A Structural Analysis of Intrusion Detection System Datasets and Practical Implications

Luiz Henrique B. A. da Silva[1], José R. de Souza Silva[1], Caio B. B. de Souza[1], Marcos R. M. Falcão, Andson M. Balieiro[1]

Centro de Informática, Universidade Federal de Pernambuco (UFPE), Recife, Brazil
{lhbas, jrss, cbbs, mrmf, amb4}@cin.ufpe.br

Abstract. Intrusion Detection Systems (IDS) rely heavily on the quality and representativeness of the datasets used for training and evaluation. However, many publicly available IDS datasets present significant challenges, such as extreme class imbalance, redundant records, and inconsistencies in feature representation. This paper presents an exploratory and comparative analysis of multiple IDS datasets, focusing on data quality, feature characterization, class distribution, and inherent limitations that may impact machine learning-based detection approaches. Our findings highlight critical issues that must be considered before applying these datasets in real-world IDS studies.

Keywords: Intrusion Detection Systems · Dataset Analysis · Class Imbalance · Data Quality · Cybersecurity

1 Introduction

Intrusion Detection Systems (IDS) are a fundamental component of modern cybersecurity infrastructures, playing a central role in identifying malicious activities across enterprise networks, Internet of Things (IoT) ecosystems, and cyber-physical environments. With the increasing adoption of machine learning (ML) and deep learning (DL) techniques, publicly available datasets have become essential for the development, evaluation, and benchmarking of detection models [1]. However, despite the growing number of released datasets, their structural properties differ substantially in terms of feature extraction methods, labeling conventions, attack taxonomies, traffic composition, and experimental design. These variations can significantly affect evaluation outcomes, yet they are frequently treated as neutral inputs within ML pipelines, without deeper examination of how dataset structure itself may influence performance [2].

Several recent studies have examined the growing dependence of Intrusion Detection Systems on machine learning techniques and publicly available datasets for training and benchmarking. Recent surveys highlight how modern IDS research increasingly relies on deep learning models and standardized datasets to evaluate detection performance and reproducibility [3,4]. In addition, analyses of network intrusion detection datasets report substantial heterogeneity in data

collection methodologies, feature representations, labeling conventions, and experimental configurations across commonly used benchmarks [5,6]. Other works have focused on developing and evaluating new datasets designed to address limitations observed in earlier benchmarks, emphasizing the importance of realistic traffic generation and comprehensive attack coverage [7,8]. Furthermore, studies on dataset assessment frameworks highlights that dataset quality, representativeness, and statistical properties such as class imbalance and redundancy can significantly influence IDS evaluation outcomes [9,10]. Finally, empirical studies using deep learning-based IDS demonstrate how dataset structure, feature composition, and class distribution directly affect model performance, generalization, and cross-dataset evaluation reliability [11,12].

Rather than assuming that datasets are inherently optimized for machine learning purposes, this work advocates for a systematic structural examination prior to model deployment, as concerns regarding dataset heterogeneity and evaluation inconsistencies have been highlighted in prior studies [13]. We present an exploratory analysis of multiple IDS datasets, focusing on their feature composition, labeling schemes, class distribution patterns, duplication behavior, and other statistical properties that may affect learning dynamics, not classifying them as deficiencies, but describing how dataset structure influences model behavior, reproducibility, and cross-dataset generalization [14].

This paper is organized as follows. Section 2 describes the proposed exploratory methodology and database-driven analysis pipeline. Section 3 introduces the evaluated datasets and their contextual characteristics. Section 4 presents the main findings obtained from the structural and statistical analysis. Finally, Section 5 concludes this paper and outlines future research directions.

2 Methodology

This study follows a data-driven exploratory approach centered on large-scale Intrusion Detection System (IDS) datasets. Given the substantial volume of records in several of these datasets, in-memory processing with conventional dataframes proved impractical. To address this, a database-oriented analysis pipeline was designed to support scalable and reproducible processing across all datasets, as illustrated in Figure 1.

2.1 Data Ingestion and Storage

The data ingestion and storage phase is a critical foundation of the proposed methodology. Given the scale of modern IDS datasets, which frequently contain several records, efficient storage and query mechanisms are necessary to ensure feasibility and reproducibility of the analysis. This stage aims to provide a scalable infrastructure capable of supporting large-volume statistical computations without memory bottlenecks. The type mapping strategy does not affect the

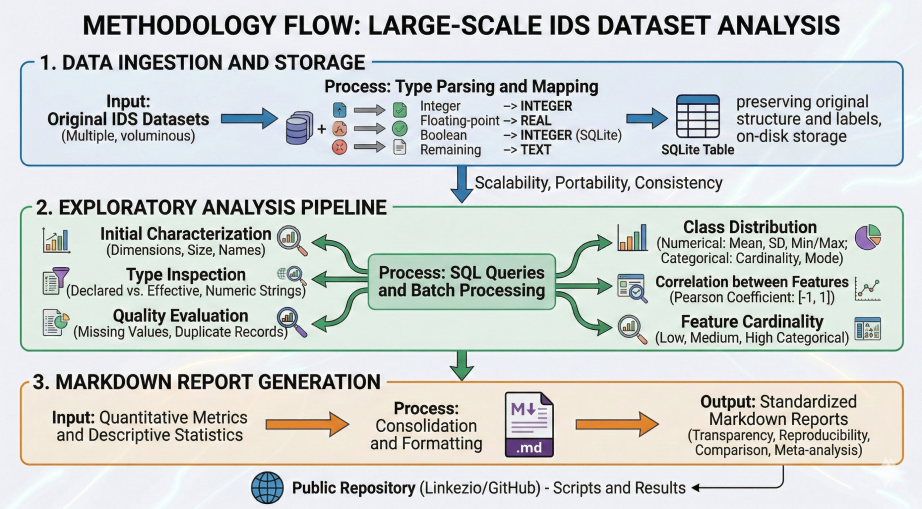

Fig. 1: Methodology Diagram

analytical objectives of this study, as it preserves the original data semantics. The script[1] used as the data processing pipeline is publicly available.

Each dataset was parsed and stored as a single SQLite[2]table, preserving the original feature structure and labels. SQLite was selected due to its lightweight nature, portability, and ability to efficiently perform aggregations, distinct counts, and statistical queries directly on disk. This approach enabled the analysis of very large datasets using commodity hardware without requiring distributed computing frameworks. During the ingestion process, dataset features were automatically mapped from their original data types to SQLite-compatible types to ensure consistency and simplify database creation. Integer-like features were mapped to `INTEGER`, floating-point features to `REAL`, boolean features to `INTEGER` since SQLite does not provide a native boolean type, and all remaining or ambiguous types were mapped to `TEXT`.

2.2 Exploratory Analysis Pipeline

The exploratory analysis pipeline constitutes the core analytical stage of this work. Its objective is to systematically extract structural, statistical, and distributional properties from each dataset in a consistent and reproducible manner. This phase is essential for identifying hidden structural artifacts such as imbalance severity, redundancy levels, feature inconsistencies, and statistical anomalies that may significantly influence machine learning behavior. The output of

[1] `https://github.com/luiz-linkezio/A-Hierarchical-Solution-based-on-MEC-and-IoT-for-Cyberattack-Detection-in-6G-Networks/blob/main/notebooks/database_creation.ipynb`

[2] `https://www.sqlite.org/`

this stage consists of a comprehensive set of quantitative metrics and descriptive statistics, which serve as the basis for the comparative discussion presented in Section 4.

Following data ingestion, an exploratory analysis pipeline was applied to each dataset. This pipeline was designed to extract both structural and statistical characteristics directly from the databases, ensuring consistent analysis across all datasets. To do so, the following activities were performed. All computations were conducted using SQL queries and controlled batch processing, ensuring that the analysis remained scalable regardless of dataset size. The full analysis script[3] pipeline is available in a public repository.

- **Initial Dataset Characterization**: measurement of dataset dimensions (number of records and features), database storage size, average record size, and feature naming consistency.
- **Data Type Inspection**: identification of declared and effective data types, with special attention to features stored as strings despite representing numeric quantities.
- **Data Quality Assessment**: evaluation of missing values, completeness ratio, and duplicate records. Duplicate analysis was conducted using full-row comparisons to quantify redundancy levels within each dataset.
- **Descriptive Statistics**: computation of summary statistics for numeric-like features (mean, standard deviation, minimum, and maximum values) and frequency-based statistics for categorical features, including cardinality, mode, and dominance ratios.
- **Numerical features**: attributes whose values belong to the set of integers or real numbers, allowing the application of arithmetic and statistical operations such as mean, standard deviation, and variance.
- **Categorical features**: attributes represented as discrete textual labels or symbolic values. Even when such features contain numeric characters, they do not represent quantities with intrinsic mathematical meaning and therefore are not suitable for direct arithmetic operations.
- **Class Distribution Analysis**: examination of label distributions, number of classes, majority-to-minority ratios, and imbalance severity at both binary and multi-class labeling levels.
- **Correlation Between Features**: Correlation quantifies the degree of linear association between two numerical variables. It evaluates both the strength and the direction of their relationship. The Pearson correlation coefficient was adopted, which assumes values in the interval [-1, 1], where 1 represents perfect positive linear dependence, -1 means perfect negative linear dependence, and 0 indicates absence of linear correlation.
- **Feature Cardinality Analysis**: classification of categorical features into low, medium, and high cardinality groups based on the proportion of unique values relative to dataset size.

[3] `https://github.com/luiz-linkezio/A-Hierarchical-Solution-based-on-MEC-and-IoT-for-Cyberattack-Detection-in-6G-Networks/blob/main/notebooks/dataset_analysis.ipynb`

2.3 Markdown Report Generation

To ensure transparency, reproducibility, and independent verification, a standardized reporting stage was incorporated into the methodology, generating structured Markdown reports for each dataset that consolidate all computed statistics, such as dataset dimensions, storage size, data types, missing values, duplicates, descriptive metrics, class distribution, and feature cardinality. This approach transforms raw analysis results into clear and comparable artifacts, while all scripts used to execute the methodology and produce the reports are publicly available to facilitate replication and further studies.

All generated reports are publicly available in a dedicated folder[4] of the project repository.

3 Datasets

This section presents five IDS datasets selected to be analyzed by applying the methodology defined in Section 2. The datasets were selected, considering the following criteria: (i) publication year equal to or later than 2023, ensuring up-to-date traffic and attack scenarios; (ii) public availability of structured and/or raw data in CSV and/or PCAP formats to enable reproducible analysis; and (iii) data volume. Additionally, the adoption by the academia of each dataset were assessed through citation counts obtained from Google Scholar (March 2026), revealing heterogeneous levels of adoption: CIC IoT 2023 (1,062 citations), CIC IoT-DIAD 2024 (39), CIC BCCC NRC IoMT 2024 (54), CIC APT IIoT 2024 (6), and CIC IIoT 2025 (3). Some datasets specifically target Industrial Internet of Things (IIoT) environments. Although all datasets are publicly distributed by the Canadian Institute for Cybersecurity (CIC), they do not represent a single homogeneous data source. Each dataset was generated under different experimental setups, network configurations, time periods, and attack scenarios. Therefore, they provide diverse and complementary perspectives for IDS evaluation rather than redundant samples from a single environment.

3.1 CIC Advanced Persistent Threat Dataset for Industrial IoT 2024 (CIC APT IIoT 2024)

This dataset[5] supports Advanced Persistent Threat (APT) detection in Industrial IoT environments. It was generated in a hybrid industrial IoT testbed, integrating simulated and real components, such as Network Simulator 3 (NS3), virtual machines, Raspberry Pi devices, IoT sensors, and OpenPLC.

It provides provenance logs and network traffic captures. The provenance data models Process and Artifact entities, while the network traffic is available

[4] https://github.com/luiz-linkezio/A-Hierarchical-Solution-based-on-MEC-and-IoT-for-Cyberattack-Detection-in-6G-Networks/tree/main/results/dataset_analysis

[5] https://www.unb.ca/cic/datasets/iiot-dataset-2024.html

in PCAP format and can be converted into structured CSV features. More than 20 attack techniques are included, organized according to MITRE ATT&CK tactics, enabling both provenance-based and network-based intrusion detection research.

3.2 A Real-Time Sensor-Based Benchmark Dataset for Attack Analysis in IIoT with Multi-Objective Feature Selection (CIC IIoT 2025)

This dataset[6] provides a real-time IIoT benchmark combining synchronized sensor measurements and network traffic data. It was collected from a laboratory testbed with 40 interconnected devices organized into IoT, Network, Edge, Cloud, and Attacker layers. It includes 50 attack types across seven categories, supporting both binary and multi-class detection. The dataset also introduces a multi-objective feature selection framework for efficient anomaly detection under resource constraints.

3.3 A Real-time Dataset and Benchmark for Large-scale Attacks in IoT Environment (CIC IoT 2023)

This large-scale IoT benchmark dataset[7] was collected from 105 interconnected IoT devices in a controlled laboratory environment. A total of 33 real-time attacks were executed across seven categories, including DDoS, DoS, Reconnaissance, Brute Force, Spoofing, and Mirai. The dataset comprises raw PCAP files and flow-based CSV features extracted for ML and DL evaluation, supporting traffic classification and intrusion detection research.

3.4 A Dual-Function Dataset for IoT Device Identification and Anomaly Detection (CIC IoT DIAD 2024)

This dataset[8] supports both Device Identification (DI) and Anomaly Detection (AD) in IoT environments. It was collected from 105 IoT devices under 33 attack scenarios grouped into seven categories. It combines packet-based and flow-based feature extraction, enabling both device fingerprinting and attack detection, and is provided in labeled CSV format for supervised ML and DL applications.

3.5 A Benchmark Dataset for Intrusion Detection in Internet of Medical Things Environments (CIC BCCC NRC IoMT 2024)

This dataset[9] targets intrusion detection in IoMT environments. It was collected from a healthcare-oriented testbed including medical devices, servers, and

[6] https://www.unb.ca/cic/datasets/iiot-dataset-2025.html
[7] https://www.unb.ca/cic/datasets/iotdataset-2023.html
[8] https://www.unb.ca/cic/datasets/iot-diad-2024.html
[9] https://www.unb.ca/cic/datasets/iomt-dataset-2024.html

attacker systems. It provides flow-based CSV features and includes 15 attack categories with both binary and multi-class labels. The dataset enables IDS evaluation in medical cyber-physical systems where reliability and latency constraints are critical.

4 Results

This section presents a concise summary of the most relevant findings for each dataset. The datasets were analyzed under three points: (1) data quality, including aspects such as data completeness, redundancy levels, and diversity; (2) features, referring to the types of traffic information they contain, and (3) class distribution, which characterizes the attack classes present in each dataset. Through this analysis, we highlight key dataset properties and discuss how they may influence the behavior and performance of ML/AI-based solutions for IDS. Next, a summary of the results of each dataset will be shown, followed by a final subsection making comparisons and reflections on the datasets. As previously mentioned, all result reports are available in a dedicated folder[10] within the repository.

4.1 CIC IoT 2023

CIC IoT 2023 contains 45,019,243 records and 40 features (39 numerical and one class label). The dataset presents high completeness, with only 1,498 missing values, but a very high redundancy level, as 53.34% of the records are exact duplicates. If not properly handled, this duplication may artificially inflate model performance. In addition, some attributes contain extreme or infinite values (e.g., `Rate`), indicating possible preprocessing inconsistencies. The features mainly correspond to flow-based traffic statistics, including packet counts, protocol indicators, and aggregated metrics, with strong correlations among size- and count-related attributes. The dataset includes 35 classes and exhibits extreme multi-class imbalance (over 765,000:1), with benign traffic representing only 2.34% of the data, which significantly challenges supervised learning and minority-class detection.

4.2 CIC APT-IIoT 2024

CIC APT-IIoT 2024 comprises 43,198,438 records and 70 attributes (67 numerical and 3 categorical). It is fully complete but contains exactly 50% duplicated records, reducing effective diversity. Most features describe traffic statistics, protocol indicators, and flow metrics, although some variables show zero variance or no protocol activity, limiting their discriminative utility. The dataset provides

[10] `https://github.com/luiz-linkezio/A-Hierarchical-Solution-based-on-MEC-and-IoT-for-Cyberattack-Detection-in-6G-Networks/tree/main/results/dataset_analysis`

three hierarchical label levels (`label`, `subLabel`, and `subLabelCat`), enabling binary and fine-grained multi-class analysis. However, it presents severe imbalance, from approximately 21,512:1 at the binary level to over 1,500,000:1 in more detailed classes, making conventional supervised multi-class learning particularly challenging.

4.3 CIC IoT-DIAD 2024

CIC IoT-DIAD 2024 includes 55,619,998 records and 84 attributes, with high completeness but 50.06% duplicate entries. The dataset follows a flow-based representation with detailed packet length, inter-arrival time, and window-based metrics. Some features exhibit zero variance, infinite rate values, and negative temporal measurements, suggesting potential collection or synchronization inconsistencies. From a labeling perspective, the dataset currently contains only a single class value (`NeedManualLabel`), meaning it is unlabeled and cannot be directly used for supervised IDS evaluation, being more suitable for unsupervised analysis or manual labeling.

4.4 CIC IIoT 2025

CIC IIoT 2025 contains 685,671 records and 94 features, combining 71 numerical and 23 categorical attributes. It stands out for its high structural quality, with no missing or duplicate records. The features capture aggregated traffic behavior and device-related information, with some variables showing high variance and heavy-tailed distributions. The dataset provides four hierarchical label columns (`label1`, `label2`, `label13`, `label4`), supporting multi-level intrusion analysis. While the binary level (`label1`) is relatively balanced, presenting a class distribution of 58.44% of benign and 41.56% of attack (1.41:1), class imbalance increases progressively at finer granularity levels (up to 4,262:1, when considering benign traffic and syn-flood attack). Tables 1 and 2 present the percentage of each class regarding `label12` and `label13`, respectively, which highlight the challenges for fine-grained multi-class classification.

4.5 CIC BCCC NRC IoMT 2024

Comprising 3,385,313 records and 85 features, the CIC BCCC NRC IoMT 2024 dataset includes 80 numerical attributes and 5 categorical ones (Flow ID, Src IP, Dst IP, Timestamp, and Attack Name). The dataset is fully complete, with no missing values detected. Duplicate analysis identified 109,180 repeated records (3.23%), leaving 3,276,133 unique samples after removal. Compared to other evaluated datasets, this low duplication level indicates better structural integrity.

From a feature perspective, the dataset adopts a flow-based representation similar to other CIC traffic collections, including packet statistics, inter-arrival times (IAT), header lengths, flag counters, bulk metrics, and window-size attributes. The Protocol feature shows zero variance (mean = 6.0, std = 0.0),

indicating that all traffic belongs to a single protocol type. Likewise, several bulk-related features (e.g., Fwd Bytes/Bulk Avg, Fwd Packet/Bulk Avg, and Fwd Bulk Rate Avg) also exhibit zero variance, limiting their discriminative value. In addition, some temporal metrics (e.g., Bwd IAT Min, Flow IAT Min, and Fwd IAT Min) present negative minimum values, suggesting possible timestamp inconsistencies or logging artifacts.

Regarding labeling, the dataset provides two columns: `Attack Name` (15 categories) and a binary `Label`. At the binary level, the distribution is highly imbalanced, with 3,352,693 samples labeled as 1 (99.04%) and 32,620 labeled as 0 (0.96%), resulting in an imbalance ratio of approximately 102.78:1. Although less severe than in some other datasets, this skew still requires mitigation strategies such as resampling, cost-sensitive learning, or anomaly-based approaches. The `Attack Name` attribute enables multi-class analysis with 15 attack categories; however, the strong dominance of DoS TCP Flood (62.24%) indicates additional imbalance at the multi-class level.

Table 1: CIC IIoT 2025 Dataset Multi-Class Distribution Regarding Label 2

Class	Count	Percent
benign	400,672	58.44%
recon	105,848	15.44%
dos	57,736	8.42%
ddos	56,692	8.27%
mitm	25,490	3.72%
malware	24,177	3.53%
web	9,040	1.32%
bruteforce	6,016	0.88%

Table 2: CIC IIoT 2025 Dataset Multi-Class Distribution Regarding Label 3

Class	Count	Percent
benign	400,672	58.44%
arp-spoofing	13,387	1.95%
mirai-udp-flood	12,869	1.88%
os-scan	12,491	1.82%
host-disc-tcp-ack-ping	12,432	1.81%
vuln-scan	12,417	1.81%
host-disc-tcp-syn-ping	12,367	1.80%
port-scan	12,340	1.80%
host-disc-tcp-syn-stealth	12,299	1.79%
Others	184,584	26.90%

4.6 Comparative Results

The comparative analysis reveals substantial structural heterogeneity across the evaluated datasets, reinforcing the importance of dataset-aware experimental design in IDS research. Rather than differing only in scale, the datasets exhibit marked variability in redundancy levels, labeling philosophy, hierarchical organization, and imbalance severity.

Table 3: Structural Properties of Evaluated IDS Datasets

Dataset	Records	Features	Missing (%)	Unique (%)	Duplication (%)
CIC IoT 2023	45M	40	≈0.003%	46.66%	53.34%
CIC APT-IIoT 2024	43M	70	0%	50%	50%
CIC IoT-DIAD 2024	55M	84	≈0.31%	49.94%	50.06%
CIC IIoT 2025	685K	94	0%	100%	0%
CIC BCCC NRC 2024	3.3M	85	0%	96.77%	3.23%

Table 3 summarizes the structural properties of each dataset, including scale, feature dimensionality, completeness, and redundancy. Although CIC IoT 2023, CIC APT-IIoT 2024, and CIC IoT-DIAD 2024 contain tens of millions of records, approximately half of their samples are exact duplicates, effectively reducing the number of unique observations to nearly 50%. In contrast, CIC IIoT 2025 presents 100% unique samples despite being considerably smaller in scale, suggesting higher effective diversity per record. CIC BCCC NRC IoMT 2024 exhibits relatively low duplication (3.23%), indicating better structural compactness among large-scale datasets.

From a labeling perspective, Table 4 highlights important differences in hierarchical organization and granularity expansion. While CIC IoT 2023 follows a flat multi-class scheme (35/35 classes), CIC APT-IIoT 2024 and CIC IIoT 2025 introduce multi-level hierarchical labels, enabling progressive refinement from coarse binary classification (2 classes) to fine-grained multi-class scenarios (26 and 84 classes, respectively). This hierarchical structure directly impacts learning complexity, as class imbalance systematically intensifies with increasing granularity. CIC IoT-DIAD 2024 stands apart, as it is currently unlabeled and therefore unsuitable for supervised classification without additional annotation.

Table 5 highlights severe imbalance and data quality issues across datasets. CIC APT-IIoT 2024 exceeds imbalance ratios of 10^6:1, while CIC IoT 2023 surpasses 10^5:1 in fine-grained levels. Even moderately duplicated datasets such as CIC BCCC NRC IoMT 2024 show significant skew (102.78:1). Only CIC IIoT 2025 remains relatively balanced at the binary level (1.41:1), though imbalance increases in deeper hierarchies.

Overall, dataset scale does not guarantee diversity, as duplication rates near 50% substantially reduce effective sample size. The recurrence of similar duplication patterns suggests methodological causes, potentially leading to train–test leakage and inflated metrics. Hierarchical labeling further amplifies imbalance, making fine-grained classification much more challenging than binary detection. Structural quality also varies: CIC IIoT 2025 is clean and complete, whereas others require preprocessing due to inconsistencies such as infinite or constant values. Additionally, heterogeneous labeling strategies, flat, hierarchical, or absent, complicate benchmarking and interoperability. These findings confirm that IDS datasets differ not only in size and attack coverage but also in structural integrity and labeling design, directly affecting evaluation reliability and generalization, reinforcing the need for structural analysis prior to benchmarking.

Table 4: Labeling Structure of Evaluated IDS Datasets

Dataset	Label	Hierarchical	Classes (Min / Max)
CIC IoT 2023	1	Flat	35 / 35
CIC APT-IIoT 2024	3	3	2 / 26
CIC IoT-DIAD 2024	1	None	1 / 1 (Unlabeled)
CIC IIoT 2025	4	4	2 / 84
CIC BCCC NRC IoMT 2024	2	2	2 / 15

Table 5: Data Quality Issues and Imbalance Analysis

Dataset	Max Imbalance Ratio	Notable Issues
CIC IoT 2023	$>$ 765,000:1	Infinite values; extreme skew
CIC APT-IIoT 2024	$>$ 1,500,000:1	Extreme skew; constant features
CIC IoT-DIAD 2024	N/A	Infinite and negative values
CIC IIoT 2025	4,262:1	High variance features
CIC BCCC NRC IoMT 2024	102.78:1	Negative IAT values

Label structure and data quality vary significantly. Hierarchical labeling increases imbalance at finer granularities, making detailed classification more challenging. Structural integrity differs as well: CIC IIoT 2025 presents clean data without missing or duplicated values, while others require preprocessing.

From a research standpoint, datasets support distinct scenarios. Overall, IDS datasets differ not only in scale and attack coverage but also in structural characteristics, directly impacting performance, reproducibility, and generalization.

5 Conclusion and Future Work

This paper presented a comprehensive exploratory analysis of multiple IDS datasets, examining their structural characteristics, feature composition, labeling schemes, and statistical properties. The study highlights that understanding dataset structure is essential before designing and evaluating machine learning–based intrusion detection systems, as traffic datasets are typically created to represent specific environments or threat scenarios rather than to optimize learning performance. The results show that properties such as class imbalance, duplication, missing values, and skewed feature distributions are often inherent to real-world network traffic. While these characteristics may reflect realistic behavior, extreme artifacts such as excessive duplication or inconsistent feature definitions can significantly influence model training and evaluation. In addition, the analysis revealed a lack of standardization across publicly available IDS datasets, including differences in feature naming, attack taxonomies, labeling conventions, and hierarchical organization, which complicates cross-dataset comparison and benchmarking. Overall, the findings emphasize the importance of performing structural dataset analysis prior to model development to ensure reliable evaluation and reproducibility. As future work, we plan to extend the

analysis to additional datasets and explore label harmonization strategies, as well as develop an automated framework for cross-dataset comparison that extracts structural metrics, normalizes labeling schemes, and generates standardized reports to support large-scale IDS research.

References

1. M. Cantone, C. Marrocco, and A. Bria. Machine learning in network intrusion detection: A cross-dataset generalization study. *IEEE Access*, 2024.
2. J.C. Mondragón, P. Branco, G.-V. Jourdan, A.E. Gutierrez-Rodriguez, and R.R. Biswal. Advanced ids: a comparative study of datasets and machine learning algorithms for network flow-based intrusion detection systems. *Applied Intelligence*, 2025.
3. B. Zou Y. Wu and Y. Cao. Current status and challenges and future trends of deep learning-based intrusion detection models. *Journal of Imaging*, 10(10):254, 2024.
4. M. Komarchesqui V. F. Schiavon M. V. O. de Assis L. F. Carvalho V. G. da Silva Ruffo, D. M. B. Lent and M. L. Proença Jr. Anomaly and intrusion detection using deep learning for software-defined networks: A survey. *Expert Systems with Applications*, 256:124982, 2024.
5. Y. A. M. Hamad L. A. H. Ahmed and A. A. M. A. Abdalla. Network-based intrusion detection datasets: A survey. In *2022 International Arab Conference on Information Technology (ACIT)*, pages 1–7. IEEE, 2022.
6. T. Li D. Wu J. Wang Y. Zhao Z. Yang, X. Liu and H. Han. A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Computers & Security*, 116:102675, 2022.
7. A. S. M. Tayeen S. Misra H. Cao J. Harikumar P. Kumar, J. Liu and O. Perez. Flnet2023: Realistic network intrusion detection dataset for federated learning. In *MILCOM 2023-2023 IEEE Military Communications Conference (MILCOM)*, pages 345–350. IEEE, 2023.
8. R. Fernandes and N. Lopes. Network intrusion detection packet classification with the hikari-2021 dataset: a study on ml algorithms. In *2022 10th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5. IEEE, 2022.
9. S. Matic G. Sánchez S. Sebastián J. Caballero, G. Gomez and A. Villacañas. The rise of goodfatr: A novel accuracy comparison methodology for indicator extraction tools. *Future Generation Computer Systems*, 144:74–89, 2023.
10. O.; Hamdi M.; Liouane N. Jablaoui, R.; Cheikhrouhou. Deep learning enabled intrusion detection system for iot security. *EURASIP Journal on Wireless Communications and Networking*, 2025:66, 2025.
11. I.; Idrees S.; Qasim M.; Khan M. J.; Khan J. Bilal, M. A.; Ul Islam. Dataset-centric evaluation of federated intrusion detection models in iot networks. *Scientific Reports*, 2026.
12. J.; Kim W. Meliboev, A.; Alikhanov. Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets. *Electronics*, 11(4):515, 2022.
13. Igal Verker et al. Security data science: The importance of data quality in intrusion detection datasets. *IEEE Security & Privacy*, 2019.
14. Rodolfo S. Miani et al. A survey of data stream-based intrusion detection systems. *IEEE Access*, 2025.