figure 4.

The UNSW-NB15 dataset encompasses more contemporary and sophisticated attack techniques compared to earlier datasets such as KDD99 and NSL-KDD. Consequently,

it is frequently employed in cybersecurity research, particularly in domains such as IoT, network security, and intrusion detection systems.

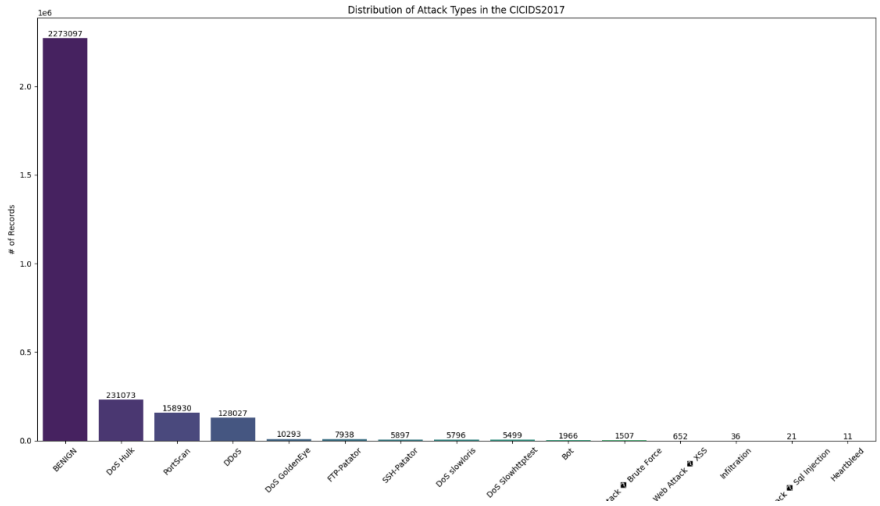
CICIDS2017

The CICIDS2017 dataset was developed by the Canadian Institute for Cybersecurity to facilitate network security research (Sharafaldin et al., 2018). Although not specifically created for IoT devices, it can be adapted for IoT IDS studies. This dataset is commonly utilized in general IDS and network security research. The dataset comprises a total of 2,830,540 records, which were generated in a simulated network environment over seven days, encompassing various attack types and normal traffic patterns that reflect real-world conditions.

The CICIDS2017 dataset includes 83 features designed to characterize each traffic instance. These features provide detailed insights into network traffic, incorporating various parameters such as protocol information, connection duration, and packet size.

An analysis of the dataset reveals the distribution of attack and benign (non-attack) traffic across different classes, as illustrated in Figure 5.

Figure 5
Distribution of Attack Types in The CICIDS-2017 Dataset



As illustrated in Figure 5, the CICIDS2017 dataset exhibits a significant imbalance, with most traffic records classified as benign (non-attack) traffic. When focusing solely on attack traffic, it is evident that the most prevalent attack categories are Denial of Service (DoS) Hulk, Port Scan, and Distributed DoS, while the Heartbleed attack demonstrates the lowest frequency. This imbalance presents challenges for the model in learning to identify such low-frequency attacks effectively.

In scenarios where there is a substantial class imbalance, models tend to favor learning the majority class (e.g., “Benign”). This tendency can hinder the accurate classification of the minority class. Additionally, classes with a low number of samples are at risk of overfitting, as the model may memorize these few instances instead of generalizing them. Consequently, while the model may perform well on the limited examples from these minority classes, it could fail when confronted with new and previously unseen instances.

TON_IoT

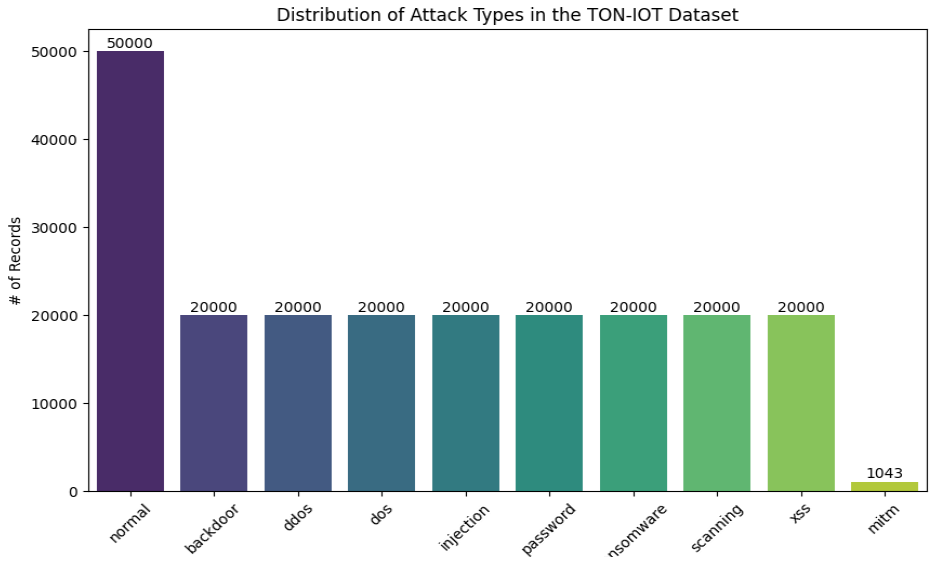
The ToN-IoT dataset, developed by UNSW Canberra, represents a next-generation resource utilized for evaluating security applications within Industry 4.0, the IoT, and Industrial IoT (IIoT) networks. This dataset serves as a benchmark for testing the accuracy and effectiveness of various cybersecurity applications, including intrusion detection systems, threat intelligence, malware detection, fraud detection, and digital forensics (Alsaedi et al., 2020; Booij et al., 2022; Moustafa, 2021).

The ToN-IoT dataset comprises a total of 12 distinct features for each record. These features encapsulate a range of parameters related to network traffic, facilitating the analysis of communication dynamics among devices. With approximately 2.8 million records, the dataset encompasses both normal traffic patterns and a variety of attack scenarios. The ToN-IoT dataset was created in a laboratory setting, where data was collected in a simulated environment that closely mirrors real-world conditions, utilizing various IoT devices (Ashraf et al., 2021; Moustafa, 2019; Moustafa, Ahmed, et al., 2020; Moustafa, Keshky, et al., 2020).

In this study, analyses were conducted using the Train Test Network dataset presented by the authors. This dataset consists of 211,043 records and encompasses 42 features. Notably, it includes not only normal traffic but also nine categories of attacks. The distribution of traffic classes is illustrated in Figure 6.

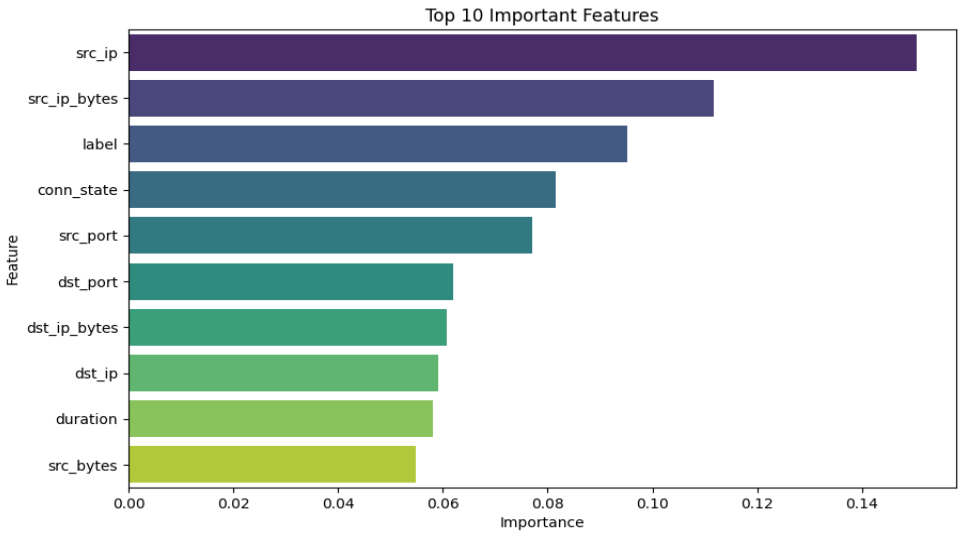
Figure 6

The Relatively Balanced Distribution of Attack Traffic in the TON-IoT Dataset.



In the dataset designed for training and testing, it has been observed that the distribution of normal traffic and all attack types, except for Man-in-the-Middle attacks, is balanced. The ten most effective features for classification purposes are illustrated in Figure 7.

Figure 7
The Top 10 Most Effective Features for Classification in the ToN-IoT Dataset.



According to the analysis, the most decisive features are primarily the source IP address, followed by features indicating the number of bytes associated with the source IP address. The TON-IoT datasets comprise telemetry data collected from IoT and IIoT sensors, as well as data obtained from Windows and Ubuntu operating systems, in addition to network traffic data. These datasets are collected from realistic and large-scale networks and encompass a variety of normal and cyber-attack events.

BoT-IoT

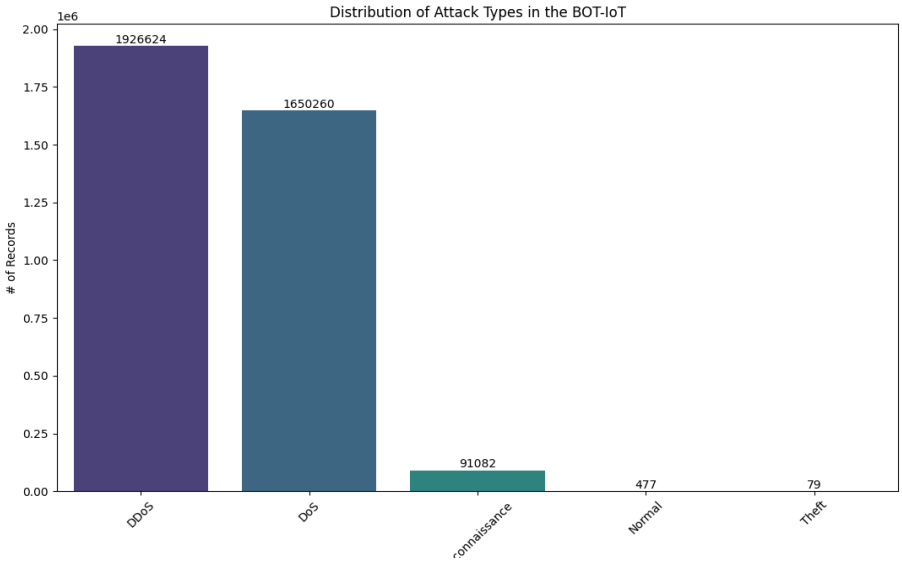
The Bot-IoT dataset, developed by UNSW Canberra's Cyber Range Laboratory in 2020, is recognized as a resource in the field of IoT security. This dataset is designed to simulate a realistic network environment that integrates normal and botnet traffic. It is primarily utilized for the detection and analysis of botnet attacks targeting IoT devices. The dataset is available in various formats, including original pcap files, argus files, and CSV files (Koroniotis et al., 2017, 2019; Koroniotis, Moustafa, Schiliro, et al., 2020; Koroniotis, Moustafa, & Sitnikova, 2020).

The Bot-IoT dataset comprises approximately 1.2 million records, with each record containing 15 distinct features. These records represent various attack scenarios and examples of normal traffic. The dataset was generated in a laboratory setting through a simulated scenario involving the use of multiple IoT devices in a real-world network environment, where botnet attacks were executed to collect data. This approach provides valuable insights into analyzing real-world cyber-attacks.

In the training dataset used in this study, there are 3,668,522 records and 46 features. The dataset is categorized into two primary target labels: normal and attack. The attack labels are further divided into subcategories that specifically represent different types of botnet attacks. The distribution of traffic by main categories is illustrated in Figure 8.

Figure 8

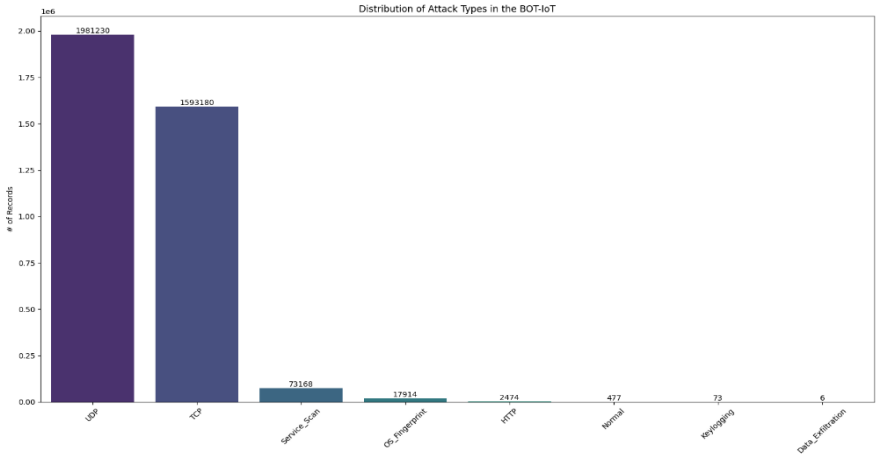
The Distribution of Normal and Attack Traffic in the BoT-IoT Dataset



The Bot-IoT dataset exhibits a significant imbalance, particularly in the representation of normal traffic compared to DDoS and DoS traffic. The volume of normal traffic records is markedly lower, which can adversely affect the model's ability to learn and generalize from the data. Similarly, the category of Theft traffic is also underrepresented, presenting challenges for effective classification. Furthermore, the dataset includes subcategories for the various types of traffic attacks. Figure 9 illustrates the distribution of attack traffic across these subcategories, highlighting the disparities in representation among different attack types. Such imbalances necessitate careful consideration during the training of models, as they may lead to biased learning outcomes favoring the more prevalent categories.

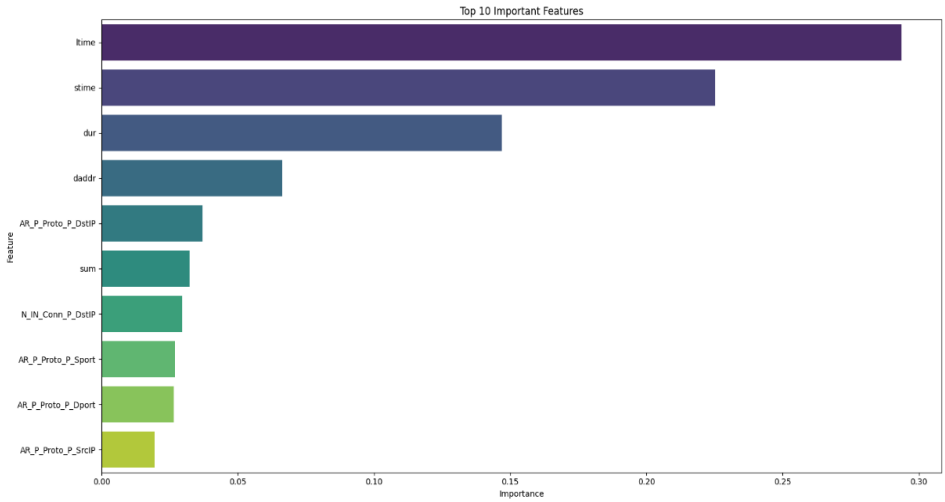
Figure 9

The Distribution of Subcategories in the BoT-IoT Dataset.



Similarly, it can be observed that the subcategories within the dataset do not exhibit a balanced distribution. The scarcity of non-attack normal traffic records is particularly concerning, as this deficiency is likely to adversely impact classification accuracy and increase the rate of false positives. Figure 10 presents the most effective features identified in the classification of the traffic types within the dataset. These features play a crucial role in enhancing the model's performance and its ability to accurately differentiate between normal and attack traffic.

Figure 10
The Top 10 Most Effective Features in Subcategory Classification in BOT-IoT.



In the context of classification, the most effective features identified are the Local Time (ltime) and System Time (stime). The ltime and stime features represent the local time at which an event occurs and the system time, respectively. These features provide critical temporal information regarding the occurrence of events, which is particularly beneficial in time series analyses. The integration of these two timestamps facilitates the synchronization of events between records from different systems, ensuring temporal alignment in the analysis of network traffic or events among IoT devices. Moreover, they play a critical role in time series modeling, enhancing the ability to discern patterns over time.

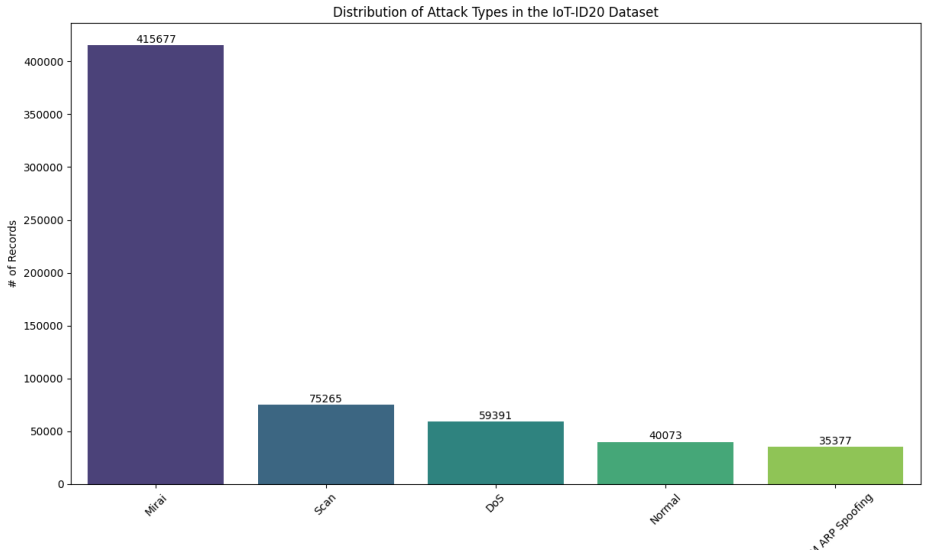
The Bot-IoT dataset is invaluable for research in areas such as machine learning and intrusion detection systems. It is frequently utilized to investigate the diversity and evolution of attacks targeting IoT devices, making it a preferred choice in relevant studies.

IoTID20

This dataset was created in 2020, focusing on attacks against IoT devices and addressing current IoT threats. It has become a frequently preferred resource in recent IoT security research due to its inclusion of up-to-date attack types, making contributions to the field (Surya & Shanthi, 2023).

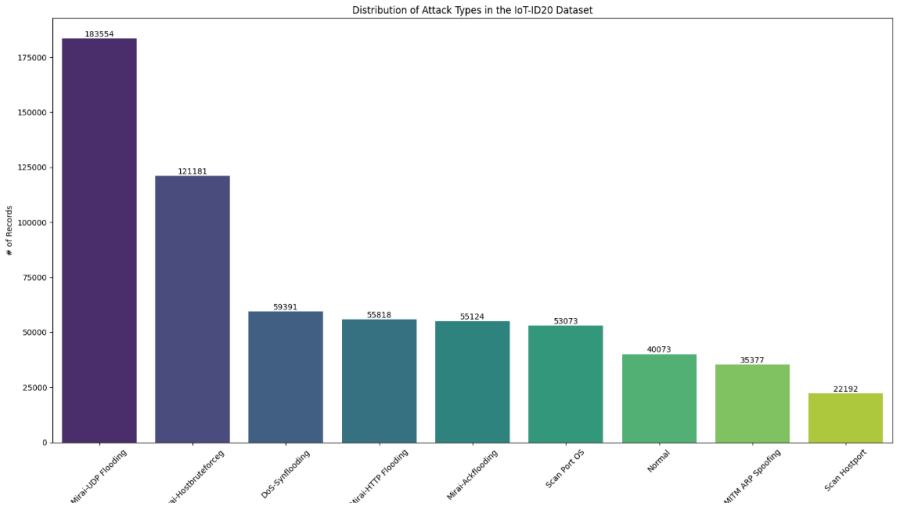
The dataset contains approximately 625,873 records, with a total of 86 different features for each record. The attacks are classified at two levels in the dataset: category and subcategory. The distribution of attack categories and normal traffic is presented in Figure 11.

Figure 11
The Distribution of Traffic Labels in the IoT-ID20 Dataset.



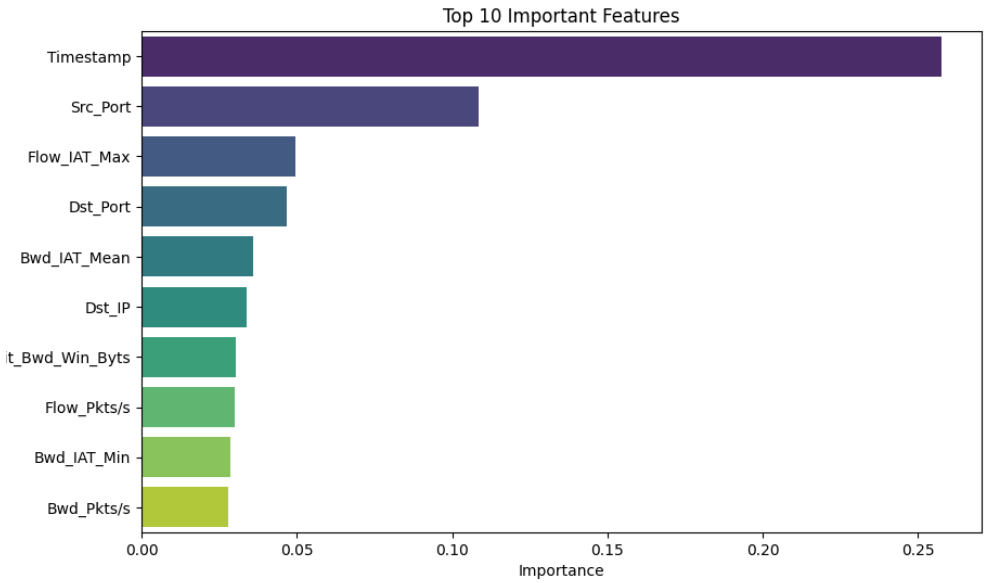
The distribution of the analysis conducted for the subcategories is shown in Figure 12.

Figure 12
The Distribution of Traffic Subcategories in the IoT-ID20 Dataset.



It can be observed that the traffic in both levels of categories is imbalanced. However, it is important that relatively recent attacks, such as the Mirai attack, are included in the dataset. The most effective features are illustrated in Figure 13.

Figure 13
The Top 10 Features in the classification of Attacks

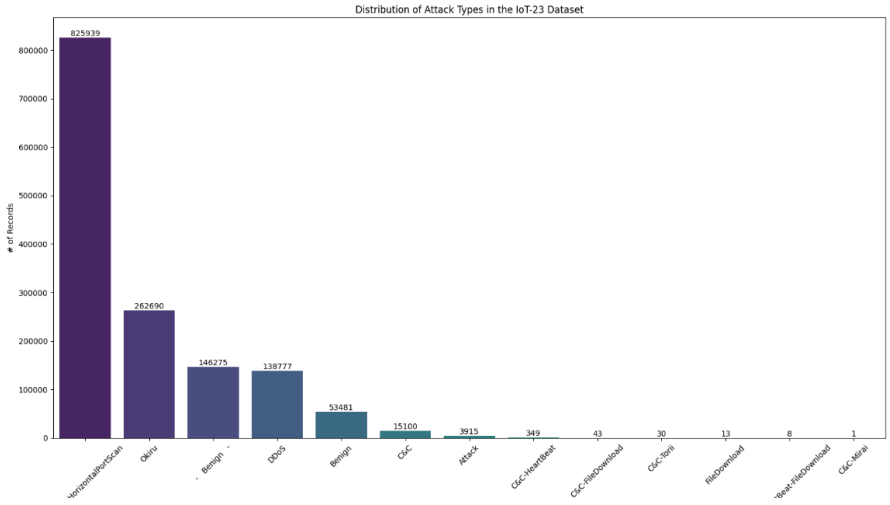


The most influential feature in classification is the timestamp, followed by the source port number.

IoT 23

This dataset has been developed to investigate malicious activities observed in IoT devices (Sharma & Babbar, 2024) (Alfares & Banimelhem, 2024). It contains 1,446,621 records and 28 different features. The dataset was generated in a laboratory environment, where real IoT devices were used to simulate specific attack scenarios and normal communication patterns. The IoT-23 dataset typically exhibits an imbalanced structure. The distribution of traffic within the dataset is illustrated in Figure 14.

Figure 14
The Traffic Classes of the IoT23 Dataset.



The dataset focuses on malware, particularly attacks such as PortScan, Command and Control (C&C), and Distributed Denial of Service attacks. It is utilized for research in

IoT security, with a specific emphasis on malware detection efforts.

General Analysis of Datasets

This section will address the general characteristics and analysis of datasets used in the field of IoT security. Datasets are critical for evaluating the effectiveness of cybersecurity applications and serve as tools for understanding different types of attacks. The differences between balanced and unbalanced datasets are important factors that affect the success of machine learning models. In this context, aspects such as the scope, features, and labeling systems of IoT security datasets will be examined

Balanced datasets are those in which each class has an equal number of examples. In contrast, unbalanced datasets contain some classes with more or fewer examples than others. From the perspective of IoT security, balanced datasets enhance the model's ability to learn each class, while unbalanced datasets may reflect more realistic scenarios.

IoT security datasets vary based on their application areas. Some datasets focus on a specific type of attack, while others include general network traffic or specific devices. For instance, the BoT-IoT dataset targets botnet attacks, whereas the TON_IoT dataset offers a wide range of attack scenarios. Different machine learning techniques may be more effective on different types of datasets. Labeled datasets are used for supervised learning, while less labeled or unlabeled datasets are utilized for unsupervised learning. For example, the NSL-KDD dataset is suitable for supervised learning models.

The sizes of datasets vary based on the number of examples and the number of features they contain. Larger datasets provide opportunities for more comprehensive analyses, while smaller datasets can be processed more quickly. The number of features in IoT security datasets directly impacts model performance. More features enable the model to learn more information, but unnecessary features can increase the model's complexity. Feature engineering plays a crucial role in these datasets. Label types in IoT security datasets can vary. Some datasets contain clear labels indicating a specific type of attack, while others may use more general or ambiguous labels. This situation can affect the accuracy and precision of machine learning models.

Research conducted using IoT datasets provides insights into the field of IoT security. These publications are valuable for identifying new threats, developing new security solutions, and improving existing methods. Table 1 presents a summary view of the datasets addressed in this study.

Table 1

Overview of Various IoT Security Datasets

Dataset	Record Count	Attack Category Count	Features Count	Categories
NSL KDD	125,973	22	42	neptune, satan, ipsweep, portsweep, smurf, nmap, back, teardrop, warezclient, pod, guess_passwd, buffer_overflow, warezmaster, land, imap, rootkit, loadmodule, ftp_write, multihop, phf, perl, spy
CICIDS-2017	2,600,000	14	79	DoS Hulk, PortScan, DDoS, DoS GoldenEye, FTP-Patator, SSH-Patator, DoS slowloris, DoS Slowhttptest, Bot, Web Attack - Brute Force, Web Attack - XSS, Infiltration, Web Attack - Sql Injection, Heartbleed

Dataset	Record Count	Attack Category Count	Features Count	Categories
UNSW-NB15	2,540,044	9	49	Generic, Exploits, Fuzzers, DoS, Reconnaissance, Analysis, Backdoor, Shellcode, Worms
TON-IoT	1,209,013	9	43	Backdoor, DDoS, DoS, Injection, Password, Ransomware, Scanning,XSS, MiTM
BoT-IoT	1,140,000	4	40	DDoS, DoS, Reconnaissance, Theft
IoT-ID20	1,049,045	8	22	Mirai-UDP Flooding, Mirai-Hostbruteforceg, DoS-Synflooding, Mirai-HTTP Flooding, Mirai-Ackflooding, Scan Port OS, MITM ARP Spoofing, Scan Hostport
IoT-23	75,000	12	23	PortScan, Okiru, Benign, DDoS, C&C, Attack, C&C-HeartBeat, C&C-FileDownload, C&C-Torii, FileDownload, C&C-HeartBeat-FileDownload, C&C-Mirai

NSL KDD is a classic intrusion detection dataset that is widely used in the field of cybersecurity. The diversity of categories allows for the learning of various types of attacks during the modeling process. However, the lower number of records compared to other datasets may offer less diversity and a less realistic environment. CICIDS-2017 has a large number of records and represents modern types of attacks. The high number of features allows for the development of more complex and effective models. UNSW-NB15 has a wide number of records and a sufficient set of features, making it suitable for analyzing various attacks. The diversity of attack categories is beneficial for understanding overall security threats. TON-IoT is a dataset specifically designed to examine threats targeting IoT devices. The variety of attacks is crucial for understanding security threats in the IoT environment. BoT-IoT focuses on a specific type of attack and provides data on fundamental IoT threats. However, the limited number of attack categories may restrict the coverage of various attack scenarios. IoT-ID20 represents Mirai-based attacks commonly observed in IoT environments. The specialization of categories is useful for focusing on specific threats. IoT-23, despite having fewer records, includes various attack categories, which is beneficial for understanding different types of attacks.

In general evaluation, CICIDS-2017 and UNSW-NB15 datasets provide a larger number of records, allowing for more comprehensive analyses compared to other datasets. The diversity of attack categories is of great importance for model development and attack detection. Particularly, IoT datasets stand out for containing up-to-date and realistic attack scenarios. Additionally, the high number of features increases the complexity of the dataset and facilitates the creation of more effective machine-learning models.

Conclusion

In this study, we explored the complexities and nuances of IoT security through the analysis of various datasets designed to detect and classify malicious activities. The hypothesis posited that the choice and characteristics of these datasets influence the

performance of machine learning models in identifying security threats within IoT environments. Our findings confirm this hypothesis, demonstrating that datasets with balanced classes and a diverse range of attack types lead to more accurate and reliable models.

The analysis revealed that while datasets like CICIDS-2017 and UNSW-NB15 provide extensive records for training models, the inherent imbalances in certain datasets, such as BoT-IoT and IoT-23, may hinder the detection capabilities of models, particularly for less frequent attack types. Additionally, the effectiveness of certain features, especially time-related attributes, underscores the importance of selecting relevant characteristics that contribute to model performance.

As IoT threats continue to evolve, the integration of synthetic datasets and the continuous refinement of existing datasets will be essential for enhancing the resilience of IoT systems against emerging security challenges. Future research should focus on developing more comprehensive datasets that reflect real-world scenarios, thus enabling researchers to train models that are not only effective but also robust in the face of diverse attack strategies.

In summary, the hypothesis that the characteristics of IoT security datasets impact the efficacy of machine learning models has been substantiated. This study highlights the critical role of dataset selection in advancing IoT security measures and sets the stage for future explorations aimed at fortifying the integrity of IoT environments.

Future efforts could focus on generating datasets with a greater variety of attack types and more balanced classes, which would address the limitations of existing datasets. Synthetic data generation could be further explored to simulate real-world attack scenarios, providing more comprehensive training data for machine learning models.

Further investigation into feature selection techniques could improve model performance by identifying the most relevant features for each attack type. This might involve applying techniques like Principal Component Analysis or Recursive Feature Elimination (RFE) to reduce model complexity while maintaining accuracy. Establishing standardized benchmarks for IoT security datasets would support comparative analysis of models across different datasets. Such benchmarks could help researchers evaluate model robustness and performance more consistently, fostering advancements in the field.

References

Aggarwal, P., & Sharma, S. K. (2015). Analysis of KDD Dataset Attributes—Class wise for Intrusion Detection. *Procedia Computer Science*, *57*, 842–851. https://doi.org/10.1016/j.procs.2015.07.490

Alfares, H., & Banimelhem, O. (2024). Comparative Analysis of Machine Learning Techniques for Handling Imbalance in IoT-23 Dataset for Intrusion Detection Systems. *2024 11th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, 112–119. https://doi.org/10.1109/IOTSMS62296.2024.10710296

Alosaimi, S., & Almutairi, S. M. (2023). An Intrusion Detection System Using BoT-IoT. *Applied Sciences*, *13*(9), 5427. https://doi.org/10.3390/app13095427

Alrayes, F. S., Zakariah, M., Amin, S. U., Iqbal Khan, Z., & Helal, M. (2024). Intrusion Detection in IoT Systems Using Denoising Autoencoder. *IEEE Access*, *12*, 122401–122425. https://doi.org/10.1109/ACCESS.2024.3451726

Alsaedi, A., Moustafa, N., Tari, Z., Mahmood, A., & Anwar, A. (2020). TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems. *IEEE Access*, *8*, 165130–165150. https://doi.org/10.1109/ACCESS.2020.3022862

Ashraf, J., Keshk, M., Moustafa, N., Abdel-Basset, M., Khurshid, H., Bakhshi, A. D., & Mostafa, R. R. (2021). IoTBoT-IDS: A novel statistical learning-enabled botnet detection framework for protecting networks of smart cities. *Sustainable Cities and Society*, *72*, 103041. https://doi.org/10.1016/j.scs.2021.103041

Bansal, K., & Singhrova, A. (2023). Review on intrusion detection system for IoT/IIoT -brief study. *Multimedia Tools and Applications*, *83*(8), 23083–23108. https://doi.org/10.1007/s11042-023-16395-6

Booij, T. M., Chiscop, I., Meeuwissen, E., Moustafa, N., & Hartog, F. T. H. D. (2022). ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets. *IEEE Internet of Things Journal*, *9*(1), 485–496. https://doi.org/10.1109/JIOT.2021.3085194

Guerra, J. L., Catania, C., & Veas, E. (2022). Datasets are not enough: Challenges in labeling network traffic. *Computers & Security*, *120*, 102810. https://doi.org/10.1016/j.cose.2022.102810

Hota, H. S., & Shrivas, A. K. (2014). Decision Tree Techniques Applied on NSL-KDD Data and Its Comparison with Various Feature Selection Techniques. In M. Kumar Kundu, D. P. Mohapatra, A. Konar, & A. Chakraborty (Eds.), *Advanced Computing, Networking and Informatics- Volume 1* (Vol. 27, pp. 205–211). Springer International Publishing. https://doi.org/10.1007/978-3-319-07353-8_24

Karamollaoğlu, H., Doğru, İ. A., & Yücedağ, İ. (2024). An Efficient Deep Learningbased Intrusion Detection System for Internet of Things Networks with Hybrid Feature Reduction and Data Balancing Techniques. *Information Technology and Control*, *53*(1), 243–261. https://doi.org/10.5755/j01.itc.53.1.34933

Kaur, B., Dadkhah, S., Shoeleh, F., Neto, E. C. P., Xiong, P., Iqbal, S., Lamontagne, P., Ray, S., & Ghorbani, A. A. (2023). Internet of Things (IoT) security dataset evolution: Challenges and future directions. *Internet of Things*, *22*, 100780. https://doi.org/10.1016/j.iot.2023.100780

Koroniotis, N., Moustafa, N., Schiliro, F., Gauravaram, P., & Janicke, H. (2020). A Holistic Review of Cybersecurity and Reliability Perspectives in Smart Airports. *IEEE Access*, *8*, 209802–209834. https://doi.org/10.1109/ACCESS.2020.3036728

Koroniotis, N., Moustafa, N., & Sitnikova, E. (2020). A new network forensic framework based on deep learning for Internet of Things networks: A particle deep framework. *Future Generation Computer Systems*, *110*, 91–106. https://doi.org/10.1016/j.future.2020.03.042

Koroniotis, N., Moustafa, N., Sitnikova, E., & Slay, J. (2017). *Towards Developing Network forensic mechanism for Botnet Activities in the IoT based on Machine Learning Techniques* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1711.02825

Koroniotis, N., Moustafa, N., Sitnikova, E., & Turnbull, B. (2019). Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. *Future Generation Computer Systems*, *100*, 779–796. https://doi.org/10.1016/j.future.2019.05.041

Liang, P., Yang, L., Xiong, Z., Zhang, X., & Liu, G. (2024). Multilevel Intrusion Detection Based on Transformer and Wavelet Transform for IoT Data Security. *IEEE Internet of Things Journal*, *11*(15), 25613–25624. https://doi.org/10.1109/JIOT.2024.3369034

Maiwada, U. D., Imran, S. A., Danyaro, K. U., Janisar, A. A., Salameh, A., & Sarlan, A. B. (2024). Security Concerns of IoT Against DDoS in 5G Systems. *International Journal of Electrical Engineering and Computer Science*, *6*, 98–105. https://doi.

org/10.37394/232027.2024.6.11

Mishra, A. K., Rajput, K., Pandey, N. K., & Pathak, A. (2023). Comparative Analysis of Classification Algorithms Using Bot_IoT Dataset. *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, 1775–1780. https://doi.org/10.1109/ICSCNA58489.2023.10370699

Moustafa, N. (2019). *A Systemic IoT-Fog-Cloud Architecture for Big-Data Analytics and Cyber Security Systems: A Review of Fog Computing* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1906.01055

Moustafa, N. (2021). A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets. *Sustainable Cities and Society*, *72*, 102994. https://doi.org/10.1016/j.scs.2021.102994

Moustafa, N., Ahmed, M., & Ahmed, S. (2020). Data Analytics-Enabled Intrusion Detection: Evaluations of ToN_IoT Linux Datasets. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 727–735. https://doi.org/10.1109/TrustCom50675.2020.00100

Moustafa, N., Keshky, M., Debiez, E., & Janicke, H. (2020). Federated TON_IoT Windows Datasets for Evaluating AI-Based Security Applications. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 848–855. https://doi.org/10.1109/TrustCom50675.2020.00114

Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). *2015 Military Communications and Information Systems Conference (MilCIS)*, 1–6. https://doi.org/10.1109/MilCIS.2015.7348942

Moustafa, N., & Slay, J. (2016). The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set. *Information Security Journal: A Global Perspective*, *25*(1–3), 18–31. https://doi.org/10.1080/19393555.2015.1125974

Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., & Ghorbani, A. A. (2023). CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors*, *23*(13), 5941. https://doi.org/10.3390/s23135941

Ozdogan, E. (2024). A Comprehensive Analysis of the Machine Learning Algorithms in IoT IDS Systems. *IEEE Access*, *12*, 46785–46811. https://doi.org/10.1109/ACCESS.2024.3382539

Qing, Y., Liu, X., & Du, Y. (2024). Mitigating data imbalance to improve the generalizability in IoT DDoS detection tasks. *The Journal of Supercomputing*, *80*(7), 9935–9960. https://doi.org/10.1007/s11227-023-05829-5

Saadouni, R., Gherbi, C., Aliouat, Z., Harbi, Y., & Khacha, A. (2024). Intrusion detection systems for IoT based on bio-inspired and machine learning techniques: A systematic review of the literature. *Cluster Computing*, *27*(7), 8655–8681. https://doi.org/10.1007/s10586-024-04388-5

Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization: *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 108–116. https://doi.org/10.5220/0006639801080116

Sharma, A., & Babbar, H. (2024). Understanding IoT-23 Dataset: A Benchmark for IoT Security Analysis. *2023 4th International Conference on Intelligent Technologies (CONIT)*, 1–5. https://doi.org/10.1109/CONIT61985.2024.10627334

Surya, V., & Shanthi, C. (2023). Cross Model Verification of Intrusion Detection

System on IoT Using Convolutional Neural Network. *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)*, 1–12. https://doi.org/10.1109/ICTBIG59752.2023.10456135

Thana-Aksaneekorn, C., Kosolsombat, S., & Luangwiriya, T. (2024). Machine Learning Classification for Intrusion Detection Systems Using the NSL-KDD Dataset. *2024 IEEE International Conference on Cybernetics and Innovations (ICCI)*, 1–6. https://doi.org/10.1109/ICCI60780.2024.10532265

Türk, F. (2023). Analysis of Intrusion Detection Systems in UNSW-NB15 and NSL-KDD Datasets with Machine Learning Algorithms. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *12*(2), 465–477. https://doi.org/10.17798/bitlisfen.1240469

Vibhute, A. D., Patil, C. H., Mane, A. V., & Kale, K. V. (2024). Towards Detection of Network Anomalies using Machine Learning Algorithms on the NSL-KDD Benchmark Datasets. *Procedia Computer Science*, *233*, 960–969. https://doi.org/10.1016/j.procs.2024.03.285

Zafar Iqbal Khan, Mohammad Mazhar Afzal, & Khurram Naim Shamsi. (2024). A Comprehensive Study on CIC-IDS2017 Dataset for Intrusion Detection Systems. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, *2*(02), 254–260. https://doi.org/10.47392/IRJAEH.2024.0041

Zoghi, Z., & Serpen, G. (2024). UNSW-NB15 computer security dataset: Analysis through visualization. *SECURITY AND PRIVACY*, *7*(1), e331. https://doi.org/10.1002/spy2.331

About the Authors

Erdal ÖZDOĞAN is an academic affiliated with Management Information Systems, Uludag University, Bursa, Turkiye. He received a Ph.D. degree from the department of Information Systems at Gazi University. He specializes in IoT (Internet of Things), cybersecurity, and network systems. He has a PhD in information sciences and his research interests include IoT, cybersecurity, networks, and artificial intelligence. He has published several papers and book chapters on these topics. He also teaches network security, cryptography, machine learning, and system analysis and design courses.
E-mail: erdalozdogan@uludag.edu.tr , **ORCID:** 0000-0002-3339-0493

Onur CERAN is an academic and researcher affiliated with Gazi University in Türkiye. He received a Ph.D. degree from Gazi University. His work primarily focuses on information security and computer and instructional technologies. He also teaches courses on network, cyber security, ethical hacking, incident response, forensics and IT law courses. He has contributed to various studies and publications on these topics.
E-mail: onur.ceran@gazi.edu.tr, **ORCID:** 0000-0003-2147-0506

Similarity Index

The similarity index obtained from the plagiarism software for this book chapter is 4%.