SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I10P107

Volume 11 Issue 10, 91-101, October 2024
© 2024 Seventh Sense Research Group®

Original Article

Enhancing Intrusion Detection System Evaluation: A Framework for Generating Comprehensive and Scalable Datasets

Faeiz M. Alserhani

Department of Computer Engineering and Networks, College of Computer and Information Sciences,
Jouf University, Saudi Arabia.

Corresponding Author : fmserhani@ju.edu.sa

Received: 07 August 2024 Revised: 08 September 2024 Accepted: 07 October 2024 Published: 30 October 2024

Abstract **-** *Intrusion Detection Systems (IDS) evaluation relies principally on the quality and broadness of datasets, which commonly have limitations, including relative scarcity, inadequate coverage of real-world attacks, imbalanced data, and difficulty reproducing with specific requirements. The advancement of IDS algorithms has created a significant gap in the availability of comprehensive and scalable datasets. Developing holistic, well-documented, and trustworthy real-world data traffic is not a simple task; it requires a great deal of effort and high cost. To tackle these challenges, we have proposed a dataset generation framework to construct a reliable dataset based on real-world and synthetic traffic data aggregation, providing a diverse range of attack methods across multiple network settings. The collected traffic records are processed and normalized to produce a consistent dataset. A wide range of multi-step intrusion instances are injected to the constructed dataset to expand the attack coverage. Several tools have been implemented to perform the required data processing steps to automate class labeling and build ground truth data. The proposed framework allows for overcoming the limitations in IDS evaluation in real-world conditions by offering scalable, reproducible, and comprehensive datasets. An experimental dataset has been generated to evaluate different IDS systems such as Snort, Zeek, and machine learning models. The study concludes that the benchmark datasets are fundamental to advancement in IDS research and toward accurate IDS evaluation for safeguarding digital ecosystems against evolving threats.*

Keywords **-** *Intrusion Detection Systems(IDS), Machine Learning (ML), Attack traffic, IDS evaluation, Benchmarking dataset.*

1. Introduction

In today's interconnected world, where the most crucial components in our daily lives are digital systems and networks, data integrity, combined with data security and communication channels, has topped the list of priorities. The sophistication level of cyber threats that come forward every next moment requires even more advanced defense mechanisms. Amongst them, the role of Intrusion Detection Systems (IDS) is a vital line of defense. It provides vigilance to monitor and timely detect any unauthorized access, malicious activities, and potential vulnerabilities within networked systems [1]. Therefore, the evolution of the cybersecurity threats landscape comes with the need to grow robust and effective IDS solutions. To assess the efficacy of Intrusion Detection System (IDS) solutions, researchers and professionals depend on benchmark datasets that simulate real-life network traffic and attack scenarios [2]. These datasets are the basis for testing, training, and benchmarking IDS algorithms and systems [3]. They play a role in evaluating detection accuracy, rates of false positives, and the overall performance of intrusion detection tools. However, these benchmark datasets must be of quality and suitability to ensure IDS research and the development of effective defense mechanisms.

The main objective of this research is to develop a framework for generating reliable and comprehensive IDS datasets with a wide range of security attacks. We start by exploring, analyzing, and investigating the commonly employed datasets to assess intrusion detection systems. It focuses on key research questions related to these datasets' characteristics, types, preprocessing, and limitations. By examining the practices and challenges linked to using benchmark datasets, we aim to offer insights into best practices and the changing landscape of IDS evaluation methodologies. Our research study is organized based on the following research inquiries;

1. What are the key characteristics of benchmark datasets commonly utilized to evaluate intrusion detection systems, including factors like dataset size, diversity,

and representativeness of real-world threats?
2. What are the different types of benchmark datasets for IDS, and how do they vary in terms of attack scenarios and methods used to collect data?
3. How do benchmark datasets for IDS differ in terms of data preprocessing, feature extraction, and labeling? Furthermore, how do these variations impact the evaluation of the IDS algorithm's performance?
4. What are the limitations and challenges inherent in existing benchmark datasets for IDS, specifically concerning their impact on accurate evaluation?
5. What is a framework's main architecture for producing an all-encompassing, dynamic intrusion detection dataset that includes various security attacks?

The proposed framework offers a comprehensive view of network security by integrating isolated attack scenarios and known benchmarks consisting of various attacks with different vectors. A combination of real-world intrusion traffic with synthetically generated attacks is utilized to improve the scalability and representativeness of IDS datasets. This approach addresses limitations in previous IDS evaluation studies and allows flexibility to adapt to emerging threats, ensuring relevance in the evolving cybersecurity landscape.

Within this research, we will set forth a thorough investigation exploration to resolve these inquiries before constructing the proposed framework. This will illuminate the situation regarding standard datasets for intrusion detection systems while providing valuable understanding that is widely beneficial for researchers, professionals, and developers in the cybersecurity field.

2. Benchmark Datasets for IDS: An Overview

Benchmark datasets are critical for assessing the performance of Intrusion Detection Systems (IDS). Understanding their key characteristics is central to effective IDS research and development.

2.1. Dataset Size

One underlying aspect impacting the realism and effectiveness of IDS evaluation is the size of the dataset. It can, therefore, be measured using its number of records or instances. Very large datasets, however, may not be well suited for conducting comprehensive evaluation, although they capture a broader range of network traffic patterns and possible attack scenarios [2]. In any case, large-size datasets would comprise additional features, therefore taking larger computation resources and more time to conduct the evaluation. As a result, a balance between dataset size and usability is often sought in IDS research.

2.2. Diversity of Attacks

To be effective, a good IDS must detect varied attacks, ranging from traditional threats to emerging and more sophisticated intrusions. Hence, the diversity of attacks becomes one of the central factors that should be recognized upon identification of the benchmark dataset. Datasets that comprise different forms of attacks, such as Denial of Service (DoS), Distributed Denial of Service (DDoS), and brute force, amongst others, are therefore required by researchers to facilitate the testing of response on IDS systems against the different threat scenarios [4]. Moreover, including known and novel attack patterns is essential to assess the adaptability of IDS algorithms.

2.3. Network Traffic Types

The nature of network traffic is multifaceted, as it includes different protocols, applications, and various types of data. Hence, IDSes must be inclined towards monitoring and detecting anomalies across these multifarious facets. For example, common types of network traffic stipulated by benchmark datasets include web traffic, IoT traffic, email traffic, and such [5]. The second thing is using datasets to examine the efficacy of IDS across various scenario networks where researchers can categorize and simulate different types of traffic.

2.4. Temporal Aspects

Temporal aspects in IDS evaluation have strong implications. Datasets can be classified into static and dynamic datasets, depending on how they represent network activity over time. Static datasets capture the state of the network traffic at a specific time and mainly provide a snapshot of the network state [6]. On the other hand, dynamic datasets change with time, allowing for the evaluation of IDS in dynamic and constantly changing network environments [7]. The choice between these two types of datasets lies in the specific research goals and the real-world applicability of the IDS being tested.

2.5. Realism and Representativeness

Realism and representativeness remain critical considerations in benchmark datasets. Realistic datasets are the ones that closely mimic real-world network traffic and attack scenarios [1]. Such datasets, for their authenticity, often turn out to be the preference. On the other hand, representativeness is the suitability of the dataset to reflect a large landscape of cybersecurity threats. Properly, intrusions in datasets should cut across various scenarios and threats to be classified as representative.

2.6. Types of Benchmark Datasets

2.6.1. IDS Models

Intrusion Detection Systems (IDS) benchmark datasets take diverse forms to cater to diverse research needs and scenarios. Signature-based IDS use predefined patterns or signatures in their data set for building detections against known attacks [8]. Developing signature-based datasets is key to testing the accuracy level of such systems. These datasets include labeled instances of commonly occurring attacks, thus making them ideal for assessing the extent to which an IDS can recognize known intrusion patterns [9]. These datasets comprise KDD99 and NSL-KDD, which have a wide range of attack signatures that help evaluate the performance of the IDS [10].

In contrast with signature-based benchmark datasets, anomaly-based IDS seeks to detect deviations from established network baselines. Anomaly-based benchmark datasets have explicitly been proposed for testing the IDS's capability in identifying unusual network activities [11]. They may include both normal and anomalous instances, and in most cases, novel, previously unseen attacks may form the focus [12]. For example, within the UNSW-NB 15 dataset, the IDS needs to detect anomalies within legitimate network traffic [9]. Hybrid IDS combines signature-based and anomaly-based detection elements. The hybrid benchmark datasets represent this hybrid approach by incorporating details of the known attacks and anomalies [13]. This hybrid approach is aimed at enhancing the effectiveness of IDS by covering a more comprehensive diverse threat category. In this category are examples comprising the datasets CICIDS2017 and CSE-CIC-IDS2018, characterized by known attacks through novelty anomalies for evaluation [14].

2.6.2. Real-World Datasets

As the name sounds, real-world benchmark datasets are a form of test data that seeks to mimic actual network traffic and the attacks themselves. The datasets often originate from authentic network logs or even capture real network traffic [1]. Given their authenticity, they are valuable in testing IDS against an operating environment. Challenges, though, arise while working with such sizes and complexities. The ISCX2012 dataset is an example based on real network data consisting of genuine network traffic [15].

2.6.3. Synthetic Datasets

Synthetic benchmark datasets are synthetically generated to simulate network traffic and key attacks. While they can never be as authentic as real-world data, they offer a channel through which the aspects of the data can literarily be manipulated towards facilitating control of experiments [16]. If the researcher is evaluating IDSs, they can use synthetic datasets in a controlled environment and subject these systems to a stress test. One of the most well-known artificial datasets is the CTU-13 dataset due to its two facts: it can be synthetic and can be used in different controlled environments through testing [17].

2.6.4. Specialized Datasets

Specialized benchmark datasets would normally be targeted to specific niches or domains within the spectrum of IDS research. For instance, such targeted datasets often cater to unique challenges presented by some phenomena, such as IoT security. BoT-IoT, for example, focuses on traffic and attacks relating to the Internet of Things [18]. The IOT-2023 dataset gives yet another specialized dataset to evaluate IDS performance in the Internet of Things shifting landscape.

2.6.5. Evolving Datasets

Evolving datasets continuously upgrade to stay in the same line as the changing nature of the threat landscape. These sets capture the most recent trends of attacks and changes in attacks in real time. They may contain new threat identifications, increasing the value attributed to IDS evaluation [19]. The CIDDS-001 dataset, which focuses on evolving network threats, exemplifies this category.

3. Data Preprocessing, Feature Extraction, and Labeling

The quality and suitability of benchmark datasets for evaluating Intrusion Detection Systems (IDS) do not depend solely on raw data content. Preprocessing of data, feature extraction, and labeling are extremely important steps during dataset preparation, significantly affecting the proper evaluation of IDS algorithm functioning.

3.1. Data Preprocessing

Data preprocessing is the transformation and refinement of raw data into quality and ready form for IDS evaluation. Normalization, handling of missing values, feature selection, and data augmentation are among the core data preprocessing steps when developing a benchmarking database for IDS [20]. Normalization is a core process ensuring consistency between features used across instances. It becomes important while dealing with different types of network traffic and attacks. Normalization ensures that no individual features dominate the analysis due to scale differences. When handling missing values, datasets may contain missing or incomplete data. Handling missing values through either imputation or deleting affected instances results in incompleteness [21]. Those might be used to find the feature selection methods in the data with the most information and keep them while discarding less informative ones. Also, feature selection could be another contributor to the comparative performance of IDS algorithms. Also, data augmentation may imply techniques such as oversampling or undersampling to produce an assurance that the intrusions present in the dataset have a representation equivalent to the normal environment instances.

3.2. Feature Extraction

Feature extraction involves transforming raw data into a relevant feature set that will express the essential characteristics of network traffic and attacks. As such, regarding benchmark datasets in IDS, the effect of feature extraction on the evaluation process can be great. Key considerations would be related to dimensionality reduction and feature engineering [1]. For instance, meaningful features are derived from raw data in feature engineering.

These could be statistical measures, traffic patterns, or even time-based statistics, depending on the nature of the data. In the case of the high-dimensional datasets, reduction of dimensionality techniques such as Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) could be applied to decrease the number of features but keep necessary information [22].

3.3. Labeling

Labeling is a vital stage of dataset preparation for IDS evaluation. It comprises constituent sub-stages, such as annotating instances to normal and intrusive (with subcategories for different intrusions). The manner of labeling significantly impacts IDS algorithm assessment [23]. The essence of labeling includes ground truth, attack scenarios, and anomaly detection. Labels have certain key features, including accuracy and reliability. A robust ground truth is necessary to evaluate IDS, ensuring that instances are correctly categorized. Besides, particular datasets might focus on only specific types of attacks or intrusion scenarios. Labeling for these should reflect these attacks and scenarios [12]. In an anomaly-based IDS evaluation case, "anomalous" labeling would correspond to all other instances not in normal network behavior. The effectiveness of labeling in identifying these anomalies is paramount.

4. Limitations of Existing Benchmark Datasets

4.1. The Lack of Realism

The most associated limitation with benchmark datasets is that such datasets cannot, in perfect form, replicate all the complexity and diversity occasioned by real-world network environments. Real-world networks show dynamic behaviors and continuously emerging new patterns, posing difficulty in creating datasets fully representative of the network activities [1]. Most often, the complexities in many existing datasets are represented simply, leading to less realistic data.

4.2. Bias Toward Specific Attacks

Most benchmark datasets are biased toward certain types of attacks or intrusion scenarios; their reason may lie behind their origination or the intentions of the researchers who have designed them [24]. Consequently, IDS performance may be more catered to certain attack categories if evaluated against such datasets, while other threats are provided with limited insights. Researchers should take caution of the potential bias in the datasets they choose for evaluating IDS performance.

4.3. Lack of Diversity

Another weakness attributable to benchmark datasets is the lack of diversity in terms of network traffic and attack scenarios. Consequently, IDS solutions tested with benchmark datasets may not be evaluated comprehensively to become adaptive against a range of threats launched [1]. In such a case, the datasets may not capture new intrusion techniques or rising trends in attacks, thus limiting the ability of IDS systems to handle new forms of threats.

4.4. Limited Size and Scalability

Another limitation could be the sizes of benchmark datasets, especially where the scalability of IDS solutions is concerned. Such limited sizes of test datasets might not effectively simulate the demands observed in large network environments, at times understating actual performance challenges that will be observed in real-world scenarios [1]. Understating is a crucial aspect of the IDS, and thus, limitation in dataset size limits effectiveness testing at scale.

4.5. Data Privacy Concerns

Datasets that contain real network traffic data pose privacy concerns. Even if such datasets offer enhanced realism, they generally contain sensitive information like IP addresses or payload data [25]. The use of such data in research needs to be conducted cautiously, with privacy regulations upheld and ethical factors considered.

4.6. Representativeness of Real-World Threats

The ability of the benchmark datasets to appropriately represent the changing landscape of cybersecurity threats may be difficult to achieve. Within a short period, the current t datasets may become outdated as new attacks emerge [1]. Maintaining and sustaining representative benchmark datasets of emerging threats is a continuous struggle for the IDS research community.

4.7. Gaps in Attack Scenarios

Some types of attacks or network scenarios may sometimes be missing within some benchmark datasets. Such gaps result in instances where the test of some factor related to IDS performance is incomplete [1]. Thus, researchers should have a clear understanding of these constraints whenever they are selecting benchmark datasets for the IDS evaluation. This will help them consider using multiple datasets or even supplementary data sources that can fill gaps in these datasets. A clear understanding of these limitations is important to researchers and practitioners since they can guide them in selecting benchmark datasets for IDS evaluation [2]. Moreover, careful IDS solutions evaluation, in terms of the bias, privacy, and scalability impacts, is needed while considering realism and diversity in the dataset.

5. Framework Architecture

This research aims to address the issues associated with dataset benchmarking by generating a dynamic and reproducible dataset. A framework to collect many network traffic records from various sources representing numerous real-world attack scenarios. The proposal is to rely upon public repositories to collect the network traffic of malware samples, different attack activities, and background traffic. This method will contribute to obtaining a larger dataset, including traffic from heterogeneous environments. We present an overview of the proposed framework, which consists of several phases.

We consider various requirements to construct the dataset appropriate for IDS evaluation, including:

- Comprehensive coverage of attacks.
- Network traffic with complete data, including encrypted payload, is to be inspected by different IDS tools.
- Anonymization to preserve privacy.
- Ground Truth to get accurate labeling.
- Complete Attack scenarios if the activity involves sequential steps.

The architectural design of the proposed framework consists of seven phases, as shown in Figure 1.

5.1. Collection of Network Traffic Data

The first phase of the framework collects network traffic data from various sources, such as public repositories, synthetic generators, or real-world sources. This data serves as the basis for building the IDS evaluation dataset. Public repositories host a wide range of network activities captured from various network environments.

The network traffic data is available in raw format and stored in pcap files as a packet capture (pcap) file or as a CSV file containing structured information extracted from the pcap file. The assessment and capabilities of the tools used in succeeding steps inform the selection of a specific format. The second source of network traffic data is synthetic data produced by simulators. Synthetic traffic allows the creation of controlled scenarios, simulates specific attack patterns, and generates different traffic patterns that may not be readily available in real-world datasets.

Additionally, data may be gathered from real network settings such as business networks, university networks, and Internet traffic collecting points. In order to ensure the resilience and efficacy of the collected dataset, it is essential to gather information from a broad range of sources, including distinct network architectures, traffic patterns, and attack scenarios. This diversity allows us to analyze overall network activity and security risks better.

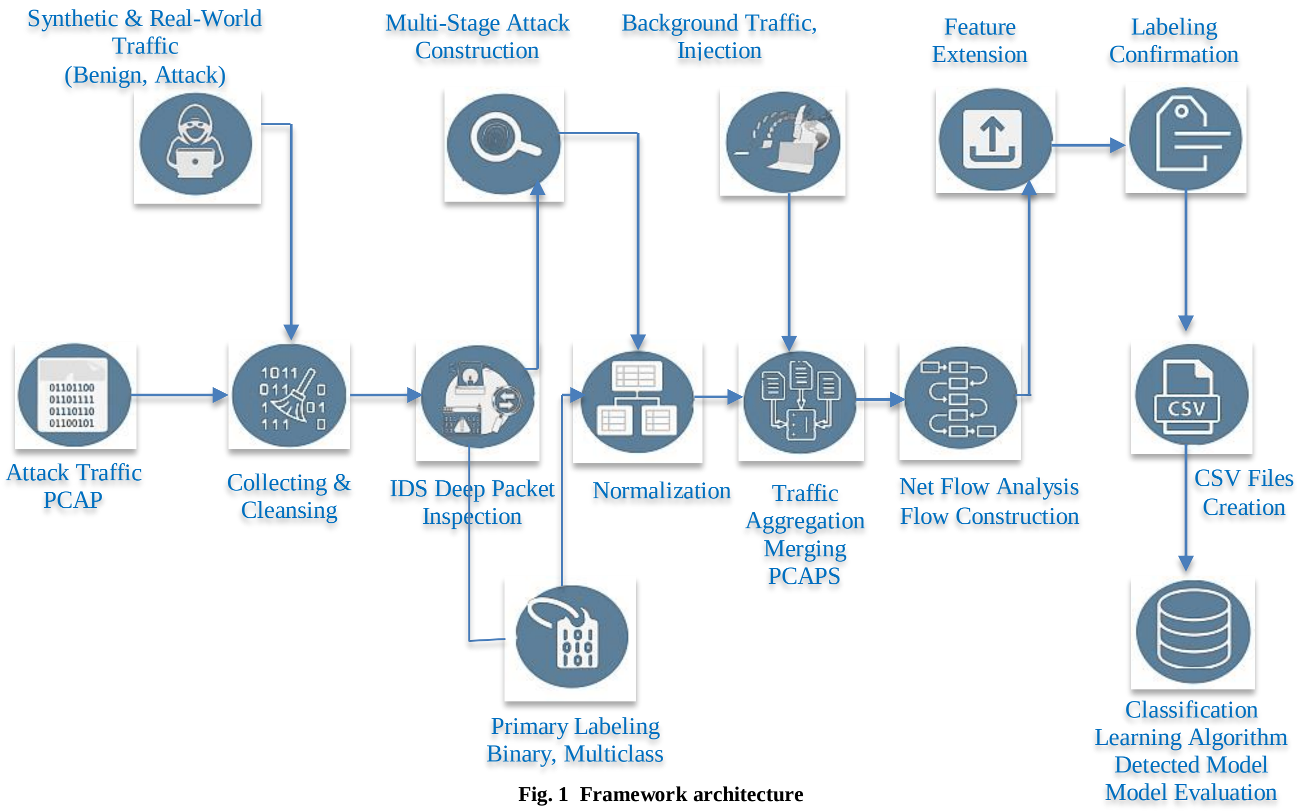
Synthetic & Real-World Traffic (Benign, Attack)
Multi-Stage Attack Construction
Background Traffic, Injection
Feature Extension
Labeling Confirmation
01101100 01101111 01110110 01100101
1011 011 0 1 01 111 0
CSV
Attack Traffic PCAP
Collecting & Cleansing
IDS Deep Packet Inspection
Normalization
Traffic Aggregation Merging PCAPS
Net Flow Analysis Flow Construction
CSV Files Creation
Primary Labeling Binary, Multiclass
Classification Learning Algorithm Detected Model Model Evaluation

Fig. 1 Framework architecture

5.2. Deep Packet Inspection (DPI) for Security Attack Detection

Deep packet inspection of the gathered network traffic in pcap format is required in the second phase of the framework to identify security breaches using a variety of detection methods. These inspection tools then provide alerts, which are gathered, aggregated, and normalized for further investigation. Deep packet inspection is a technique for analyzing the contents of packets passing over a network at a granular level. It involves examining the payload of each packet to detect patterns that indicate security threats or malicious activity. DPI methods use signature-based detection, anomaly detection, machine learning algorithms, or a mix of these approaches to detect suspicious behavior or known attack patterns in network traffic. The detection models generate alerts when deep packet inspection detects suspicious or malicious activities. These alerts provide information about the identified activity, such as the kind of attack, severity level, source and destination IP addresses, and timestamps. These notifications are collected and normalized to guarantee consistency and uniformity in format and presentation. This includes arranging alerts systematically, standardizing field names and values, and deleting duplicate or superfluous data. In addition to primary aggregation and normalization, data fusion and enrichment techniques may improve the quality and value of alert data. This might include combining alerts from several sources to detect relevant occurrences, enhancing alerts with extra contextual information, or supplementing alerts with threat intelligence feeds. Correlation and analysis are performed on the aggregated and normalized alerts to detect patterns, trends, and links between them. This helps to evaluate the network's overall security posture, detect repeated attack patterns, and prioritize security issues for further investigation or action.

5.3. Construction of Multi-Stage Attacks Using Flow and Session Analysis

Moving further through the framework phases, multi-stage attacks are created by analyzing flows and sessions using the kill chain model. This step entails analyzing network data to determine sequential stages of attack execution and mapping them to the stages of the kill chain model. The kill chain model is a paradigm for describing the stages of cyber attacks, such as early reconnaissance, data exfiltration, and system penetration. The process normally involves several steps, including reconnaissance, weaponization, delivery, exploitation, installation, command and control, and action on targets. Each stage represents a different step in the attack lifecycle, with attackers moving through them to achieve their goals. Flow and session analysis examines the patterns and features of network flows and sessions to detect Indications of Compromise (IOCs) and attack behaviors. Analyzing flows and sessions allows for identifying abnormal behavior, unexpected traffic patterns, and potential signs of multi-stage attacks. Multi-stage attacks are created by mapping observable network events to phases of the kill chain model. Each level of the kill chain model correlates to certain actions and behaviors seen in network traffic. For example, reconnaissance efforts may take the form of port scanning or enumeration, whereas delivery may include the transfer of malicious payloads. By linking observed network activities to the steps of the kill chain model, multi-stage attacks may be rebuilt and analyzed better to understand adversaries' tactics, methods, and procedures. Attack scenarios are developed using kill chain model phases and network behaviors to replicate real-world threats. These attack scenarios are used to examine the effectiveness of IDSs, measure the resilience of network defenses, and test incident response capabilities.

5.4. Normalization and Aggregation of Network Traffic Data

Normalization standardizes the format and structure of network traffic data for consistent processing and analysis. This process guarantees that data from many sources or formats is translated into a consistent representation, allowing for easy comparison, correlation, and information analysis. Network traffic data is transformed into an identical format, simplifying processing and analysis. Redundant or duplicate data is removed during the aggregation process, minimizing data redundancy and increasing the efficiency of subsequent analytic operations. Standardising the data's format and structure and reducing duplication makes the dataset more comprehensible and more straightforward to deal with, increasing productivity and performance.

5.5. Flow Analysis and Traffic Attribute Extraction

Flow analysis and traffic attribute extraction are critical steps in the dataset creation. The normalized and aggregated network traffic data is processed using flow analysis tools, which allow traffic sessions to be created. CSV files are then created, providing a full collection of traffic attributes gathered from network traffic and Intrusion Detection System (IDS) tools. Flow analysis facilitates the identification of communication patterns, session formation, and data transfer activities in the network. Using flow analysis techniques, traffic sessions are created by combining related network flows into cohesive communication sessions. A traffic session is a logical link or interaction between network endpoints that involves the exchange of data packets over a period of time. Session generation entails grouping together packets from the same communication stream, such as TCP connections, UDP sessions, or application-layer transactions. Once traffic sessions have been created, CSV files are generated to hold the extracted traffic characteristics in an organized fashion. Comma-Separated Values (CSV) files are generally used for storing tabular data. Therefore, they are appropriate for expressing traffic characteristics derived from network traffic and IDS tools. Each row in the CSV file represents a traffic session, and the columns correspond to various attributes collected from network traffic and IDS alerts. Traffic characteristics derived from network traffic and IDS tools contain a diverse set of information and features that describe the behavior and qualities of network communications. These attributes might include source and destination IP addresses, ports, protocols, packet sizes, timestamps, session length, packet payloads, IDS alert kinds, severity levels, and attack classes. Extracting various traffic parameters allows for a complete examination of network behavior, anomaly detection, and security threat identification. Feature extraction translates raw data into higher-level features or representations that capture network traffic patterns, trends, or characteristics.

5.6. Feature Retrieval and Selection

Extracting useful features from traffic data may be accomplished using feature engineering approaches, statistical analysis, and machine learning algorithms. The feature retrieval and selection phase is a vital stage in IDS assessment. Features are retrieved from network traffic data depending on their properties. The goal is to obtain a broad set of attributes that capture essential elements of traffic behavior. This step guarantees that a complete feature set is accessible for selection, which meets the needs of ML algorithms. These features might include statistical metrics, frequency counts, time-based measurements, protocol-specific qualities, payload data, and security indications. Feature extraction approaches strive to convert raw data into an organized set of features that can be fed into machine learning algorithms. Relevant features are those that can discriminate between different types of communication or detect irregularities. Redundant features that give duplicate or overlapping information can be removed to minimize computational complexity and enhance model performance. Feature engineering may be used to create new or improve current features for better discrimination. Dimensionality reduction, feature scaling, feature aggregation, and transformation are all strategies that use domain knowledge or expert insights.

5.7. Labeling of Network Traffic Records

Network traffic records are labeled based on specified documentation and IDS analysis. During this step, network traffic instances are labeled to identify whether they are normal traffic or other attacks. Binary and multi-classification methodologies and descriptions of the attack stages are used. The labeling process starts with a review of the available material, which might contain planned attack scenarios, recognized threat signatures, or descriptions of harmful behaviors. Documentation is used as a reference to detect and categorize various forms of attacks and abnormal traffic patterns. IDS tools analyze network traffic and identify possible security risks or abnormalities. IDS alerts generated during this step give significant information about suspicious behaviors and indications of compromise in network traffic. Binary classification categorizes network traffic data as legitimate or harmful. Normal traffic logs show legal communication and no evidence of malicious activities. Malicious traffic records show suspicious or unauthorized behavior, suggesting a possible security problem. Multi-classification expands the labeling strategy by categorizing network traffic into numerous attack categories or classes. Different attacks, such as Denial-of-Service (DoS), intrusion attempts, reconnaissance operations, malware infections, and data exfiltration, may be discovered and labeled independently. Each attack class represents a unique security threat or attack vector, allowing for more precise analysis and response. Each labeled instance includes a description of the attack stage and binary and multi-classification labels. The attack stage description specifies which step or stage of the attack lifecycle corresponds to the observed network behavior. This information helps to contextualize discovered threats within the larger context of the attack lifecycle, which aids incident response and mitigation efforts. Security analysts or domain experts can undertake manual labeling and annotation to confirm the outcomes of automatic labeling. Using their experience and judgment, analysts analyze network traffic examples, IDS alerts, and supporting data to validate or alter the given labels. The labeled network contains traffic logs, attack stage descriptions, and accompanying labels. The labeled dataset is invaluable for developing and testing machine learning models, verifying intrusion detection algorithms, and monitoring network security posture. Network traffic labeling requires regular monitoring and upgrading to keep up with emerging threats and attack strategies. As new attack vectors arise or threat landscapes shift, labeling criteria and attack definitions may need to be updated to ensure the labeled dataset's correctness and relevance.

By appropriately labeling network traffic records based on documentation and IDS analysis, organizations may create effective intrusion detection systems, threat intelligence feeds, and security analytics platforms capable of identifying and mitigating a wide range of cyber threats. The labeled dataset is vital for boosting cybersecurity posture and incident response skills and mitigating new security concerns.

5.8. System Design

Network traffic packets are digital recordings of online network activity essential to cybersecurity across various fronts. These packets transmit diverse data across the intricate network of servers and devices that compromise the internet. Packet analysis experts scrutinize payloads and headers for abnormal behavior, a crucial task for IDS to thwart attacks in real time. IDS efficiently neutralizes cyber threats by identifying anomalies and signatures. Normalizing alarms from multiple IDS enhances threat detection across networks. Techniques like packet capture and protocol analysis uncover communication complexities, trends, and vulnerabilities. This thorough examination enables proactive threat mitigation, significantly boosting organizational digital defenses. In order to implement the proposed framework, a detailed systems design can be adopted with full consideration of the general representation in Figure 2. This design will embed the relationships of all these sub-systems and all their elements in a coordinated manner to ensure proper integration and the best operational efficiency within the framework's design.

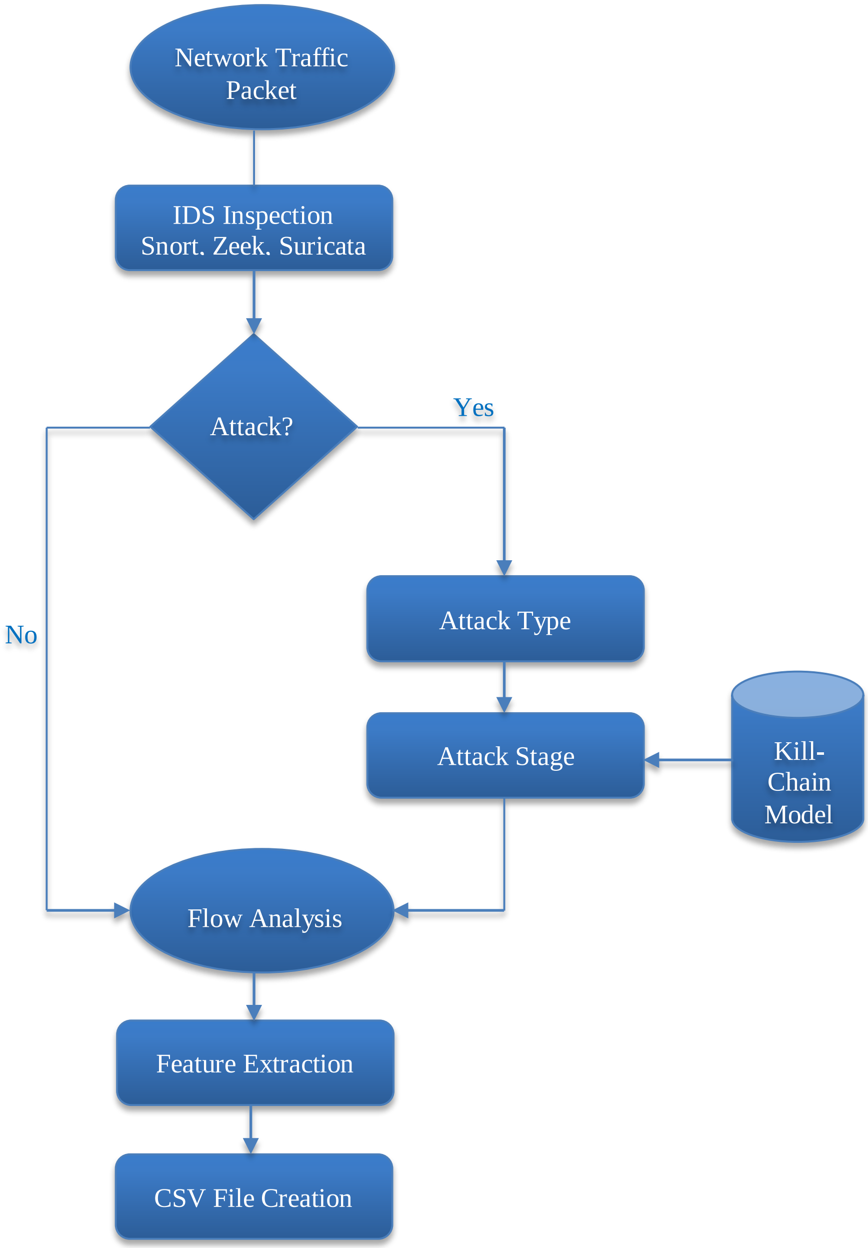

Fig. 2 System design

Traffic records and flow analysis are employed to build a dataset for IDS evaluation. Cybersecurity tools, such as those involving Snort and Zeek, are essential for attack detection in IDS. Snort excels in signature-based detection, identifying known attack patterns quickly, while Zeek offers deep insights into communication patterns through protocol analysis. These tools allow cybersecurity teams to efficiently manage security incidents by analyzing packet payloads and headers, standardizing alerts, and evolving to counter new threats. Machine learning models may assist with attack detection and alert correlation to obtain a global view of security posture. If no ongoing attack is detected, the process moves to flow analysis, examining network traffic to identify patterns and anomalies with tools like NetFlow. Machine learning and behavioral analytics enhance this analysis, helping to detect and respond to threats, protect data, and prioritize actions. Upon confirming an attack, the next step is to classify and understand its characteristics, strategies, and potential impacts. Analysts use automated tools and manual investigation to customize response strategies and mitigate future attacks. Feature extraction follows, identifying attributes from network traffic data to develop machine learning models and anomaly detection systems. This phase converts raw data into insights for threat detection, enhancing cybersecurity posture. Organized into CSV files, extracted features and metadata facilitate data analysis and visualization, integrating with analytics platforms and incident response workflows. This structured data format supports collaboration and actionable intelligence. Flow analysis, feature extraction, and strategic frameworks collectively enhance the ability to detect, respond to, and mitigate cyber threats, protecting organizations' digital assets. This system design enables the generation of datasets for IDS based on the collection of security alert information.

6. Case Study

In order to validate the effectiveness of the proposed framework, we have conducted an experimental evaluation of three types of IDS systems, including Snort, Zeek, and a machine learning model. Our framework has generated the dataset comprising a variety of multi-stage attack scenarios, including Distributed Denial of Service (DDoS) and brute-force attacks. Three categories of network traffic are included in the dataset:

- Normal Traffic: 60% undermining prevalent common communication protocols and patterns.
- Attack Scenarios: 40% consisting of a variety of attack scenarios containing DoS, DDoS, brute force, and malware injection.

Different network traffic sources include real-world and synthetic sources.

The evaluation process was conducted based on typical performance metrics, which were evaluated based on the following metrics: Detection Rate (DR), False Positive Rate (FPR), and processing time. Table 1 demonstrates the efficacy of using a dataset generated by the proposed framework and the evaluation results for the three IDS. Figure 3 illustrates a comparison between the evaluated IDS systems.

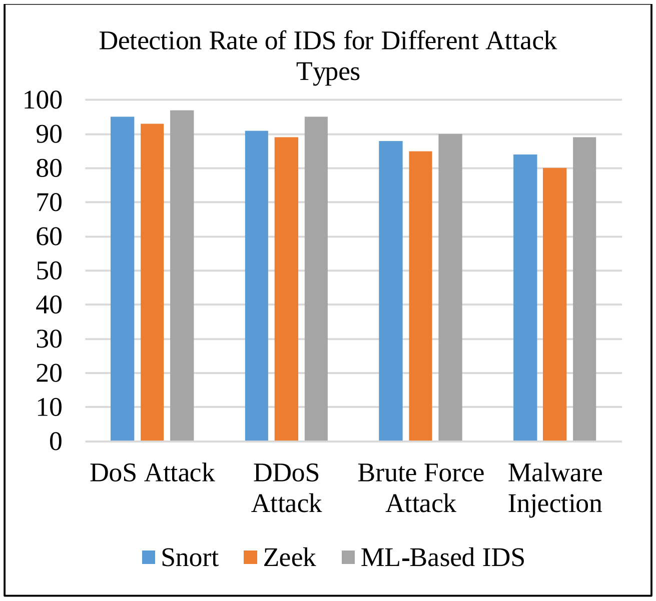

Fig. 3 IDSs detection rates

Table 1. IDSs evaluation of a constructed dataset

Attack Type	Snort (DR / FPR)	Zeek (DR / FPR)	ML-based IDS (DR / FPR)
DoS Attack	95% / 2%	93% / 1.8%	97% / 1.5%
DDoS Attack	91% / 3%	89% / 2.5%	95% / 2%
Brute Force Attack	88% / 4%	85% / 3%	90% / 2.8%
Malware Injection	84% / 5%	80% / 4.5%	89% / 3.2%

7. Discussion

To compare the resulting dataset built using the proposed framework with the current and previous benchmark datasets utilized in IDS evaluation, we discuss several factors affecting the evaluation process.

7.1. Enhancing Realism and Diversity

Future benchmark datasets would be more realistic while modeling real-world network environment complexities. These include other types of traffic in networking systems, such as IoT, industrial control systems, and new communication protocols that may emerge [26].

The datasets should represent all possible networking scenarios, including but not limited to various industries, network architectures, and traffic patterns.

7.2. Evolving Attack Scenarios

The development of benchmark datasets should be agile and responsive to evolving new threats and attack scenarios. It should require regular updating and expanding datasets with the latest variations of existing attacks and techniques [7]. Collaboration with threat intelligence organizations can provide valuable insights into evolving threats.

7.3. Privacy-Preserving Datasets

Concerns about the protection of data privacy are increasing, and future benchmark datasets should embed compliance with various data protection regulations like GDPR mandates. The scope should be to generate datasets that do not violate sensitive information using techniques like data anonymization and synthetic data generation approaches.

7.4. Large-Scale Datasets

Scale-related issues are a critical concern for IDS solutions, especially in the era of big data. Big data scales mean that the benchmark datasets of the future should be characterized by including large-scale datasets within them, and the stress that IDSs should be able to handle scales of big data [26]. These datasets would mimic the demands of large, complex network environments.

7.5. Multimodal Datasets

As attacks become more sophisticated, IDS systems should be capable of handling multiple data sources and modalities. Future benchmark datasets shall comprise combinations of multimodal data sources, such as network traffic, system logs, and user behavior [8]. Considering the evaluation of the performance of the IDS in a multimodal context is critical.

7.6. Attack Attribution and Threat Intelligence

Including information on attack attribution and threat intelligence as metadata in benchmark datasets would be useful as future IDS evaluation may be conducted with information about threat grouping.

For instance, such metadata regarding the source and intent of an attack might allow a researcher to evaluate an IDS to ascertain the extent of its ability to attribute a threat.

7.7. Adversarial Learning and Robustness Testing

Adversarial learning and robustness testing can be the direction to go with the IDS benchmark datasets in the future. For example, a possible way to develop benchmark datasets is by designing adversarial attacks that aim to evade the IDSs while developing adversarial datasets [26].

This way, better evaluation will be done not only on detection capabilities but also on the general resilience of IDS against adversarial threats.

7.8. Community Collaboration

Community collaboration proves vital in solving the problem of developing more complex and larger benchmark datasets in the future. Collaboration helps pool resources with experts from different domains and data sources to develop a more encompassing and representative dataset [26].

Community-driven creation can also be applied to help aid the development of charters that bring about better representation of their dataset for real-world challenges.

7.9. Evaluation of Explainability

With the advancement of AI and machine learning, explainability is an area that further comes to the limelight. Future benchmark datasets must consider the evaluation of IDS explainability so that decisions bestowed by these systems are understandable and explicable to security professionals [8]. Proactive planning is required in benchmark dataset development in the future of IDS research due to the evolving threat landscape and technology advancements.

8. Conclusion

Intrusion Detection Systems (IDS) play pivotal roles in securing our digital environments against a variety of cyber threats. Continuous monitoring, as well as developing the area of intrusion detection systems, requires benchmark datasets to be created and used in the process. This research study draws the key aspects of IDS benchmark datasets and starts recognition of what they are, their characteristics, limitations, best practices, and future directions. Benchmark datasets provide a platform for the assessment of IDS solution performance. We have explored the fundamental characteristics of the benchmark datasets, including their size and diversity, as well as their representation of real-world threats. Current benchmark datasets are unrealistic, limited in diversity, and, besides the associated privacy issues, biased. Analysis of the limitations consequential to current benchmark datasets should be considered constraints to guide researchers in choosing the best dataset to use while interpreting evaluation results. To address these challenges, we created a dynamic and scalable framework for producing benchmark datasets for IDS evaluation. The proposed system allows for fully gathering various types of network traffic, including network data for regular traffic, attack traffic, and generated attack scenarios. Finally, we explored potential research options for benchmark datasets, including improving realism, variety, and scalability.

Funding Statement

This research was supported by Jouf University, Scientific Research Support, Project No. 698-39.

References

[1] Mossa Ghurab et al., “A Detailed Analysis of Benchmark Datasets for Network Intrusion Detection System,” *Asian Journal of Research in Computer Science*, vol. 7, no. 4, pp. 14-33, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Monowar H. Bhuyan, Dhruba K. Bhattacharyya, and Jugal K. Kalita, “Towards Generating Real-Life Datasets for Network Intrusion Detection,” *International Journal of Network Security*, vol. 17, no. 6, pp. 683-701, 2015. [Google Scholar] [Publisher Link]

[3] Yasir Hamid et al., “Benchmark Datasets for Network Intrusion Detection: A Review,” *International Journal of Network Security*, vol. 20, no. 4, pp. 645-654, 2018. [Google Scholar] [Publisher Link]

[4] Sajal Bhatia, Sunny Behal, and Irfan Ahmed, *Distributed Denial of Service Attacks and Defense Mechanisms: Current Landscape and Future Directions*, Versatile Cybersecurity, Advances in Information Security, Springer, Cham, pp. 55-97, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jorge Luis Guerra, Carlos Catania, and Eduardo Veas, “Datasets are not Enough: Challenges in Labeling Network Traffic,” *Computers & Security*, vol. 120, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] R. Vinayakumar et al., "Deep Learning Approach for the Intelligent Intrusion Detection System," *IEEE Access*, vol. 7, pp. 41525-41550, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[7] Ansam Khraisat et al., "Survey of Intrusion Detection Systems: Techniques, Datasets, and Challenges," *Cybersecurity*, vol. 2, pp. 1-22, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Ankit Thakkar, and Ritika Lohiya, "A Review of the Advancement in Intrusion Detection Datasets," *Procedia Computer Science*, vol. 167, pp. 636-645, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Nour Moustafa, and Jill Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set)," *2015 Military Communications and Information Systems Conference (MilCIS)*, Canberra, ACT, Australia, pp. 1-6, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ananya Devarakonda et al., "Network Intrusion Detection: A Comparative Study of Four Classifiers Using the NSL-KDD and KDD'99 Datasets," *Journal of Physics: Conference Series*, vol. 2161, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Suzan Hajj et al., "Anomaly-Based Intrusion Detection Systems: The Requirements, Methods, Measurements, and Datasets," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 4, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Zhen Yang et al., "A Systematic Literature Review of Methods and Datasets for Anomaly-Based Network Intrusion Detection," *Computers & Security*, vol. 116, pp. 1-20, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Imtiaz Ullah, Ayaz Ullah, and Mazhar Sajjad, "Towards a Hybrid Deep Learning Model for Anomalous Activity Detection in Internet of Things Networks," *IoT*, vol. 2, no. 3, pp. 428-448, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Joffrey L. Leevy, and Taghi M. Khoshgoftaar, "A Survey and Analysis of Intrusion Detection Models Based on CSE-CIC-IDS2018 Big Data," *Journal of Big Data*, vol. 7, pp. 1-19, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Wang Yan, Han Dezhi, and Cui Mingming, "Intrusion Detection Model of the Internet of Things Based on Deep Learning," *Computer Science and Information Systems*, vol. 20, no. 4, pp. 1519-1540, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Steven M. Bellovin, Preetam K. Dutta, and Nathan Reitinger, "Privacy and Synthetic Datasets," *Stanford Technology Law Review*, vol. 22, no. 1, 2019. [Google Scholar] [Publisher Link]

[17] Brandon Williams, Xishuang Dong, and Lijun Qian, "Data-Driven Network Monitoring and Intrusion Detection Using Machine Learning," *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Paris, France, pp. 1-7, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Patrick Russell et al., "On the Fence: Anomaly Detection in IoT Networks," *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, Miami, FL, USA, pp. 1-4, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mohamed Amine Daoud et al., "Convolutional Neural Network-Based High-Precision and Speed Detection System on CIDDS-001," *Data & Knowledge Engineering*, vol. 144, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Cheng Fan et al., "A Critical Review on Data Preprocessing Techniques for Building Operational Data Analysis," *Proceedings of the 25th International Symposium on Advancement of Construction Management and Real Estate*, pp. 205-217, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade, "A Review: Data Preprocessing and Data Augmentation Techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91-99, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Anna C. Belkina et al., "Automated Optimized Parameters for T-Distributed Stochastic Neighbor Embedding Improve Visualization and Analysis of Large Datasets," *Nature Communications*, vol. 10, pp. 1-12, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] Ryan Mills et al., "Practical Intrusion Detection of Emerging Threats," *IEEE Transactions on Network and Service Management*, vol. 19, no. 1, pp. 582-600, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Amandalynne Paullada et al., "Data and its (dis) Contents: A Survey of Dataset Development and Use in Machine Learning Research," *Patterns*, vol. 2, no. 11, pp. 1-14. 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Geeta Singh, and Neelu Khare, "A Survey of Intrusion Detection from the Perspective of Intrusion Datasets and Machine Learning Techniques," *International Journal of Computers and Applications*, vol. 44, no. 7, pp. 659-669, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Huseyin Ahmetoglu, and Resul Das, "A Comprehensive Review on Detection of Cyber-Attacks: Data Sets, Methods, Challenges, and Future Research Directions," *Internet of Things*, vol. 20, 2022. [CrossRef] [Google Scholar] [Publisher Link]