IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 March 2024, accepted 17 April 2024, date of publication 22 April 2024, date of current version 2 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3392338

TOPICAL REVIEW

From Bytes to Insights: A Systematic Literature Review on Unraveling IDS Datasets for Enhanced Cybersecurity Understanding

AKBAR KHANAN, (Member, IEEE), YASIR ABDELGADIR MOHAMED, (Member, IEEE), ABDUL HAKIM H. M. MOHAMED, (Senior Member, IEEE), AND MOHAMED BASHIR
College of Business Administration, A'Sharqiyah University, Ibra 400, Oman

Corresponding author: Yasir Abdelgadir Mohamed (Yasir.abdulgadir@asu.edu.om)

ABSTRACT In the wake of the expanding digital realm, the imperative for robust cybersecurity measures has burgeoned significantly. This extensive investigation digs into the complicated realm of cybersecurity datasets, with the goal of improving our understanding and implementation of these critical tools. This study's comprehensive evaluation of 37 distinct datasets shows a complicated world in which no one dataset stands out as totally suitable for all uses. A precise balance must be struck between crucial dataset qualities such as diversity, authenticity, and usefulness. Using a complete assessment technique, this paper illuminates the challenges and possibilities that developers and researchers face in the field of cybersecurity datasets. Although some databases accurately identify certain forms of cyberattacks, their coverage may not include the whole range of cyber threats. On the other hand, datasets with a strong emphasis on accurate portrayal may forgo comprehensiveness or practical use. This intricacy is heightened by the dynamic and sophisticated nature of cyber threats, emphasizing the delicate balance required between accuracy and practicality. The study emphasizes the necessity of selecting datasets strategically and contextually for cybersecurity studies, with the goal of matching research objectives with the most appropriate dataset selections. Furthermore, it emphasizes the need of continual cooperation and innovation within the cybersecurity community in developing datasets that accurately represent the ever-changing nature of cyber threats. After analyzing 37 cybersecurity datasets, it is obvious that no one dataset can meet all of the field's unique demands, demonstrating the need of a flexible, adaptable, and developing dataset for intrusion detection systems (IDS). This inquiry offers a critical assessment of dataset characteristics and their related issues, providing essential insights for academics, professionals, and dataset creators, enabling the construction of a more resilient and adaptable cybersecurity infrastructure.

INDEX TERMS Attacks, intrusion detection system, datasets.

I. INTRODUCTION

The use of the internet has seen a substantial rise, accompanied by the presence of linked networks encountering several vulnerabilities. Consequently, the internet landscape is fraught with several security vulnerabilities. Security firms throughout the globe are actively engaged in the ongoing development of novel methodologies aimed at safeguarding peripherals and sensitive data from cyberattacks. Various security techniques include network-based security systems as well as host-based mechanisms [1], which safeguard peripheral devices against unauthorized infiltration.

These systems are comprised of a collection of devices, mostly including firewalls, intrusion detection systems (IDS), threat prevention mechanisms, basic control over system activities, and a flag enhancement feature that operates depending on the specified detection priority. The prioritization of detection is significantly influenced by the

The associate editor coordinating the review of this manuscript and approving it for publication was Lei Shu.

function that intrusion detection plays. In the realm of information security, intrusion detection is vital since it assists in identifying unauthorized access, modifications, and information systems disruption [2]. In the intricate realm of Intrusion Detection Systems (IDS), experts recognize three predominant categories, each distinguished by its approach to safeguarding the cyber domain. As illustrated in the pivotal Fig. 1 [3], these categories are: the meticulous signature-based, the analytical statistical anomaly-based, and the hybridized combination techniques.

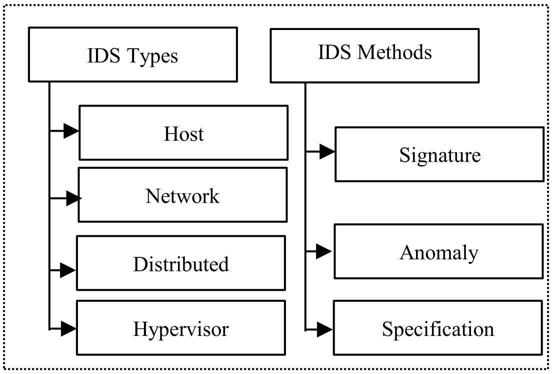


FIGURE 1. Intrusion detection types and methods.

However, the integration of both abuse detection and anomaly detection approaches is used in the combined approach of detection methods. Furthermore, it is common practice for several manufacturers to use signature-based methods in order to detect and classify malware. Attacks with ransomware, zero-day attacks, unauthorized, denial of service (DoS) data breaches, phishing, social engineering [4], and other similar cyber threats are prevalent in contemporary times.

This foundational understanding of IDS types and methods will provide the necessary context for the subsequent analysis and discussion of cybersecurity datasets tailored for these systems.

Security events and cybercrimes have a tremendous effect on both businesses and people, resulting in interruptions and huge financial losses. The repercussions go beyond the immediate aftermath, impacting the overall stability and expansion of the organization.

Deep learning algorithms have become a potent tool for tackling the intricacies of cybersecurity. Deep learning (DL), a specific branch of machine learning (ML), utilizes complex layers of processing to extract high-level representations from large datasets. By using deep learning, cybersecurity companies may transform their security systems, improving efficiency while maximizing cost-efficiency. Deep learning facilitates the early detection and prevention of potential threats, enhancing the overall security stance and decreasing the probability of successful assaults. The cost-effectiveness of deep learning enables optimal deployment of resources while maintaining financial stability. The use of deep learning algorithms will have a crucial impact on protecting digital assets and sensitive information as the field of cybersecurity continues to develop [5].

A. RESEARCH PROBLEM

Researchers in several disciplines encounter a substantial and pervasive obstacle that impedes their advancement. Given the wide range of datasets accessible, researchers have the intricate challenge of selecting the dataset that most effectively aligns with their particular research inquiries [6]. This intricate selection process is profoundly influenced by a confluence of factors, encompassing dataset dimensions, diversity, pertinence, caliber, and alignment with research objectives, thereby magnifying the complexity of the challenge [7]. Moreover, the conspicuous void in a comprehensive dataset expressly tailored for the exigencies of IDS research underscores a substantial deficiency. The core research problem highlighted in this study underscores the compelling necessity to institute a systematic methodological apparatus for dataset selection [8]. In parallel, it vehemently advocates for the conception of a versatile IDS dataset, positioned as a pivotal cornerstone, to be harnessed by researchers spanning diverse domains [9]. Unveiling the layers of this research problem has the potential to amplify the precision and potency of research endeavors, arming researchers with indispensable tools for meticulous dataset curation, and by extension, propelling the evolution of datasets engineered to fulfill precise research objectives [10].

With its far-reaching effects in academia as well as industry, this study might change the face of data-driven research forever. With a systematic and rigorous framework guiding the selection and exploitation of datasets, it opens the way for a new era of scientific discovery and technological growth. The importance of this undertaking is emphasized by the need for a joint effort that draws on the combined knowledge of scholars from many domains. With the help of collaborative efforts, this study hopes to pave the way for researchers to have the resources they need to effectively choose and analyze datasets, allowing them to uncover the secrets contained inside massive amounts of data.

B. RESEARCH MOTIVATION

The thoughtful choice of datasets serves a crucial role in the field of IDS research, as it greatly impacts the quality and effectiveness of research findings within this ever-evolving domain [11]. Academic researchers sometimes encounter the difficulty of establishing their research aims with the datasets at their disposal, which may result in possible contradictions and restrictions within their studies. The main goal of this research is to identify issues through thorough System Literature Research which will help the researcher by selecting the best appropriate datasets for their research pursuits. The research will helps in improving the accuracy and resilience of IDS.

One of the main goals is to build a framework that fosters variety and appropriateness of datasets. Through a methodical process of categorization and analysis of pre-existing datasets, researchers will acquire the ability to make well-informed judgments, effectively aligning the features of the datasets with their specific research objectives. This strategy would not only enhance the efficiency of resource allocation but also guarantee that research endeavors are focused on datasets that faithfully depict the real-world circumstances and difficulties under investigation.

Moreover, this systematic literature review (SLR) aims to examine the essential elements of data quality and consistency. The presence of inconsistent or erroneous data has the potential to compromise the validity of study conclusions. This study seeks to make a contribution to the generation of dependable and replicable research results by delineating the characteristics of datasets of superior quality and suggesting principles for data gathering procedures, labeling accuracy, and data display.

Another significant incentive is to enable the process of benchmarking and comparison within the area of IDS. The SLR aims to find and evaluate benchmark datasets that have been widely recognized and accepted in the research community. This will assist researchers in choosing datasets that are most suitable for their assessment purposes, which may include anomaly detection, intrusion categorization, or network traffic analysis. The use of System Literature review will also boost the effectiveness of benchmarking process and improve the comparison process of various intrusion detection algorithms.

Last couple of decade witness a growing emphasis on the ethical and legal aspects of associated with data privacy [12]. The SLR will explore the areas related to privacy and legal concerns to dataset by providing insight into the possible compliance that may raised for researcher.

The primary objective of this study is to provide guidance to researchers in selecting datasets that align with established ethical and legal norms, hence resolving the aforementioned difficulties.

Within a wider framework, the results of this SLR have the potential to greatly enhance research on IDS. The use of a methodical methodology for dataset selection would enhance researchers' ability to devise novel IDS and make valuable contributions towards the broader objective of improving cybersecurity. Furthermore, the suggested methodology has the potential to function as an instructional asset, offering advantages to students, novices, and professionals interested in exploring IDS study. Consequently, this may contribute to the development of a proficient cohort of cybersecurity experts.

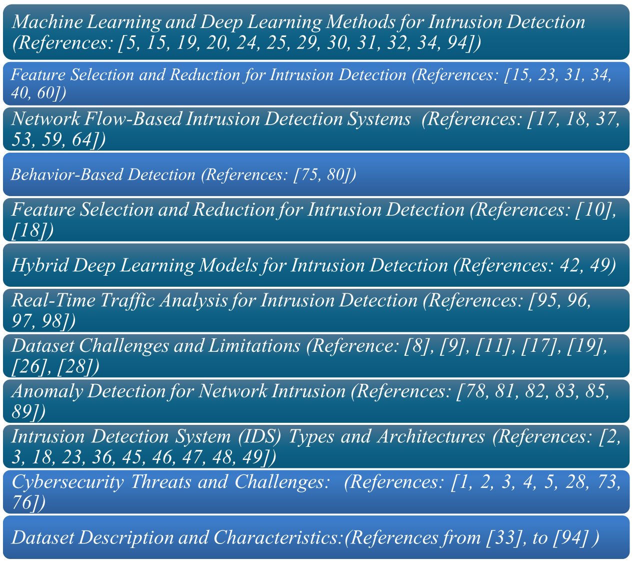
All in all, this SLR aims to address the disparity between IDS research and the process of selecting datasets by proposing a methodical methodology that matches the goals of research with the features of datasets. This study seeks to empower researchers and contribute to the continued development of cybersecurity solutions by focusing on boosting dataset variety, assuring data quality, encouraging benchmarking, addressing ethical issues, and expanding the area of intrusion detection.

C. METHODOLOGY

Through a methodical examination of the complex domain of IDS datasets and the associated academic research, we have diligently undertaken the task of carefully classifying the collected references into distinct clusters. The clusters have been carefully designed as systematic frameworks to categorize the references according to their relevance to the defined aims and core domains of our comprehensive SLR. Each unique cluster functions as a storehouse for a distinguishable group of references that exhibit significant thematic similarities and interrelated subject matter. This enables us to undertake a comprehensive examination of many aspects of IDS dataset research, so providing us with an unparalleled opportunity for nuanced inquiry. The careful utilization of this clustering framework not only provides a well-organized method for presenting our comprehensive literary analysis, but also and potentially more significantly, offers a logical means for extracting and identifying important insights and overarching patterns from the diverse range of academic investigations. In the following sections, we will elaborate on the particular characteristics of each cluster, revealing and discussing their unique traits and significant contributions within the larger scope of our thorough analytical assessment.

1) CLUSTER 1: COMPREHENSIVE SURVEYS AND DEEP LEARNING

The articles in this group are particularly pertinent to the review needs. They conduct extensive literature reviews and in-depth analysis of IDS datasets with an emphasis on DL methods for intrusion detection. These citations include in-depth discussions of taxonomy, assessment, methodologies, and difficulties associated with IDS datasets. Details of this cluster is presented in Fig. 2.



Machine Learning and Deep Learning Methods for Intrusion Detection (References: [5, 15, 19, 20, 24, 25, 29, 30, 31, 32, 34, 94])

Feature Selection and Reduction for Intrusion Detection (References: [15, 23, 31, 34, 40, 60])

Network Flow-Based Intrusion Detection Systems (References: [17, 18, 37, 53, 59, 64])

Behavior-Based Detection (References: [75, 80])

Feature Selection and Reduction for Intrusion Detection (References: [10], [18])

Hybrid Deep Learning Models for Intrusion Detection (References: 42, 49)

Real-Time Traffic Analysis for Intrusion Detection (References: [95, 96, 97, 98])

Dataset Challenges and Limitations (Reference: [8], [9], [11], [17], [19], [26], [28])

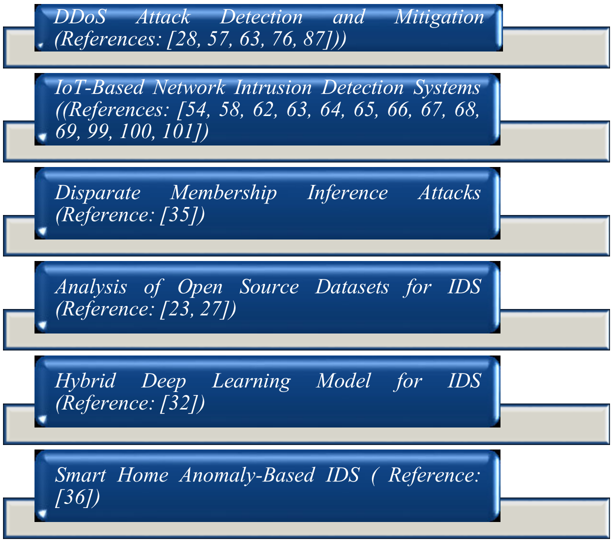
Anomaly Detection for Network Intrusion (References: [78, 81, 82, 83, 85, 89])

Intrusion Detection System (IDS) Types and Architectures (References: [2, 3, 18, 23, 36, 45, 46, 47, 48, 49])

Cybersecurity Threats and Challenges: (References: [1, 2, 3, 4, 5, 28, 73, 76])

Dataset Description and Characteristics:(References from [33], to [94])

FIGURE 2. **Cluster 1: Comprehensive surveys and deep learning articles.**



DDoS Attack Detection and Mitigation (References: [28, 57, 63, 76, 87]))

IoT-Based Network Intrusion Detection Systems ((References: [54, 58, 62, 63, 64, 65, 66, 67, 68, 69, 99, 100, 101])

Disparate Membership Inference Attacks (Reference: [35])

Analysis of Open Source Datasets for IDS (Reference: [23, 27])

Hybrid Deep Learning Model for IDS (Reference: [32])

Smart Home Anomaly-Based IDS (Reference: [36])

FIGURE 3. Diverse topics and applications.

2) CLUSTER 3: DIVERSE TOPICS AND APPLICATIONS

These citations are more peripherally related to the principal research interest (IDS datasets). Approaches, attacks, feature optimization, anomaly detection, and case studies in cybersecurity are just a few of the many areas covered. These citations, depicted in Fig. 3, may be useful for gaining a wider understanding of intrusion detection and cybersecurity; however, they may marginally related to IDS dataset research topics.

3) CLUSTER 4: OTHER TOPICS

This cluster has articles that address various themes such as:

- *Taxonomy and Ontology of Intrusion Detection Systems.*
- *Evaluation of IDS Implementation and Performance.*
- *An Understanding of Dataset Vulnerabilities.*
- *Composition of Hybrid Deep Learning Model.*

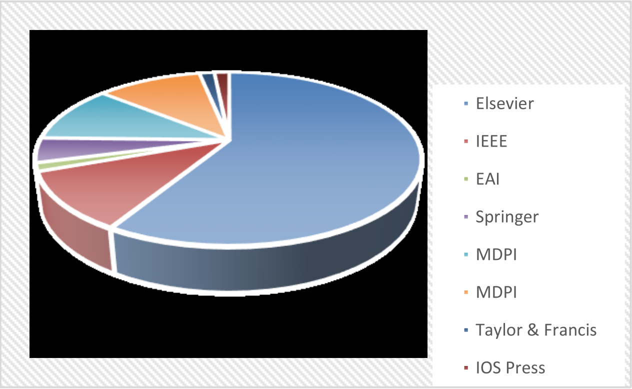


FIGURE 4. Distribution of reviewed articles across various publishers.

Figure 4 depicts the global collaborative nature and diverse origins of research efforts in the intrusion detection domain, showcasing the distribution of publications across renowned academic publishers and organizations. The study encompassed scholarly works from prestigious sources such as IEEE, Elsevier, and MDPI, among others, resulting in a comprehensive assemblage of contributions that reflects the worldwide cooperation and multidisciplinary character of this field.

This systematic literature review transcends the mere exploration of diverse intrusion detection methodologies, recognizing the pivotal role that datasets play in the evaluation and validation of these techniques. By seamlessly integrating sources that delve into IDS datasets and their intricate analysis, we forge a profound understanding of the pragmatic challenges and factors inextricably linked to intrusion detection scenarios in real-world environments. This inclusive approach underscores an unwavering commitment to provide a holistic perspective of the discipline, harmoniously blending theoretical advancements with their practical ramifications, thus illuminating the path towards a comprehensive mastery of the subject matter.

The inclusion of IDS datasets in the review enhances the credibility and relevance of the study. This statement suggests that our attention is not limited to theoretical discourse, but rather extends to the practicality and efficacy of intrusion detection approaches in real-world settings. This method increases the use of our evaluation for academics, practitioners, and decision-makers who are interested in obtaining both theoretical insights and practical assistance for the implementation of IDSs.

By meticulously examining IDS datasets, we exhibit a profound grasp of the intricate complexities interwoven with the assessment of intrusion detection techniques, underscoring the indispensable need for empirical verification. This conscientious acknowledgment of datasets as indispensable assets further reinforces the seminal significance of the systematic literature review, solidifying its status as an authoritative and comprehensive reference for all who traverse the labyrinthine realms of intrusion detection research.

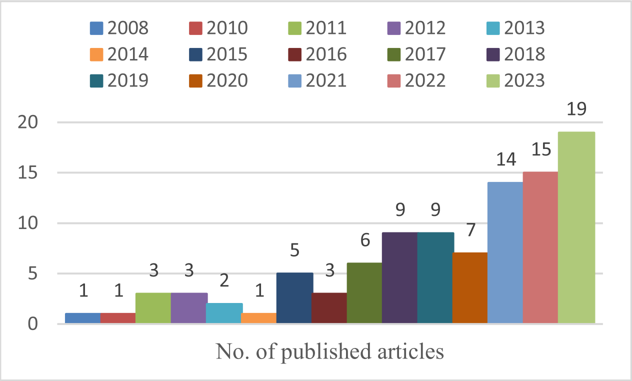


FIGURE 5. Distribution of reviewed articles by year of publication from 2008 to 2023.

The systematic literature review encompasses a comprehensive array of publications spanning the years 2011 to 2023, as vividly depicted in Fig. 5. This expansive temporal scope encapsulates the dynamic evolution of intrusion detection and network security research, encompassing both

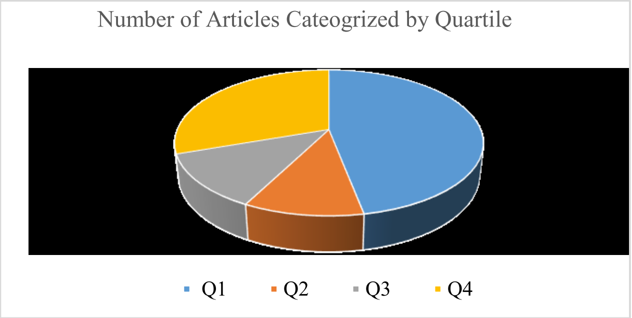


FIGURE 6. **Quartile-wise distribution of reviewed articles based on journal or conference rankings.**

historical milestones and contemporary achievements that have sculpted the discipline. The judicious incorporation of publications across this extensive period proffers an unparalleled panoramic perspective, illuminating the evolutionary trajectories, patterns, approaches, strategies, and obstacles that have been meticulously explored by scholars and professionals over the years, thus fostering a holistic understanding of this ever-evolving domain.

The research articles meticulously examined in this study have been judiciously classified according to their quartile rankings, providing invaluable insights into the academic influence and eminence of these esteemed publications. The quartile categorization, elegantly portrayed in Fig. 6, spans the spectrum from Q1 to Q4, serving as a discerning lens through which to assess the profound significance of the journals or conferences where these seminal works have graced the scholarly realm. This astute evaluation illuminates the intricate tapestry woven by the distribution of research quality and relevance across diverse publishing venues within the intrinsic domains of intrusion detection and network security. By deftly stratifying the articles into quartiles, it unveils a panoramic vista, affording profound insights into the prominence and myriad contributions of various erudite sources. The judicious utilization of this quartile-based analysis in the present study profoundly augments the comprehension of the academic milieu, while simultaneously serving as a beacon, guiding the identification of preeminent contributors to the inexorable progression of knowledge within this hallowed field.

II. RELATED WORKS

This section provides an overview of current efforts pertaining to the development of a benchmark dataset for intrusion detection. The intrusion detection datasets provide a handy platform for the research community to assess and evaluate their methodologies and models pertaining to intrusion detection [13]. Nonetheless, evaluating the proposed detection methodologies using outdated datasets may not accurately represent the effectiveness of these methodologies in identifying contemporary attack types, perhaps yielding inaccurate results [14]. The following is a summary of current efforts that are pertinent to the creation of benchmark datasets for intrusion detection.

The authors [15] presents a thorough and systematic examination of several methodologies and datasets used in the context of anomaly-based network intrusion detection. The research thoroughly addresses several aspects of the IDS, including application domains, preprocessing approaches, detection algorithms, and datasets. Internet-Centric Networking (ICN), Software-Defined Networking (SDN), Internet of Things (IoT), and the Internet are some of the topics covered in the article Nevertheless, despite a small number of papers did explore other application areas, the vast bulk of the publications that were taken into account for this research focused mainly with the Internet. The majority of studies in the field of study mostly emphasize preprocessing methods, feature selection, and feature extraction. The dataset plays a crucial role in an IDS. The author provides a list and overview of 52 commonly used datasets. However, further investigation is necessary to delve into details such as attack descriptions, dataset instances, features, labeling, and comparisons with real-time network data. This comprehensive analysis can assist researchers in enhancing the performance of IDS. The author also highlights that the absence of accessible datasets and the inadequate labeling of datasets pose significant obstacles to the efficacy of IDS.

The authors in [16] underline the rising expertise and motivation of cybercriminals, who target computer users using sophisticated tactics and social engineering schemes. To combat these attacks, improved IDS are critical for successfully identifying contemporary malware. The study provides a thorough examination of IDS approaches, kinds, and technologies, as well as their benefits and drawbacks. It also examines several ML algorithms for detecting zero-day threats. However, several of these techniques have difficulty producing and updating information on new assaults, leading in a large number of false alarms or low accuracy. To address the limits of current datasets, such as DARPA/KDD99, which do not cover newer malware activities, the research underlines the need for newer and more comprehensive datasets encompassing a broad spectrum of malware behaviours. The study also emphasizes the need of building IDS capable of identifying assaults that use evasion strategies, which remains a key research issue. The article sheds light on the significance of advanced IDS as well as the limits of current datasets and methodologies. It may, however, benefit from giving more particular data regarding the detected attacks' tactics as well as the possible consequences for IDS efficacy. Furthermore, the article might investigate alternative remedies or future research areas to solve the issues described in efficiently identifying current and elusive cyber-attacks. Furthermore, discussing real-world case studies or practical examples to substantiate the statements stated in the article may be valuable.

Network based Dataset play a significant role in the training and evaluation for NIDS. The study provides thorough analysis for the training and evaluation of anomaly-based network IDS. This paper presents a comprehensive review of relevant literature pertaining to datasets, specifically focusing on packet- and flow-based network data. A thorough explanation of both types of data is provided. This article [17] does a comprehensive examination of conventional network-based data formats and delineates 15 fundamental characteristics that may be classified into five distinct groups, serving as a framework for evaluating the appropriateness of datasets. This study provides a thorough examination of 34 datasets, with a particular emphasis on their unique characteristics. The analysis primarily centers on attack scenarios and the interconnectedness between them. The research assesses each dataset by considering the indicated qualities, such as features, among others. The study furthermore discovered that there is a rising acknowledgment among the academic community regarding the need of publically accessible network-based datasets. This has led to a rise in the quantity of released datasets in recent times. The paper proposes that the advancement of intrusion detection research may be achieved via the promotion of various dataset assessment, adherence to standard formats, and better cooperation.

Datasets for training and testing Network Intrusion Detection Systems (NIDS) have been analyzed in a literature review by Ghurab [18]. The authors conduct an in-depth study of common network-based data formats and deduce 15 features that may be used to evaluate a dataset's usefulness. General Data Properties, Data Nature, Data Volume, Recording Context, and Data Quality are the subcategories in the stated article. The paper's major contribution is the synthesis of 34 datasets, emphasizing their individual characteristics and concentrating on attack possibilities within the data. In order to help the reader choose the most appropriate dataset for their needs, each dataset is rated according to the discovered qualities. Based on the provided analysis, it seems that researchers have recognized the need for more freely accessible datasets and have been making an effort to publish additional datasets in recent years. The study goes on to address how data repositories and traffic producers contribute to the overall network traffic picture. The study concludes with observations and suggestions for the usage and development of network-based intrusion detection datasets, arguing for standardized formats with preset training and test subsets and stressing the need of assessing algorithms on numerous datasets. In light of the abundance of high-quality datasets now available, the authors advocate for tighter cooperation within the academic community. This research contributes significantly by giving a comprehensive review of datasets for network-based intrusion detection. It would be helpful, though, if further explanation or examples of the qualities used to evaluate datasets were provided. Carrying out more studies focused on resolving issues with dataset creation, labelling, and authenticity should be considered. While the topic of traffic generators and data repositories is informative, it would be helpful to have additional data on the sources' possible shortcomings and biases when it comes to creating or collecting network traffic. In addition, the publication might benefit from more analysis of the difficulties and opportunities presented by researchers working together to share and analyze datasets. Overall, the work is helpful for researchers looking for NIDS assessment datasets, but it might be even better if it explored more depth on certain points and addressed some possible limits.

The identification of emerging attack methods has been a significant problem for researchers in the field. The purpose of the IDS is to identify and detect the emergence of potential cyber attacks. IDS datasets have been used for the purpose of training and simulating IDS pertaining to various assault types. This research conducted by [19] investigated the use of IDS as a means of mitigating cyber-attacks, which are persistently adapting alongside the widespread adoption of the internet and intelligent technologies. Various IDS datasets were used in this study to mimic a range of attack types. The datasets used were CSE-CIC IDS-2018, UNSW-NB15, ISCX-2012, NSL-KDD, and CIDDS-001. The findings of the study revealed that the classifiers used in the research exhibited performance levels that were comparable to or greater than those documented in previous investigations. The Decision Tree (DT) classifier emerged as the most efficacious among the classifiers used, with success rates ranging from 99% to 100% for the CSE-CIC IDS-2018, ISCX-2012, NSL-KDD, and CIDDS-001 datasets, consistent with prior research. A classification technique was created for the UNSW-NB15 dataset. Consequently, the performance of the classifiers on this dataset exceeded that of prior research, indicating the accurate execution of the classification of the UNSW-NB15 dataset. Nevertheless, it is worth noting that existing IDS datasets sometimes overlook the inclusion of attack techniques that are regularly used inside local networks, like Mac flooding, DHCP snooping, and ARP spoofing. The next research endeavors to include novel forms of assaults. However, the absence of prior testing of these attack techniques introduces a level of uncertainty about the successful retrieval of property data via these attacks. This uncertainty is seen as a possible limitation of the future investigation.

The potential value of network-based datasets for training and assessing intrusion detection algorithms, notably NIDS, is emphasized by the authors in [20]. It gives a thorough examination of eight publicly accessible datasets that are extensively utilized in the area. Because of its applicability to present attack scenarios, the authors advocate testing NIDS using recent datasets such as CIDDS-001, CICIDS2017, and CSE-CIC-IDS2018. The research emphasizes the need of avoiding overfitting by examining approaches across many datasets and in a broad context. The primary goal of the investigation is to give an overview of existing datasets for NIDS and suggestions for their usage.

The study effectively emphasizes the importance of network-based datasets and offers useful advice. It might, however, have value in outlining potential challenges or constraints connected with each dataset and how they may impact

NIDS assessment. Furthermore, additional information about the exact characteristics and classes available in each dataset, as well as their relevance to various sorts of assaults, would be beneficial. Furthermore, the article might go into detail on the advantages and disadvantages of various datasets, as well as their applicability for certain assessment circumstances. While the study provides useful insights, it might be improved by integrating additional comparison studies across datasets and giving a more detailed discussion of future research goals in the area of intrusion detection. Overall, the study is a great resource for academics looking for relevant network-based datasets for NIDS assessment, although it may benefit from more development and contextualization.

In their study, [21] investigated the intricacies associated with multi-class classification of network intrusions in datasets characterized by a high degree of data imbalance. The authors specifically used the CIC-IDS2017 and CSE-CIC-IDS-2018 datasets for their analysis. In order to assess the potential enhancement in categorization outcomes for 28 kinds of incursions, a range of ML models with varying degrees of complexity were examined. A comparative analysis was conducted on six ML models, whereby various accuracy metrics were used.

The difficulty in assessing approaches as well as contrasting their effectiveness is highlighted by the authors in [22], who note the existing dearth of meaningful datasets for anomaly identification. In order to keep up with the constantly shifting nature of the internet and malware, the authors stress the need of new projects that supply frequently updated publically accessible statistics. For the purpose of benchmarking new and current anomaly detection ideas, it is important that these datasets include evidence of a wide variety of malware families and malicious actions. The authors also recommend extending the testing set by including tried-and-true techniques for injecting synthetic malware traces into real-world traffic patterns. Anomaly detection may reveal previously unknown malware, helping researchers to keep up with the ever-evolving nature of cyber threats, therefore the potential rewards are tremendous despite the substantial work necessary to attain this aim. The article makes a strong case for using representative datasets when studying anomalies. However, it would be helpful if additional information was provided concerning the difficulties caused by the existing scarcity of appropriate datasets. The study may also examine methods or techniques to promote the development and dissemination of high-quality datasets among researchers. Understanding how the assessment of anomaly detection algorithms is impacted by the possible biases introduced by synthetic malware insertion techniques might potentially be useful. While this article serves a good job of highlighting the advantages of having a broad dataset, it could have benefited from more examples or case studies demonstrating these advantages in practice. The essay successfully highlights the need to handle the dataset difficulties in anomaly detection as a whole, however it might be strengthened by elaborating on several points.

This research paper [23] presents a comprehensive examination of the categorization of IDS with regards to their architecture, detection techniques, decision-making processes, and localization. This research covers a range of IDS methods that are based on ML techniques. Additionally, it provides a thorough examination of DL models and their numerous classifications. The paper also examines publicly accessible datasets used for research focused on IDS. The existing body of research suggests that IDS are of paramount importance in identifying and mitigating various forms of assaults, hence safeguarding networks and systems.

The study has found two primary issues, namely efficiency and performance. The majority of prior studies primarily focus on a restricted collection of datasets. With the ongoing progress of artificial intelligence (AI) models, the susceptibility of networks to assaults is steadily rising. Consequently, the research asserts that there is a pressing need to revise and augment existing databases in order to satisfy contemporary security requirements.

The publicly accessible IDS datasets have limitations in that they do not reflect genuine cyber threats, do not represent real-time network situations, do not represent current malware assaults, and do not consider layer 3 (L3) information. The authors of [24] offer a new realistic, real-time, low-footprint, and up-to-date benchmarked IDS dataset to solve this problem. This dataset's visualization aids in identifying data distortion prior to developing optimal and highly accurate classifier models. To depict L3 information, the research leverages the Eigen Centrality (EC) approach from graph theory, as well as additional techniques such as Principal Component Analysis (PCA) and Gaussian Mixture Model (GMM). The research indicates that lower packet lengths of 1000 to 2000 bytes are predictive of attack features, and the results show that the centrality graph successfully visualizes IPs compromised by recent assaults in real-time. Although the study's goal is to analyze a new realistic benchmarked IDS dataset, it does not go into detail on the dataset's properties or sources, which may be useful for readers looking for further information. The use of complex visualization methods such as PCA and GMM is also mentioned briefly in the paper, however it would be good to highlight how these approaches contribute to the analysis and understanding of the IDS dataset. While the article clearly delivers the key ideas of the research, further detail on the visualization approaches and dataset features would improve its comprehension.

The study conducted by [25] emphasizes on the use of DL models in the identification of phishing assaults, a growing concern in the field of cybersecurity. This study highlights the importance of every DL model, starting from the preprocessing of input data through the generation of the model's output. The primary emphasis of this research is on the process of data preparation. The performance of the model, particularly in real-time detection applications, is significantly influenced by data processing.

The preprocessing phase encompasses three distinct processes, namely cleaning, tokenization, and embedding. Although DL models have shown to be successful, they pose challenges in terms of resource consumption and time requirements. This may be particularly troublesome when it comes to real-time phishing detection systems that are needed by end-users. The paper proposes the use of a Convolutional Neural Network (CNN) to decrease the size of the model, together with the incorporation of Long Short-Term Memory (LSTM) to effectively capture the long-term dependencies of the inputs, as a means to address this concern. However, a significant problem in this field is the exploration of further model limits using real-time data for the detection of zero-day assaults, as well as the development of a model to effectively address these limitations.

The transformation of the Internet into essential infrastructure has increased society's vulnerability to security challenges, resulting in a flood of cybersecurity threats. To address these difficulties, several stakeholders, including business, government, and academia, have worked to reduce risks and install defenses [26]. To assess the efficacy of these projects, novel analytic tools for large-scale empirical data gathered through Internet measuring procedures are necessary. Third-party researchers confront challenges when it comes to acquiring data or making Internet measures, demanding customized procedures for accuracy and completeness. Alternatively, investigations that are more extensive may be carried out by using numerous perspectives, connecting multiple data sources, and experimenting with novel approaches. Researchers attempted to compile disparate literature on measuring methodologies in the cyber security arena, giving an in-depth evaluation of this critical study subject. The authors look at dangers within certain application areas and gives a taxonomy of Internet measuring studies connected to cyber security. It compares the scope, measurement breadth, viewpoint size, and analytic methodologies of each investigation using macroscopically collected data analysis of cyber assaults. The report also explores the limitations of Internet measuring and suggests possible future research directions.

The review gives an in-depth account of the major aims and findings of the research. It may, however, benefit from additional detailed data regarding the domains and application areas addressed in the study. Furthermore, although the article discusses evaluating the breadth and techniques of various research, it does not dwell on the particular results or insights acquired from this comparison. Furthermore, although the review briefly highlights issues in Internet measurement, it would be beneficial to emphasize some of the study's significant challenges and their possible influence for cyber security research. Overall, it successfully delivers the essential elements of the study, but it might be improved with additional detailed material and insights to provide readers with a better grasp of the subject topic.

The study in [27] offers a comprehensive examination of three datasets: the CIC-IDS2018, which is an enhanced iteration of existing open-source datasets, as well as the well-established KDD99 and NSL-KDD datasets. The proposed research aims to improve the IDS via the implementation of several techniques, including model training, model comparison, and the implementation of the Extraction and Prediction procedures. Nevertheless, the research also revealed that the replication of stated findings sometimes present difficulties. However, the conducted data analysis enables readers to reproduce the findings provided and fosters opportunities for future comparisons. The possibility for misclassification error in predicting attack classes, as shown by the research, suggests that IDS may have the capability to identify zero-day attacks, provided that these assaults have a network profile similar to that of known attacks. Misclassification errors were seen in the categorization of DoS, DDoS, Probe, INF, FBF, SBF, and BFW attacks as a result of the presence of comparable Transport Layer network traffic patterns.

The use of the Internet has seen a significant rise, both businesses and individuals conduct a multitude of everyday transactions in the virtual realm rather than the tangible one. This process has been accelerated by the coronavirus (COVID-19) pandemic. Traditional crimes have transferred to the digital area as a consequence of extensive use of the digital environment. Cloud computing, the Internet of Things (IoT), social media, wireless communication, and cryptocurrency are all generating security problems in cyberspace. Cyber thieves have recently begun to employ cyber assaults as a service to automate attacks and maximize their damage. Attackers take advantage of flaws in the hardware, software, and communication layers. DDoS attacksm man-in-the-middle,, privilege escalation phishing, password, remote,, and malware are all of cyber-attacks. The sophistication of these attacks has increased, and new evasion techniques have rendered traditional protection methods like firewalls, IDS, antivirus software, access control lists, etc., ineffective. Consequently, we must immediately begin to address cyber dangers with more innovative and practical strategies. The research article [28] commences by thoroughly explaining the primary causes of cyber assaults. Then it goes into the most current attacks, attack patterns, and detection methods. Third, the essay addresses current technological and nontechnical options for anticipating assaults. Using cutting-edge technology such as ML, DL, cloud platforms, big data, and blockchain to combat present and future cyber assaults might be a potential option. These technical solutions may help detect malware, detect intrusions, identify spam, classify DNS attacks, detect fraud, recognize hidden channels, and differentiate sophisticated persistent threats. However, certain promising methods, including ML and DL, are vulnerable to evasion strategies, which must be taken into account when providing solutions to clever cyber assaults.

This study [29] investigates how well supervised ML intrusion detection models, known for their high accuracy in identifying attacks, perform when exposed to new samples from different datasets. These datasets, including CIC-IDS2017, CIC-DoS2017, and CIC-DDoS2019, share some attack types but have distinct attacks and network settings. The research suggests that pre-trained models, initially trained on the CSE-CIC-IDS2018 dataset, can effectively classify samples from the other datasets. This assumption is commonly made in intrusion detection studies using ML, but it hasn't been extensively tested due to a lack of compatible labeled datasets. Models focusing on attacks involving network interactions showed the least decline in performance when tested on different datasets. The accuracy of classifying brute force, DoS, and DDoS attacks decreased by 5-15 points. The drop in performance was usually less pronounced in terms of balanced accuracy, indicating a larger impact on identifying malicious classes compared to benign ones. However, the extent of performance loss varied based on the dataset. Traditional dataset generalization often overestimates the actual classification capability, highlighting the need for a more thorough validation approach that includes testing. This analysis provides valuable insights into the current limitations of DL in cybersecurity and points towards future research directions to further leverage DL in tackling cybersecurity issues. It also underscores the necessity for ongoing innovation and improvement in this rapidly changing field.

The authors in [30] takes a comprehensive look at cybersecurity applications through DL approaches, given the increasing complexity of securing systems due to rapid technological changes. The authors underline the need for innovative solutions to keep pace with current cybersecurity challenges and the evolving landscape of cyber threats. Three key aspects of DL techniques are examined and discussed in the study. Firstly, the research delves into common cyber-attacks, utilizing publicly available datasets for this purpose. The goal here is to gain a deeper understanding of the nature and characteristics of these attacks, and how they might be best addressed using DL models. Secondly, the study proposes a framework for cybersecurity that leverages DL techniques for a broad range of applications. This framework is likely designed to detect, prevent, and respond to cybersecurity threats, potentially offering an improved and more effective approach to managing these threats. The third aspect involves a practical setup in a lab environment, where live network packets are captured for real-time analysis of cyber security attacks. This hands-on testing allows for an assessment of various important metrics such as false alarm rate, detection rate, accuracy, precision, recall, among others. The intent is to evaluate the effectiveness of the DL models in a dynamic, real-world context. Lastly, the study also reviews the challenges faced by researchers in this field, including both technological and operational aspects.

Datasets utilized in the area of IDS for assessing the efficacy of ML and Data Mining (DM) based IDS are investigated by [31]. Since cybercriminals employ a wide variety of techniques, it is imperative that there is an access to up-to-date datasets that can reveal the most current assaults, as shown by this study. This attack diversity pattern highlights the need for realistic network scenario datasets. The CIC-IDS-2017 and CSE-CICIDS-2018 datasets were developed in response to this need. The limits of these datasets are also discussed, along with a discussion of their features. It is planned as a future work to analyze how well these datasets perform with other ML and DM methods. To address the drawbacks of these datasets, the researchers will also investigate feature engineering and data sampling. The use of ML and DM approaches for performance assessment is also mentioned, however the article does not go into detail on these methods or how they would remedy the problems that have been found. Potential research paths in feature engineering and data sampling to improve the performance of the datasets would be good to highlight, besides, emphasizing the possible consequences or advantages of using modern and realistic datasets in the study of IDS.

Network attacks are the most pressing problem in today's society, thus it's crucial that every network has a reliable IDS to identify and stop harmful intrusions. By monitoring for anomalies in traffic and letting through only legitimate data, DL may improve the ID system's efficiency [32]. To identify malicious network attacks, the study proposes a hybrid ID system built on a CRNN. A convolutional neural network (CNN) and a recurrent neural network (RNN) are used in this model to capture local features and HDLNIDS features, respectively. Positive outcomes in terms of accuracy and data loss have been shown in experiments employing publicly accessible intrusion detection data, in particular CICIDS-2018. In order to identify other forms of malicious network traffic, the researchers want to add more parameters to the model and develop efficient algorithms. In addition, they want to include real backbone network traffic into the validation process and increase the model's ability to deal with zero-day attacks.

More information regarding the HDLNIDS system's performance metrics and how they stack up against those of competing approaches would be valuable. The description also discusses improving the model by adding additional parameters and using efficient techniques but does not specify what these changes would be applied to. The analysis would be more helpful to readers if it included more information on the difficulties or constraints encountered during the trials and how they could have affected the findings. In addition, the summary might emphasize the study's possible relevance in combating the pressing problem of current network threats and in helping to improve cybersecurity measures.

New guidelines for creating datasets are proposed in [33], which aim to remedy deficiencies in previous efforts. These include encryption, since current attacks often employ encrypted communication, anonymization to preserve privacy, payload capture to improve detection of harmful activity inside encrypted data, and ground-truth data to guarantee there are no unlabeled assaults. The authors provide HIKARI-2021, a new IDS dataset that includes missing ground-truth data and covers network traffic with encrypted traces. Each flow in the dataset is either categorized as benign or assault, with multiple attack types defined. The dataset has more than 80 attributes. The article provides researchers with a framework for building their data sets. It provides scripts for collecting and producing synthetic assaults, as well as instructions on developing synthetic attacks and network setups. Human interaction simulation tools, critical for generating fresh data and adjusting traffic according to researchers' demands, are also at their disposal. Researchers may modify datasets to fit their own network setups by following the procedure criteria and creating them in a regulated setting. The authors used ML methods to undertake a basic review of the HIKARI-2021 dataset, scoring it on measures of Accuracy, Balanced Accuracy, Precision, Recall, and F1.

Observations are scheduled to be expanded in the future with ambient traffic, and an assessment will be included. Due to the unlabeled and unknown nature of background traffic, evaluations may make use of unsupervised learning techniques. Since malicious traffic may disguise itself utilizing reserved ports to evade firewalls or IDS by blending in with regular network activity, it is also desirable to compare performance with existing datasets and conduct an analysis of application identification. The article does not explain how a dataset is updated and maintained over time. Furthermore, since network traffic and attack patterns vary, an up-to-date and continually updated dataset is required to stay relevant and valuable in the quickly changing cybersecurity field.

The authors of [34] present a thorough description of cyber security applications that use DL approaches. Three DL algorithms are investigated and explained, beginning with a look at frequent cyber-attacks utilizing publicly available datasets. Following that, a suggested cybersecurity framework is shown, demonstrating the wide application of DL approaches. A lab setup is used to record live network packets, analyze real-time cyber security assaults, and evaluate key attributes such as false alarm rate, detection rate, accuracy, precision, and recall. Researchers have investigated several DL methods to detect, categorize, and forecast various cybersecurity risks. The key areas where DL may be used efficiently were shown, including dealing with needless security alerts and drawing incorrect conclusions. Improving DL approaches in situations with low confidence instances, as well as resolving difficulties linked to faulty or irrelevant components and limited training capacity, are critical concerns. Most study findings are based on publically accessible datasets, however real-time settings are required to evaluate DL techniques against novel cybersecurity assaults. Researchers should concentrate on scenarios in which criminals use DL methods to breach previously safeguarded processes. Accessing real-time information is difficult, and future research should focus on examining diverse open-source datasets to increase dataset quality.

The expenses of applying DL techniques for cybersecurity, as well as the difficulty of mistake correction in DL models, are notable. To build successful cybersecurity knowledge methodologies, researchers should focus knowing the prominent features of incursions. Strong production infrastructure, fast CPUs, huge data repositories, and enough expertise are critical components for effective DL applications in cybersecurity. To improve performance, researchers may combine DL designs with other ML approaches and investigate different built-in DL models. Finally, the paper underscores the relevance of incorporating DL approaches into cybersecurity, emphasizing the need for real-time settings, high-quality datasets, and a targeted approach to addressing the constraints and complexity of DL implementation in the cybersecurity domain.

Author in [35] investigates the susceptibility of datasets to divergent membership inference attacks, a sort of attack that targets certain classes rather than the whole dataset, adding to the current literature on model vulnerability. The researchers built a vulnerability-classification model employing over 100 datasets, many of which have been frequently referenced in the AI security literature. In identifying datasets as susceptible or safe to various threats, the ensemble model, which used logistic regression, Naive Bayes, and random forest models, obtained an amazing testing accuracy of 84.5%. According to the findings, in-class area distribution consistency and a higher concentration of binary characteristics are important variables determining dataset risk. The report also gives information on early hardening solutions for mitigating these vulnerabilities. To address information abundance and class-region sparsity, feature reduction approaches and clever over- and undersampling strategies were used. The best-performing strategy coupled multiple theory-based feature reduction with CTGAN-based oversampling, resulting in a large decrease in divergent attack accuracy without impacting victim model performance much. Notably, just 1% of class-based sub-datasets were more susceptible, whereas 19% grew more safe. This study allows data owners to analyze the susceptibility of their datasets to divergent membership inference attacks and provides insights into potential mitigating measures. Future research should investigate how class-level hardening affects total hardening and investigate other possible hardening approaches to reach even better outcomes.

The researchers created a generic architecture that is passive, adaptable, and efficient for detecting assaults in Smart Home scenarios [36]. Their method is divided into two modules: one for detecting network attacks using Indicators of Compromise (IoCs) obtained from traffic flow reports, and another for detecting abnormalities in IoT application data, which requires customized per-case detectors. This modular design facilitates modification and interaction with third-party components such as Specific Intrusion Detection Systems (SIDS). They identified 10 effective IoCs and two application-level detectors after evaluating their system using a case study containing frequent dangers in the smart home sector. For most devices, testing on a public dataset gave a detection capability of more than 90% and an acceptable false positive rate, successfully recognizing all tested assaults.

Furthermore, the architecture's versatility allows it to be used in a variety of situations by including additional detectors or developing IoCs using the available public dataset as a reference. The article suggests three possible research directions for further work. First, in cases involving third-party processing of user data, lightweight cryptographic approaches may assure information confidentiality and safety. Second, investigate how assaults on smart devices may be tracked by evaluating each device's electrical usage footprint. Finally, unique application-specific anomaly detection algorithms are being developed to supplement current detectors and widen the spectrum of detected attacks. While the proposed generic architecture for detecting attacks in Smart Home environments shows promise, there are some potential criticisms to consider. The validation is limited to a case study and a single dataset, and real-world deployment and scalability challenges have not been extensively addressed. The architecture's ability to detect emerging threats, handle resource constraints, and withstand adversarial attacks also remains unexplored. A comparative analysis with existing IDSs in the Smart Home domain would provide valuable insights. Addressing these concerns through broader validation, real-world testing, scalability considerations, and evaluations against adversarial attacks would strengthen the architecture's credibility and practicality in securing Smart Home environments.

Based on the comprehensive review provided, several research gaps and limitations have been identified in the field of intrusion detection systems (IDS) and cybersecurity. One of the primary challenges is the lack of up-to-date and comprehensive datasets that accurately represent contemporary attack types and real-time network situations. Many existing datasets suffer from insufficient labeling and accessibility, which hinders the effectiveness of IDS. Additionally, there is limited exploration of advanced evasion techniques employed by attackers, leading to difficulties in detecting sophisticated attacks. The scarcity of real-world case studies and practical examples to validate the effectiveness of proposed methodologies is another notable limitation.

Furthermore, current datasets often fail to incorporate attack techniques commonly used within local networks, such as Mac flooding, DHCP snooping, and ARP spoofing. The potential biases introduced by synthetic malware insertion techniques in anomaly detection datasets have not been thoroughly analyzed, which may impact the reliability of the results. Moreover, the performance and scalability of proposed intrusion detection architectures in real-world deployments have not been extensively evaluated. Comparative analyses between proposed architectures and existing IDS in specific domains, such as Smart Home environments, are also lacking.

To address these research gaps and limitations, several future directions have been proposed. Researchers should focus on developing newer and more comprehensive datasets that encompass a wide range of contemporary malware behaviors and attack scenarios. Incorporating real-time network traffic and live network packets for validating the effectiveness of proposed intrusion detection methodologies is crucial. Exploring feature engineering and data sampling techniques can help address the limitations of existing datasets and improve their performance with machine learning and data mining methods.

In scenarios involving third-party processing of user data, investigating lightweight cryptographic approaches to ensure data confidentiality and security is essential. Developing specialized application-specific anomaly detection algorithms can complement existing detectors and expand the range of detected attacks. Evaluating the electrical consumption footprint of smart devices can provide insights into tracking and detecting attacks in Smart Home environments.

Researchers should also consider exploring unsupervised learning techniques for evaluating datasets with unlabeled and unknown background traffic. Continuously updating and maintaining datasets is crucial to keep pace with the rapidly evolving cybersecurity landscape and ensure their relevance and usefulness. Investigating scenarios where attackers employ deep learning techniques to breach previously secured processes and developing countermeasures is another important direction.

Integrating deep learning models with other machine learning approaches and exploring hybrid architectures can potentially improve performance in cybersecurity applications. Conducting broader validation, real-world testing, and evaluations against adversarial attacks is necessary to strengthen the credibility and practicality of proposed intrusion detection architectures.

By addressing these research gaps, limitations, and future directions, researchers can contribute to the advancement of intrusion detection systems and enhance the overall cybersecurity landscape. Collaborative efforts among researchers, industry partners, and government agencies can facilitate the development of robust and effective solutions to combat the ever-evolving cyber threats.

III. INTRUSION DETECTION DATASETS

The aggregation of data from diverse sources, such as network traffic flows, which include information about the host, user behavior, and system specifications, may be used to create a dataset for intrusion detection purposes [37]. The acquisition of data is crucial in order to assess malevolent patterns and unforeseen behaviors linked to diverse network threats. In order to document network operations, a router or switch is used. The use of network flow analysis is employed to assess network traffic subsequent to the acquisition of incoming and outgoing network data. Flow analysis refers to the systematic examination of network packet data.

TABLE 1. **KDD cup 1999 dataset: Attack types.**

Attack Type	Description	Approximate Ratio in Dataset
Normal	Regular network activities without any malicious intent.	60%
Denial of Service (DoS)	Attacks aimed at disrupting the availability of network resources. Examples include 'Syn flood', 'ping of death', and 'neptune'.	25%
Remote to Local (R2L)	Unauthorized access from a remote machine, exploiting vulnerabilities. Examples include password guessing and session hijacking.	5%
User to Root (U2R)	Attacks where a normal user account is exploited to gain root access. Involves various forms of privilege escalation.	2%
Probing	Surveillance activities like port scanning to gather information. Examples include port scans and ping sweeps.	8%

Within this part, we provide a thorough overview of publicly available datasets that pertain to the specific parameters and objectives of our research. The datasets have been provided to academics, analysts, and the wider community, acting as significant resources for diverse inquiries and analysis. Our objective is to enhance comprehension of the existing data environment by providing a comprehensive analysis of the content, structure, and possible uses of each dataset. This collection may be used by both researchers and practitioners to locate appropriate datasets for their study, hence improving the quality and breadth of their investigations.

A. KDD CUP 1999

The KDD Cup 1999 dataset is widely recognized for evaluating the effectiveness of IDS. As part of the 1999 ACM SIGKDD Conference on Knowledge Discovery and Data Mining [38] it was used in the KDD Cup 1999 competition. The dataset was developed with the aim of evaluating IDS within the framework of cybersecurity and network traffic analysis [39]. Its primary function is to evaluate and contrast different IDS feature selection algorithms.

The KDD Cup 1999 dataset is comprised of simulated network traffic data. There are 4,898,431 network connections (instances) in total. Each network instance is defined by 41 characteristics, all of which are generated from network traffic data.

The length of the connection, source and destination IP addresses, source and destination port numbers, protocol type, service being used, amount of bytes transmitted, number of packets exchanged, are some of the characteristics included in the dataset. Some characteristics are categorical, whereas others are continuous or discrete [40].

The dataset contains a mix of regular and attack traffic, as well as many sorts of attacks as illustrated in Table 1 [41]. The attacks are divided into numerous categories, including Denial of Service (DoS) assaults, Probe attacks, User-to-Root (U2R), Remote to Local (R2L) Attacks, Normal Traffic: This class represents non-attacked network connections [42]. Although the KDD Cup 1999 dataset has been extensively utilized for research and evaluation purposes, it is not without limitations [43]. Some of the patterns found in the dataset may not completely reflect real-world network traffic or the most recent forms of cyber-attacks due to its synthetic character and the unique settings under which it was gathered [44].

The KDD Cup 1999 dataset, a seminal resource in IDS evaluation, pioneered standardized benchmarking and spurred innovation. Despite its comprehensive features and large-scale simulation, its synthetic nature and age limit its relevance to modern threats. While valuable for understanding IDS fundamentals, researchers should complement it with recent, representative datasets capturing the complexity of current network environments. Building upon its legacy while adapting to evolving threats, the community can push the boundaries of intrusion detection and develop effective, resilient solutions. The KDD Cup 1999 dataset's significance lies in its groundbreaking role, but its limitations underscore the need for continuous adaptation in the face of ever-changing cybersecurity landscapes.

B. NSL-KDD

NSL-KDD is an abbreviation for "NSL-KDD dataset," which is a better version of the KDD Cup 1999 dataset. The NSL-KDD dataset was developed to solve some of the shortcomings and limitations of the original KDD Cup 1999 dataset, making it a better candidate for assessing intrusion IDSs [45].

While extensively utilized, the KDD Cup 1999 dataset had significant flaws that rendered it less typical of real-world network traffic and created obstacles for reliable assessment of IDS [46]. The inclusion of repetitive records, a lack of a clear divide between training and testing sets, and an imbalance between normal and attack classes were all issues with the original dataset [47]. To address these concerns, the NSL-KDD dataset with the some proposed enhancements. Redundant entries in the NSL-KDD dataset were eliminated,

TABLE 2. NSL-KDD attacks types and description.

Attack Type	Description	Approximate Ratio	Difference from KDD Cup 1999	Research Implications
Normal	Regular activities without malice.	55%	+5%	More balanced, enhancing model generalization for normal behavior.
Denial of Service (DoS)	Disruption of service availability.	20%	-5%	Reduced to prevent overfitting on DoS attacks, aiding in a more generalized model.
Remote to Local (R2L)	Unauthorized remote access.	10%	+5%	Better representation improves model sensitivity to diverse R2L attack strategies.
User to Root (U2R)	User account exploitation for root access.	5%	+3%	Increased presence enhances model training for these critical but rare events.
Probing	Reconnaissance,	10%	+2%	Increased focus aids in better detection of subtle reconnaissance activities.

resulting in a more compact and efficient dataset. The dataset was correctly separated into training and testing sets, ensuring that intrusion detection models were evaluated fairly [48]. The NSL-KDD dataset aimed to balance the normal and attack classes, making it more appropriate for ML techniques. To make it simpler to analyse various elements of intrusion detection, the assaults were classified into four major classes: DoS, Probe, U2R, and R2L, this is depicted in Table 2 [49]. In addition, some of the unnecessary and duplicated features have been deleted, resulting in a more concentrated collection of characteristics for analysis. In comparison to the original KDD Cup 1999 dataset, the NSL-KDD dataset seeks to offer a more realistic and difficult environment for assessing IDS.

There are several issues in the NSL-KDD dataset. It is a synthetic dataset, which might lead to overfitting, and it was released in 2009, therefore it is out of date and does not completely represent current cyber-attack patterns [50]. The representation of innovative assaults in the dataset may be restricted, and class imbalance may bias models. The selection procedure for features may induce bias, and the dataset's scope is largely focused on network traffic analysis, limiting its application to larger cybersecurity concerns. Furthermore, using a rule-based method for labeling may result in mistakes. To guarantee a more full knowledge of intrusion detection performance, researchers should consider these constraints and supplement the assessment with real-world data [51].

The NSL-KDD dataset, an improvement over the KDD Cup 1999 dataset, addresses limitations like redundant records, unclear separation of training and testing sets, and class imbalance. It provides a more streamlined and balanced environment for evaluating intrusion detection systems. However, its synthetic nature, age, and limited representation of contemporary attacks pose challenges. The dataset's narrow focus on network traffic analysis may restrict its applicability to broader cybersecurity concerns. Researchers should exercise caution and supplement the NSL-KDD dataset with real-world data and more recent datasets to develop effective and robust IDS solutions that adapt to the evolving cybersecurity landscape.

C. CICIDS 2017

The CICIDS 2017 dataset, advanced by the Canadian Institute for Cybersecurity (CIC), is a considerable contribution to the field of cybersecurity, it facilitates the evaluation of IDS and related research. With a focus on addressing limitations found in previous datasets like NSL-KDD and KDD Cup 1999, CICIDS 2017 aims to provide a more realistic representation of real-world network traffic and cyber-attacks, making it a valuable resource for developing and testing advanced intrusion detection techniques [52].

Comprising approximately 2.8 million instances, each representing an individual network connection, the dataset offers a diverse set of features to describe the network traffic data. With 80 attributes for each instance, the features cover various aspects, including source and destination IP addresses, port numbers, transport protocols, flow duration, and total bytes transferred [53]. This comprehensive set of features enhances the dataset's ability to represent complex and dynamic network behaviors.

CICIDS 2017 includes a wide range of cyber-attacks, which are categorized into specific attack classes. The dataset encompasses various types of attacks, such as Denial of Service (DoS) attacks, port scanning,, brute force attacks, Distributed Denial of Service (DDoS), SQL injection-like attacks, and cross-site scripting (XSS), botnet activity, and infiltration with data exfiltration attempts [54]. This broad depiction of attacks aids in evaluating the effectiveness of intrusion detection methods against different threat scenarios.

An extensive range of cyber-attacks, categorized into several attack types, are included in CICIDS 2017. The dataset comprises different kinds of attacks, including port scanning, brute force attacks, DoS attacks, DDoS attacks, botnet activity, web application attacks (such as SQL injection and XSS), and penetration attempts with data exfiltration. This broad depiction of attacks aids both researchers and practitioners in assessing the efficacy of intrusion detection technologies against various threat scenarios.

Researchers interested in utilizing the CICIDS 2017 dataset for intrusion detection evaluations and cybersecurity studies should consider its advantages, such as the use of real-world network data, comprehensive features, and the representation of various attack types [55]. Nevertheless, it's essential to acknowledge that even with the improvements offered by CICIDS 2017, no single dataset can fully capture the complexity and ever-evolving nature of cyber-attacks. Hence, it is advisable to complement evaluations with additional real-world data and remain cautious of potential biases or limitations inherent to any dataset [10]. The CICIDS 2017 dataset, along with other datasets and methodologies, contributes to advancing the development of robust and effective intrusion detection techniques to protect network infrastructures from emerging cyber threats [56].

An extensive and varied resource for assessing intrusion detection systems is the CICIDS 2017 dataset. Researchers may evaluate the efficacy of detection algorithms against diverse threats since the dataset includes a wide range of attack types. However, there are several constraints to consider, such as the fact that it may be biased or fail to adequately account for the dynamic nature of cyberattacks. To guarantee the transferability and applicability of created methods, researchers should augment CICIDS 2017 with further real-world data. Although there are certain limitations, CICIDS 2017 makes a substantial contribution to the advancement of intrusion detection technologies that are both resilient and effective.

D. UNSW-NB15

As it contains a vast volume of network traffic data that closely resembles real-world environments, the UNSW-NB15 dataset is a useful means in network security [57]. This dataset is especially valuable for evaluating the efficacy of IDS. The dataset, which was created by academics affiliated with the University of New South Wales (UNSW), is of significant significance since it serves as a standard for evaluating the efficacy of intrusion detection approaches. The dataset contains a substantial number of records, around 2.5 million, including a wide array of attack types such as Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. Additionally, each network flow in the dataset is accompanied by other attributes. The aforementioned characteristics comprise specific facts about protocols, source and destination IP addresses, port information, duration, as well as packet and byte counts, among other relevant factors [58].

The UNSW-NB15 dataset contains well annotated records that are classified as either "normal" or associated with distinct attack types. This comprehensive labeling enhances the dataset's value for studying both regular network activity and malicious conduct. The dataset is often divided into training and testing subsets, which allows for the evaluation of IDS. This enables practitioners to reliably assess the performance of these systems. The assessment of these systems often involves the use of a variety of measures, including accuracy, precision, recall, F1-score, and ROC-AUC, which cooperatively capture the breadth and depth of the dataset.

The dataset known as UNSW-NB15 consists of a total of 42 features. Among these characteristics, three occurrences are categorized as non-numeric or categorical, while the other 39 features are of a numeric nature. The UNSW-NB15 dataset is categorized into the following primary datasets: The UNSW-NB15-TRAIN dataset is used for training different models, while the UNSW-NB15-TEST (100%) dataset is used for evaluating the performance of the learned models [59].

UNSW-NB15 dataset offers a realistic representation of network traffic for evaluating intrusion detection systems (IDS). It provides a comprehensive benchmark. The dataset's strengths lie in its real-world resemblance, diverse attack coverage, and well-annotated labels. However, its limitations include potential biases, lack of recent attack types, and a focus on specific network environments. Researchers should use UNSW-NB15 for evaluating IDS performance against known attacks but complement it with other datasets to assess novel threats and ensure robustness in different network settings.

E. ADFA-LD & ADFA-WD

The dataset known as the Australian Defense Force Academy Linux Dataset (ADFA-LD) is a well-regarded and extensively used dataset for host intrusion detection. It was first released by Xin et al. [60]. System call traces are often used by host intrusion detection systems (HIDSs) for the purpose of identifying attacks directed at target systems. The ADFA-LD dataset has a total of 833 normal training traces, 4372 normal validation traces, and 746 attack traces. These traces were obtained specifically from the Linux system. The integer representation is used to denote each system call in the traces. Within the ADFA-LD dataset, there are a total of seven distinct class labels. These labels are normal, adduser, hydra-ftp, hydra-ssh, java-meterpreter, meterpreter, and webshell. Table 3 depicts the attack vectors used to construct the ADFA-LD attack dataset.

ADFA-LD strengths include its focus on host-level intrusion detection and the use of real system call data. However, its limitations lie in the relatively small size compared to network-based datasets and the specific focus on Linux systems. Researchers should use ADFA-LD when evaluating HIDS for Linux environments but consider complementing it with other datasets for a more comprehensive assessment.

F. BoT-IoT

The main objective of the dataset is to provide a diverse range of attack scenarios related to the Internet of Things (IoT), hence highlighting the distinct vulnerabilities and

TABLE 3. Attack vectors used to generate ADFA-LD attack dataset.

Attack	Payload/EFFECT	Vector	Trace Count
Hydra-FTP	Password bruteforce	FTP by Hydra	162
Hydra-SSH	Password bruteforce	SSH by Hydra	176
Adduser	Add new superuser	Client side poisoned executable	91
Java-Meterpreter	Java based Meterpreter	TikiWiki vulnerability exploit	124
Meterpreter	Linux Meterpreter Payload	Client side poisoned executable	75
Webshell	C100 Webshell	PHP remote file inclusion vulnerability	118

problems that exist within IoT systems. The range of attacks includes several types, such as Distributed Denial of Service (DDoS) attacks that overwhelm IoT resources, as well as brute-force attacks that seek to compromise devices by repeatedly trying to authenticate. Additionally, the BoT-IoT dataset aims to capture instances of Command and Control (C&C) communication patterns, along with activities such as network scanning and reconnaissance with the purpose of identifying vulnerable targets [61]. This study also examines the dissemination of malware across Internet of Things (IoT) ecosystems, demonstrating the capacity for devices to get infected and then spread risks to other interconnected devices. The diverse range of attack types shown in this dataset exemplifies its comprehensive nature, serving as a valuable resource for academics and cybersecurity experts seeking to develop robust tactics for protecting IoT settings. The BoT-IoT dataset is essential for progress of IoT security, servind as a fundamental resource for comprehending, addressing, and combating nascent risks within the dynamic field of IoT security. To get the most up-to-date and comprehensive information, it is recommended to consult authoritative sources, scholarly articles, and academic literature specifically focused on the subject of Internet of Things (IoT) security and cybersecurity. Bot-IoT consists of multiple sets and subdivisions that vary in file format, size, and number of features [62].

The second set, referred to as the Full Set, comprises CSV files containing approximately 73 million instances that were generated by the Argus network security utility. It is essential to remember that each instance represents a network session. All the bytes and packets associated with a single communication session between two sites are represented by a session's characteristics. The Full Set has the fewest number of total features among all processed sets and subsets. It has 26 independent characteristics and 3 dependent characteristics. The 14 additional calculated features developed by Koroniotis are not included in the 26 independent features, which contain only the network flow data from Argus and not the 14 independent features [63].

Although the "BoT-IoT" dataset is a valuable resource for IoT security research, it does have some limitations. The fact that the dataset was collected in a controlled environment is a notable shortcoming, as it may lack the complete spectrum of real-world variation that IoT devices and network traffic exhibit. This environment may not accurately represent the complexities and subtleties of real IoT ecosystems [64]. In addition, the dataset's coverage of IoT devices and attack categories may be insufficient, omitting emergent devices and novel attack methods. As IoT technologies and hazards evolve, the immobility of the dataset may reduce its relevance over time. Additionally, when utilizing the dataset, concerns regarding accurate labeling, potential privacy breaches, and ethical considerations must be addressed. As IoT environments continue to expand, the extent and representation of the dataset may struggle to keep up with the escalating complexity [65]. Additionally, researchers should be aware of potential imbalances in the distribution of benign and malicious instances, which could affect the efficacy of ML models trained on the data. Lastly, the absence of extensive contextual information may limit the depth of analysis. To mitigate these limitations, researchers frequently augment the dataset with additional sources, validate findings in real-world contexts, and employ techniques to account for biases and errors.

With millions of instances, BoT-IoT serves as a valuable resource for developing IoT security strategies. However, its controlled environment may not fully capture real-world IoT complexities, and its coverage of devices and attack types may be limited. As IoT evolves, the dataset's relevance may diminish. Researchers should use BoT-IoT to study known IoT threats but complement it with real-world data and consider potential biases and ethical concerns. Augmenting the dataset and validating findings in real-world contexts can help mitigate limitations and ensure the development of robust IoT security solutions.

G. MIRAI DATASET (CICIDS-MIRAI)

Comprises several files of IoT network traffic data. Each file contains both benign, or normal, network traffic data, as well as malicious traffic data associated with the prevalent IoT botnet assaults often referred to as the Mirai botnet [66]. Three distinct forms of IoT botnet attacks, namely SYN-Flooding, ACK-Flooding, and HTTP-Flooding, were the subject of research emphasis. The consideration of normal/benign traffic statistics was also undertaken. The dataset employed in this study was derived from the IoT network intrusion dataset, previously originated in-house for binary attack classification purposes [67]. This dataset facilitates the classification not only of samples as benign or associated with specific attack types such as ACK-Flooding, SYN-Flooding, and HTTP-Flooding, but also enables the formulation of multi-class classification models. Through analysis of network traffic data, three categories of Mirai-based botnet assaults were discerned. The creation of the recent dataset involved extraction from the original PCAP-formatted dataset, followed by conversion into CSV format, culminating in a dataset encompassing 16 features. The comprehensive attributes of the newly developed dataset, utilized for multi-class attack classification in this study, will be meticulously examined and discussed within the segment dedicated to the analysis and discourse of experimental outcomes [68].

CICIDS-MIRAI strengths lie in its specific focus on Mirai-based attacks and its ability to facilitate the development of IoT botnet detection models. However, its limitations include a narrow scope of attack types and potential lack of generalizability to other IoT botnets. Researchers should use CICIDS-MIRAI when studying Mirai-specific attacks but consider complementing it with datasets covering a broader range of IoT threats to develop comprehensive detection solutions.

H. CIC-AndMal2017

The CIC-AndMal2017 dataset comprises a total of 957 malicious APKs originating from four distinct families, with 647 benign binary files sourced from a variety of applications found on the Play Store. Specifically, the dataset includes 99 instances of Adware, 101 instances of Ransomware, 112 instances of Scareware, and 647 instances of benign files [69]. Below, the descriptions of each malware family is provided:

- The advent and widespread use of mobile banking has engendered the emergence of malicious software designed to intercept users' transactions and illicitly acquire sensitive banking-related information stored on their devices.
- Adware, irrespective of internet connectivity, is a kind of undesirable software that exhibits advertisements on the user's display interface. This Android virus is quite prevalent. In accordance to some experts the aforementioned PUP (potentially undesirable program may have functioned as a precursor to contemporary iterations. Malicious software, commonly plays a role of genuine application or attaches itself to another program in order to deceive users into unwittingly downloading it on their computer, tablet, or smartphone.
- Ransomware refers to a kind of malicious software that use encryption techniques to coerce victims into surrendering their data. The first phase involves the acquisition of system access by the malware. Ransomware has the capability to encrypt either the whole of the operating system or specific files, depending upon the specific variant of ransomware in question. Subsequently, the targeted individual is subjected to intimidation tactics using extortion, whereby they are coerced into making a monetary payment as a kind of ransom.
- The use of scareware involves the deceptive practice of inducing individuals to acquire or get hazardous software, typically devoid of value, by means of trickery. Scareware capitalizes on the psychological vulnerability of users by deceiving them into downloading counterfeit antivirus software, often via the use of a pop-up advertisement. Users may become prey to malware, which has the potential to harm their data, generate financial gains, or facilitate the download of further infections when this program is used.

The CIC-AndMal2017 dataset presents a valuable resource for studying Android malware, offering a diverse collection of 957 malicious APKs from four distinct families (Adware, Ransomware, Scareware) alongside 647 benign files sourced from the Play Store. This dataset's strengths lie in its real-world relevance to the evolving mobile banking landscape, its substantial sample size, and its inclusion of benign applications for balanced analysis. However, limitations include a potentially narrow focus on static characteristics, a lack of temporal data, and a platform-specific nature that might exclude cross-platform threats. Despite these drawbacks, the dataset sheds light on notable trends: the escalation of mobile banking threats, the enduring prominence of adware, the sophistication of ransomware techniques, and the use of psychological manipulation in scareware [70].

Researchers utilizing the CIC-AndMal2017 dataset can delve into the evolution of Android malware, explore mitigation strategies against each malware family, and potentially uncover insights into the shifting tactics of cybercriminals. Addressing the dataset's limitations, such as supplementing static analysis with dynamic behavioral insights and accounting for temporal trends, will contribute to a more comprehensive understanding of the Android malware landscape and bolster the development of effective cybersecurity measures [71].

The CIC-AndMal2017 dataset strengths lie in its real-world relevance, substantial sample size, and inclusion of benign applications. However, its focus on static characteristics, lack of temporal data, and Android-specific nature may limit its applicability to evolving threats. Researchers can use CIC-AndMal2017 to study Android malware evolution, develop mitigation strategies, and uncover insights into cybercriminal tactics. Addressing limitations by incorporating dynamic analysis and temporal trends will enhance the dataset's value in understanding the Android malware landscape and developing effective cybersecurity measures.

I. CAIDA2007

The information presented covers an estimated duration of one hour, namely from 20:50:08 UTC to 21:56:16 UTC, and comprises anonymized traffic traces originating from a Distributed Denial of Service (DDoS) attack that occurred on August 4, 2007. This kind of denial-of-service attack seeks to obstruct access to the designated server by decreasing the server's computational capabilities and saturating the network bandwidth that connects the server to the Internet [72].

The trace duration of one hour is divided into individual pcap files, each spanning a duration of five minutes. The dataset's overall size amounts to 5.3 gigabytes when compressed, and expands to 21 gigabytes when uncompressed [73]. The traces only consist of attack traffic sent towards the victim and the corresponding reactions from the victim. Efforts have been made to eliminate non-attack traffic to the greatest extent practicable. The dataset in question undergoes anonymization by the use of CryptoPAn prefix-preserving anonymization technique, using a singular key. The payload has been extracted from all packets [74]. In the study, it is contended by the researchers that the suitability of the dataset employed is challenged for Denial of Service (DoS) research due to several limitations. Notably, the anonymization of IP addresses was executed to safeguard user privacy; however, this measure impedes precise analysis of attack origins. Furthermore, the complete removal of packet payloads precludes the examination of transmitted data content. The omission of crucial data concerning routine traffic is underscored, a factor potentially conducive to a more comprehensive comprehension of attack characteristics. The DoS attacks found in CAIDA are categorized as flood attacks and can be further classified into two types: stealth attacks, referred to as Low Rate attacks, and High Rate attacks. Low Rate attacks aim to create a minimal number of connections to avoid detection by detection systems. However, these connections are kept open for an extended period, consuming all available resources of the victim. On the other hand, High Rate attacks follow the traditional approach used in DDoS attacks, involving the rapid transmission of numerous packets to deplete the victim's resources as quickly as possible [75].

J. MAC-CDC 2012

The MAC-CDC 2012 [76] study highlights several significant aspects of network traffic behavior, offering insights into the prevalence of malicious traffic versus benign traffic, particularly in contrast to live broadcasts. This emphasis on prevalence aids in understanding the distribution and impact of potential threats within network environments. By showcasing the higher prevalence of malicious traffic, the dataset underscores the importance of robust intrusion detection and prevention mechanisms, prompting researchers and practitioners to focus on refining cybersecurity strategies to counteract these threats effectively.

A notable strength of the MAC-CDC dataset lies in its contribution to the evaluation of rule-based systems. The dataset serves as a foundation for assessing the effectiveness of such systems in mitigating threats, particularly those that deviate from established signatures. This evaluation encourages a critical examination of the limitations of rule-based approaches and drives innovation toward more adaptive and sophisticated threat detection methods. Furthermore, the dataset's approach of constructing a model to deduce absent conditions or antecedents from rules highlights the potential for expanding the capabilities of rule-based systems. This has the potential to drive advancements in rule-based methodologies, fostering the development of more flexible and context-aware security solutions [77].

However, the MAC-CDC dataset also presents certain limitations that warrant consideration. While the study reveals prevalence trends, the dataset's specific size and diversity could impact its generalizability to broader network scenarios. This limitation calls for cautious extrapolation of findings to real-world environments, urging researchers to conduct further investigations with a more extensive range of network conditions and attack types. Additionally, the focus on rule-based systems might inadvertently overshadow the exploration of other emerging approaches, potentially hindering the dataset's ability to comprehensively address the dynamic nature of modern cyber threats.

Trends indicated by the MAC-CDC dataset echo the ongoing evolution of network security practices. The dataset's emphasis on prevalence aligns with the contemporary imperative of understanding the prevalence and distribution of various threat vectors. The exploration of model-based deductions from rules reflects a broader trend toward more intelligent and adaptive security frameworks. It signals the growing recognition of the need to incorporate ML and artificial intelligence techniques to enhance threat detection and response capabilities. This dataset encourages the ongoing exploration of hybrid methodologies that integrate rule-based systems with advanced analytics, fostering a holistic approach to cybersecurity that adapts to the evolving threat landscape.

K. MALWARE TRAFFIC ANALYSES

Malware Traffic Analyses include a collection of Capture the Flag (CTF) tasks designed to facilitate the analysis of network traffic. These challenges serve as a valuable means of honing threat hunting skills, using tools like as Wireshark and Suricata. In this context, we engage in the examination of a third Capture The Flag (CTF) challenge, whereby we undertake the analysis of a Packet Capture (PCAP) file originating from a compromised computing device [78].

Malware Traffic Analysis (MTA) is the examination of network data produced by malware attacks to comprehend their behavior and goals. It involves capturing and inspecting network traffic, analyzing malware actions, and studying communication patterns. The payload, or the malicious part of the software, is scrutinized to understand its capabilities. By identifying indicators of compromise (IoCs) like IP addresses or file hashes, MTA aids in threat detection and response. This process contributes to threat intelligence, enriching our understanding of cyber threats and bolstering cybersecurity measures [79].

L. NUMENTA ANOMALY BENCHMARK (NAB)

The Numenta Anomaly Benchmark (NAB) is a comprehensive framework that has been specifically developed to evaluate the efficacy of anomaly detection algorithms in the context of time-series data. The anomaly detection dataset, created by Numenta, has a wide range of datasets that include labeled abnormalities. These anomalies are representative of real-world situations, hence making the dataset very relevant for assessing the effectiveness of anomaly detection methods in different fields. The National Assessment Battery (NAB) provides a standardized assessment system that incorporates predetermined criteria, baseline algorithms, and a scoring mechanism. This method allows academics and practitioners to objectively quantify the effectiveness of their algorithms. The use of this benchmark enables an equitable evaluation of various methodologies and encourages progress in the field of anomaly detection specifically in the domain of time-series data analysis [80].

The NAB (Network Anomaly Benchmark) is specifically developed for the purpose of evaluating the capabilities of anomaly detection algorithms in the context of streaming data. The objective is to identify and assess the effectiveness of these algorithms in real-world scenarios and their potential usefulness in practical applications. In order to get high scores on the NAB assessment, it is recommended that anomaly detection algorithms operate in an unsupervised manner. There is no need for any special dataset tweaking, engage in ongoing or online learning, and the process relies only on real-time data and does not rely on any kind of anticipation or future projections.

The data included inside the NAB corpus encompasses a diverse range of measures, spanning from IT measurements like as network use, to sensors deployed on industrial machinery, to social media conversations. Additionally, we include a selection of artificially-generated data files into our study. These files serve to assess anomalous behaviors that have not yet been captured in the corpus's authentic data. Furthermore, we add numerous data files that do not contain any abnormalities. The existing NAB dataset comprises 58 data files, with each file containing a range of 1000 to 22,000 data instances. In all, the dataset has 365,551 data points [81].

The NAB dataset is manually labeled, according to a rigorous and well-documented approach. Labelers are required to follow a prescribed set of guidelines while examining data files for irregularities. Additionally, a label-combining technique is used to provide a standardized set of labels that reflect consensus among labelers. The procedure has been created with the intention of minimizing human mistake to the greatest extent feasible. The user did not provide any text. Furthermore, the use of a refined scoring function, as elucidated subsequently, guarantees that minor inaccuracies in labeling will not result in significant fluctuations in the reported scores. An essential component of the NAB dataset is to the incorporation of real-world data that encompasses abnormalities whose underlying causes are known [82].

M. NSL-KDD

The dataset NSL-KDD is designed to address the challenges inherent in the KDD'99 dataset. Statistics of redundant records in the KDD train set includes 4,898,431Original records where 1,074,992 record is Distinct and total Reduction ratio is about 78.05%. Out of total 972,781is normal and 812,814 is normal Distinct records and normal Reduction ratio is 16.44%. 3,925,650 include attack records, where 3,925,650 is attack distinct record and the ratio of attack reduction is 93.32%.the details are shown in the [83]. Statistics of redundant records in the KDD test set includes 311,027Original records where 77,289 record is Distinct and total Reduction ratio is about 75.15%. Out of total 60,591 normal and 47,911 is normal Distinct records and normal Reduction ratio is 20.92%. 250,436 include attack records, where 29,378 is attack distinct record and the ratio of attack reduction is 88.26%.

N. ISCX 2012

The primary concern revolves around the dearth of appropriate datasets that can facilitate precise evaluation, comparative analysis, and effective implementation. A considerable portion of such datasets remains inaccessible due to stringent privacy regulations. Conversely, others are excessively anonymized, rendering them unrepresentative of prevailing trends. Additionally, they might lack essential statistical properties. The shortcomings contribute significantly to the absence of an ideal dataset in this field, an issue that continues to challenge researchers and professionals alike. To overcome the shortcomings a dataset ISCX 2012 created by the Information Security Centre of Excellence at the University of New Brunswick aims. The dataset contains round 2M labeled data sample for 20 attributes/feature. The data for the dataset is captured from network activities for seven days with normal and malicious network traffic. Attack includes HTTP Denial of Service, Infiltrating, Distributed Denial of Service using RC Botnet, Brute Force SSH. The dataset is created with help of 21testbed interconnected windows-based machines. These workstations were divided into 4 LAN, and a main server network which includes web, email, DNS, and Network Address Translation (NAT) services [84]. The dataset is particularly designed for anomaly detection. However, the database can enhance by improving the accuracy features, instances, attack classes, and attack multiclassification to meet the current cyber security challenges. Table 4 and Table 5 depict ISCX-IDS-2012details.

O. CTU-13 DATASET

Botnets pose a persistent menace to internet security, instigating distributed denial of service attacks and spam propagation that drain network resources. Typically, botnets function under a client-server model, where hackers establish a Command and Control Server (C&C Server) following the creation of botnet malware. This server acts on behalf

TABLE 4. ISCX-IDS-2012 details.

Id	Duration (hrs)	# Packets	#NetFlows	Size	Bot	#Bots
1	6.15	71,971,482	2,824,637	52GB	Neris	1
2	4.21	71,851,300	1,808,123	60GB	Neris	1
3	66.85	167,730,395	4,710,639	121GB	Rbot	1
4	4.21	62,089,135	1,121,077	53GB	Rbot	1
5	11.63	4,481,167	129,833	37.6GB	Virut	1
6	2.18	38,764,357	558,920	30GB	Menti	1
7	0.38	7,467,139	114,078	5.8GB	Sogou	1
8	19.5	155,207,799	2,954,231	123GB	Murlo	1
9	5.18	115,415,321	2,753,885	94GB	Neris	10
10	4.75	90,389,782	1,309,792	73GB	Rbot	10
11	0.26	6,337,202	107,252	5.2GB	Rbot	3
12	1.21	13,212,268	325,472	8.3GB	NSIS.ay	3
13	16.36	50,888,256	1,925,150	34GB	Virut	1

TABLE 5. Iscx-IDS-2012 normal and malicious day to day description.

Day	Description	Size (GB)
01	Normal Activity, No malicious activity	16.1
02	Normal Activity, No malicious activity	4.22
03	Infiltrating the network from inside + Normal Activity	3.95
04	HTTP Denial of Service and Normal Activity	6.85
05	Distributed Denial of Service using an IRC Botnet	23.4
06	Normal Activity, No malicious activity	17.6
07	Brute Force SSH and Normal Activity	12.3

of the hackers, disseminating the botnet source code and subsequently receiving reports from successfully infiltrated bots. Once infected, these bots engage with the server to transmit data, update the source code, or receive directives to launch attacks on specified targets. In 2011, the CTU University in the Czech Republic introduced the CTU-13 dataset, a meticulously curated collection capturing both real botnet traffic and conventional background traffic. This dataset is segmented into thirteen distinct scenarios, each representing the deployment of a specific malware variant utilizing diverse protocols and functions.

Nevertheless, while the dataset is comprehensive in terms of botnet attacks, its utility as an IDS benchmark dataset is limited. This is due to its exclusive focus on botnets, and the consequent absence of data on various other types of attacks. This dataset is segmented into thirteen distinct scenarios, each representing the deployment of a specific malware variant utilizing diverse protocols and functions [85].

P. DDoS SIMULATION DATASET (CIC-DDOS2019-2018)

A Large number of devices are now connected to the internet, causing a lot of fast network traffic. Meanwhile, launching DDoS attacks is getting cheaper and the harmful traffic from these attacks is growing. DDoS attacks, ongoing threats, and malware harm the safety and availability of online services. The CIC-DDoS2019 dataset is made by the Canadian Institute for Cyber Security. It has information on DDoS cyber-attacks. The dataset shows 15 types of DDoS attacks from different tools and targets. All these attacks were carried out in a safe space, and details from the attacks were recorded. There are 80 million records in this dataset, both good and anomalies traffic. The records have 88 details like IP addresses, packet size, and more. This dataset is for researchers to study and make better security tools. The dataset include various modern reflective DDoS attacks like PortMap, NetBIOS, LDAP, and others. The study ran 12 DDoS attacks such as NTP, DNS, and WebDDoS. The dataset description is depicted in Table 6 [86].

On the testing day, 7 attacks were executed, including PortScan and UDP-Lag. This is illustrated in Table 7.

TABLE 6. CICDDOS-2019 description.

	Benign	Attack
Benign	110,931	365
Attack	26	453,919

TABLE 7. CIC-DDOS-2019 testing and training attack description.

Testing Set	PortMap , NetBIOS, LDAP, MSSQL , UDP UDP-Lag SYN
Training Set	NTP, DNS, LDAP, MSSQL, NetBIOS, SNMP, SSDP, UDP, UDP-Lag,WebDDoS (ARME), SYN TFTP

TABLE 8. CIC-DDOS-2019 features.

Feature No.	Feature Name	Feature No.	Feature Name
1	Flow Duration	27	Fwd Header Length
2	Total Fwd Packets	28	Bwd Header Length
3	Total Backward Packets	29	Fwd Packets/s
4	Fwd Packets Length Total	30	Bwd Packets/s
5	Bwd Packets Length Total	31	Packet Length Min
6	Fwd Packet Length Max	32	Packet Length Max
7	Fwd Packet Length Min	33	Packet Length Mean
8	Fwd Packet Length Mean	34	Packet Length Std
9	Bwd Packet Length Max	35	FIN Flag Count
10	Bwd Packet Length Min	36	SYN Flag Count
11	Bwd Packet Length Mean	37	RST Flag Count
12	Bwd Packet Length Std	38	PSH Flag Count
13	Flow Bytes/s	39	ACK Flag Count
14	Flow Packets/s	40	URG Flag Count
15	Flow IAT Mean	41	CWE Flag Count
16	Flow IAT Std	42	ECE Flag Count
17	Flow IAT Max	43	Down/Up Ratio
18	Flow IAT Total	44	Average Packet Size
19	Fwd IAT Total	45	Avg Fwd Segment Size
20	Fwd IAT Mean	46	Avg Bwd Segment Size
21	Fwd IAT Std	47	Fwd AVg Bytes/Bulk
22	Fwd IAT Max	48	Fwd AVg Packets/Bulk
23	Fwd IAT Min	49	Fwd AVg Bulk Rate
24	Bwd IAT Total	50	Bwd AVg Bytes/Bulk
25	Bwd IAT Mean	51	Bwd AVg Packets/Bulk
26	Bwd IAT Std		

Table 8 summarizes the -DDOS-2019 key features extracted for network traffic analysis.

Q. CICEV2023 DDoS ATTACK DATASET

Recent studies on DoS or DDoS detection predominantly focus on general networks, leaving a gap in datasets specific to electric vehicle (EV) charging infrastructure. Unlike conventional datasets that mainly record packet reception counts, *CICEV2023* dataset introduces broader ML features such as packet access counts and system status data on charging facilities as depicted in Table 9. The dataset designed by [87] at Canadian Institute for Cybersecurity. The novel dataset, as claimed, derived from a simulator replicating multiple EVs, charging stations, and a charging infrastructure network, encompasses four attack scenarios. It promises to enhance EV charging system analyses and support the development of DoS or DDoS detection tools. The dataset features four attack scenarios related to electric vehicle (EV) charging systems. The dataset encompasses four specific attack scenarios related to electric vehicle (EV) charging systems. In the "Correct EV ID" scenario, an attacker attempts to authenticate using a legitimate ID but does not possess the correct key. Conversely, in the "Wrong EV ID" scenario, the attacker tries to gain access using an incorrect ID, even though they have the genuine key. The "Wrong EV Timestamp" scenario sees the attacker tampering with the timestamp between the EV and the Charging Station (CS), setting it to an outdated value, which triggers an authentication failure within the CS. Similarly, in the "Wrong CS Timestamp" scenario, the attacker modifies the timestamp between the CS and the Grid System (GS) to a previous value, causing an authentication disruption at the GS level.

The CICEV2023 DDoS attack dataset fills a critical gap in DoS/DDoS detection research for electric vehicle (EV) charging infrastructure. With diverse features like packet access counts and system status data, it offers a comprehensive representation of EV charging systems. The dataset's simulation-based approach enables realistic attack scenarios, focusing on authentication and timestamp manipulation. While it may not fully capture real-world complexities, CICEV2023 is valuable for developing targeted detection solutions. Researchers should use it alongside real-world data and consider a wider range of attacks to ensure robustness. CICEV2023 has the potential to drive advancements in securing EV charging systems against DoS/DDoS threats.

R. ANDROID ADWARE AND GENERAL MALWARE DATASET (CIC-AAGM2017)

Sophisticated Android malware has evolved the ability to detect the presence of emulators employed by malware

TABLE 9. CICEV2023 DDoS attack dataset features.

Feature	Description
“Processed_Data”	This directory contains preprocessed data for our research on the attack detection model regarding DDoS attacks on the EV-CS-GS environment.
“Raw_Data”	This directory is the root directory of the dataset.
“Correct_ID”	The correct id attack scenario data belongs in this directory.
“Wrong_CS_TS”	The wrong timestamp data on the charging station belongs in this directory.
“Wrong_EV_TS”	The wrong timestamp data on the EVs belongs in this directory.
“Wrong_ID”	The wrong id data of the EVs belongs in this directory.
“Random_CS_Off”	The data without the random attack targeting strategy belongs in this directory.
“Random_CS_On”	The data with the random attack targeting strategy belongs in this directory.
“Gaussian_Off”	The data without the Gaussian attack strategy belongs in this directory.
“Attack”	The attack data belong in this directory.
“Normal”	The normal data belongs in this directory.
“cs/gs_record”	The Perf record data of CS or GS belongs in this directory.
“cs/gs_stat”	The Perf STAT data of CS or GS belongs in this directory.
“cs/gs_top”	The Perf TOP data of CS or GS belongs in this directory.
“acn_data.csv”	This file contains real EV charging schedules from the ACN-Network.
“attack_config.csv”	This file contains the information on the attack scenario, normal authentication trials and attack trials.
“attack/normal_mode.txt”	This file contains the information on simulation environment settings.
“attack/normal_time_diff.txt”	This file contains the CS ID list and data points of the intervals on DDoS attacks or normal EV authentication trials.
“authentication_results.csv”	This file contains the results of the normal EV authentications and DDoS attacks.
“cs_id_pid.csv”	This file contains the CS IDs matched with specific process IDs in the Linux kernel.
“cs_installation.csv”	This file contains the CS list installed legitimately.
“date.csv”	This file contains the start date and end date of the simulation.
“ev_authentication.csv”	This file is similar to “authentication_results.csv.”
“ev_count.txt”	This file shows how many EV authentication and DDoS attack trials are made through different CS. CS ID, attack count and normal authentication count are in order.
“ev_installation”	This file shows normal authentication or attack sequences, session ID, CS ID, session key information and the result of successful CS installation.
“gaussian_attack_count.txt”	This file is used for the paper of this work.
“mean_std.txt”	This file is used for the paper of this work.

analysts. In doing so, the malware can change its behavior, effectively evading detection. To address this challenge, instead of relying on emulators, the study opted to install the Android applications directly onto real devices and then monitored their network traffic while creating CICAAGM dataset. CICAAGM dataset is meticulously assembled by semi-automatically installing Android apps on actual smartphones. This dataset comprises data from as many as 1,900 applications [88].

The research examines three categories of Android applications. First, Adware, which includes apps like Airpush and Dowgin, both notorious for unsolicited ads and information theft; Kemoge and Shuanet, designed to hijack devices; and Mobidash that displays ads while compromising user data. Second, General Malware, consisting of apps such as AVpass, a disguised Clock app; FakeAV, a deceptive software upsell; FakeFlash/FakePlayer, a counterfeit Flash app redirecting users; GGtracker, which conducts SMS fraud and theft; and Penetho, a fake hacktool for WiFi passwords that can also infect computers via multiple channels. Lastly, Benign apps, sourced from GooglePlay's 2015 and 2016 top free listings, totaling 1,500 apps [89].

The public Android Malware Dataset comprises 24,553 malware samples, each of which is scrutinized using 55 antivirus tools via Virus Total. For an app to be categorized as malware within this dataset, it must be identified by more than 28 of the antivirus tools employed in the three Virus Total checks.

The dataset's strengths lie in its use of real devices, ensuring the captured behavior reflects real-world scenarios. It covers a diverse range of malware types, including adware, device hijackers, information stealers, and deceptive apps. The inclusion of benign apps allows for a balanced analysis and evaluation of detection methods. However, the dataset's limitations should be considered. The rapid evolution of Android malware means that the dataset may not include the most recent threats. Additionally, the focus on network traffic analysis may not capture all aspects of malware behavior, such as local device interactions. Researchers should use CIC-AAGM2017 when studying known Android malware behaviors and developing detection methods that can handle evasive techniques. However, it is essential to complement this dataset with more recent samples and consider a holistic approach that includes both network and device-level analysis.

S. MACCDC2012

The MACCDC2012 represents the Mid-Atlantic Collegiate Cyber Defense Competition held in the year 2012. Organized annually, the CCDC events offer a unique platform for collegiate scholars to experience real-world cybersecurity challenges. In these competitions, participating teams are charged with the mission of defending simulated corporate networks against threats from professional "red team" attackers. This setup not only evaluates their defensive strategies but also assesses their capacity to sustain regular business operations amidst cyber threats.

What distinguishes the MACCDC from typical capture-the-flag cybersecurity contests is its emphasis on the operational aspect of managing and safeguarding a network over just tactical offense or defense. This comprehensive approach ensures that participants get a holistic view of network security. The datasets emerging from these events, encompassing network logs, PCAP files, and more, serve as invaluable resources for researchers. Such datasets allow for the in-depth examination of network traffic, potential vulnerabilities, and contemporary attack methods [90].

The MACCDC2012 dataset is notable for its extensive coverage and genuine nature, providing researchers with a complete perspective on network operations, vulnerabilities, and attack techniques. Nevertheless, its drawback is that it only portrays a particular incident that occurred in 2012, and the potential risks and challenges may have changed since that time. This dataset is very helpful for examining network defense strategies, incident response procedures, and forensic analysis. However, it may not provide an accurate representation of the most current methods of attack or security measures. In order to have a comprehensive grasp of current cybersecurity concerns, it is crucial to supplement this dataset with the latest threat information and best practices.

T. CRIME DATASET (CICIDS-CRIME2018)

Anomaly detection, vital for identifying novel cyberattacks, faces challenges transitioning to real-world applications due to system complexity and the exhaustive pre-deployment testing needed. The ideal test environment would use genuine labeled network traces, rich in intrusions and anomalies. However, finding suitable datasets is arduous: many are privately held due to privacy concerns, while publicly available ones may be overly anonymized, outdated, or lack crucial statistical traits. As a result, researchers often work with suboptimal datasets. Given the rapid evolution of network behaviors and cyber threats, there's a pressing need for dynamic datasets that are current, modifiable, and reproducible [2].

In the CSE-CIC-IDS2018 dataset, "profiles" are introduced as a structured method for generating datasets. These profiles encapsulate intricate details about intrusions and offer abstract distribution models for various applications, protocols, and foundational network entities. Agents or human operators can harness these profiles to instigate specific events within the network. Owing to the profiles' abstract foundation, they can be universally applied across a plethora of network protocols and distinct topologies. The dataset encompasses seven distinct attack scenarios, namely: Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and internal network infiltration. The setup involves an attacking infrastructure of 50 machines targeting a victim organization that is structured into 5 departments, consisting of 420 machines and 30 servers. Comprehensive data is provided in the form of captured network traffic and system logs from each machine. Moreover, 80 specific features have been extracted from this traffic using the CICFlowMeter-V3 tool [91].

U. CIC-DARKNET2020

The Darknet, also referred to as the dark web, is a segment of the IP space where routed but inactive services and servers reside. This includes systems designed to passively receive messages without giving any active response. These systems might be part of an overlay

TABLE 10. CICDarknet2020 dataset features and description.

Category	Feature	Description
Flow duration and Packet timing	Fl-dur	Flow Duration
	Fl_iat_avg	Average Time Between two flows
	Fl_iat_std	Standard deviation of time between two flows
	fl iat max	Maximum time between two flows
	fl iat min	Minimum time between two flows
Packet Counts and	tot fw pk	Total packets in the forward Rates direction
	tot bw pk	Total packets in the backward Direction
	fl pkts	Flow packets rate
Packet Sizes	fw pkts	Number of forward packets per second
	bw pkts	Number of backward packets per second
	tot I fw pkt	Total size of packet in forward direction
	fw pkt I max	Maximum size of packet in forward direction
	fw pkt I min	Minimum size of packet in forward direction
	fw pkt I avg	Average size of packet in forward direction
	fw pkt I std	Standard deviation size of packet in forward direction
	Bw pkt I max	Maximum size of packet in backward direction
	Bw pkt I min	Minimum size of packet in backward direction
	Bw pkt I avg	Mean size of packet in backward direction
	Bw pkt I std	Standard deviation size of packet in backward direction
	fw iat tot	Total time between two packets in forward direction
	fw iat avg	Mean time between two packets in forward direction
Inter-Arrival Times	fw iat std	Standard deviation time between two packets in forward direction
	fw iat max	Maximum time between two packets in forward direction
	fw iat min	Minimum time between two packets in forward direction
	bw iat tot	Total time between two packets in backward direction

network, accessible through non-standard communication ports and protocols. Classifying Darknet traffic is crucial for categorizing real-time applications. While many approaches utilize existing datasets and ML classifiers, there's a limited exploration in employing DL for darknet traffic detection and characterization. The CICDarknet2020 dataset traffic is generated through a two-layered approach: the initial layer produces both benign and darknet traffic, while the second layer generates specific darknet traffic types such as Audio-Stream, as depicted in Table 10, Browsing, Chat, Email, P2P, Transfer, Video-Stream, and VOIP. To create a comprehensive dataset, the study integrated previously established datasets, ISCXTor2016 and ISCXVPN2016, merging the VPN and Tor traffic into their respective Darknet categories [92].

V. REAL-TIME GENERATED TRAFFIC

Using real network traffic for training and testing IDS, as outlined in the reviewed research, offers distinct advantages over using publicly accessible datasets [93] and [94]. Real network traffic provides a level of real-world relevance that synthetic datasets lack. It captures the intricate dynamics of actual network environments, reflecting the diverse behaviors of legitimate users and potential attackers. This authenticity enables the IDS to be evaluated under conditions that closely resemble those of production environments, leading to more accurate performance assessments.

Furthermore, real network traffic introduces a complexity and diversity that synthetic datasets may not fully replicate. The multifaceted nature of real traffic exposes the IDS to a wide array of patterns and interactions, allowing it to better adapt and generalize its detection capabilities. This exposure to complexity enhances the IDS's robustness, ensuring its effectiveness against a broader spectrum of potential threats [95].

An essential aspect where real network traffic excels is in dealing with emerging threats. The ever-evolving threat landscape continually introduces new attack methods. By utilizing real traffic, the IDS becomes more attuned to identifying novel attack patterns that might not be present in publicly accessible datasets [96]. This adaptive learning positions the IDS to recognize and counter emerging threats effectively.

Moreover, real network traffic accounts for scenario variability. It encompasses various situations such as shifts in user behavior, evolving application usage, and network upgrades. This variability provides a comprehensive testing and training ground for the IDS, enabling it to learn and perform well in a wide range of scenarios.

On the other hand, publicly accessible datasets, although useful for standardized comparison between different IDS implementations, often lack the complexity and authenticity of real-world network behavior. They are generated in controlled environments, which might not fully mirror the intricacies of actual networks. This controlled nature could lead to an incomplete representation of certain attack vectors or patterns encountered in real-world scenarios.

In conclusion, using real network traffic for IDS training and testing offers substantial benefits. It bridges the gap between laboratory-controlled environments and the dynamic reality of network operations. While it introduces challenges related to data privacy, ethical considerations, and potential noise in real traffic, the advantages of realism, complexity, adaptability, and responsiveness to emerging threats make this approach superior to relying solely on publicly accessible datasets.

W. MQTT-IOT-IDS2020: MQTT INTERNET OF THINGS INTRUSION DETECTION DATASET

The MQTT protocol is increasingly pivotal in IoT machine-to-machine communications, prompting an urgent need for robust IDS. Addressing the requirement for relevant IoT-specific datasets, this research introduces a novel dataset, drawn from a simulated MQTT network framework consisting of twelve sensors, a broker, a camera model, and an attacker. Five scenarios, ranging from standard operations to brute-force attacks, are meticulously documented [97]. From the primary pcap logs, the study delineated features across three distinct layers: packet-level, unidirectional flow, and bidirectional flow. In another study the author stated that the MQTT-IoT-IDS2020 dataset encompasses three feature abstraction levels specific to MQTT-integrated IoT: Packet-flow, Bi-flow, and Uni-flow. Each of these feature sets contains five distinct files, representing both standard operations and attack scenarios. The dataset has been used by the researches [98] and [99].

IV. DISCUSSION

When analyzing the strengths, weaknesses, limitations, and use concerns of different IDS datasets, it becomes evident that each dataset has distinct attributes and compromises. The KDD Cup 1999 and NSL-KDD datasets, while extensively used and facilitating comparison analysis with other research, are obsolete and may not accurately represent contemporary attack types and network traffic patterns. These datasets are artificially created and may not correctly depict real-world situations. As a result, they are better suited for early proof-of-concept studies or for comparing with previous Intrusion Detection System (IDS) research, rather than for assessing IDS against current threats or in real-world environments.

In contrast, datasets like as CICIDS 2017 and UNSW-NB15 provide a wide range of contemporary attack types and more authentic network traffic, accompanied by a substantial number of attributes for analysis. Nevertheless, these reports may not include the latest forms of attacks or network protocols and are produced in artificial settings that may not accurately reflect the intricacies of actual networks. These datasets are ideal for evaluating the performance of Intrusion Detection Systems (IDS) against a variety of attack types and in network settings that resemble real-world scenarios. However, they may not be adequate for analyzing IDS performance against the most current or sophisticated attacks.

Specialized datasets, such as ADFA-LD and ADFA-WD for host-based intrusion detection, BoT-IoT for IoT-specific attacks, CAIDA2007 for DDoS attacks, and CIC-DDoS2019 for modern DDoS attacks, offer valuable resources for studying specific attack characteristics and assessing the performance of intrusion detection systems in those particular contexts. Nevertheless, these tools could not include a wide array of attack categories and might be restricted in terms of the size of the dataset or the inclusion of the latest attack methodologies.

The Malware Traffic Analysis (MTA) dataset provides authentic malware traffic data, which is valuable for assessing the effectiveness of Intrusion Detection Systems (IDS) in detecting and mitigating malware-related risks. Nevertheless, its scope is restricted to traffic related to malware and may not include other forms of attacks or incorporate the latest malware samples or methodologies.

The ISCX-IDS-2012 dataset comprises genuine network traffic and a diverse range of attack methods, accompanied with labeled data for assessment purposes. Although it may not include the latest attack types or network protocols and is comparatively lower in size compared to other IDS datasets, it is well-suited for assessing the efficacy of IDS systems against various attack types under real-world network settings.

In the area of intrusion detection system (IDS) research, one of the most crucial and difficult challenges researchers make is picking the best relevant dataset that matches with their specific study aims and target settings. With so many dataset alternatives accessible, each with its own set of strengths, flaws, and limits, researchers are sometimes faced with a confusing terrain of choices, attempting to choose which dataset would deliver the most relevant insights and successfully support their research aims.

It takes a thorough familiarity with the benefits, drawbacks, and constraints of each choice for assessing and choosing a dataset, thus it's not a simple undertaking. Every dataset has its own unique qualities that researchers need to evaluate thoroughly. These include the variety and applicability of the included attack types, the accuracy of the network traffic patterns, the dataset's size and complexity, and whether or not it contains labelled data for assessment. In order to build reliable and efficient IDS models, it is essential to do this evaluation to guarantee that the chosen dataset is appropriate for the study goals.

Nevertheless, the existing state of intrusion detection datasets is a disjointed and diverse assortment, with each dataset providing a distinct collection of attributes and concentrating on certain areas of intrusion detection. This disjointed nature makes it more difficult to compare and validate study results across various studies and contexts, which is a major obstacle for researchers. Researchers also struggle to create IDS models that are really generalizable and relevant to different network environments due to the absence of a single, complete dataset that includes a broad variety of attack kinds, network protocols, and real-world situations.

Researchers urgently want a single dataset that can serve as a comprehensive resource for several purposes in order to tackle this problem and develop IDS research. This combined dataset has to include many characteristics for analysis, realistic patterns of network traffic, and a wide variety of modern attack methods. Researchers should be able to test their models against the most current and sophisticated threats, and the system should be built to make it easy to evaluate IDS performance in different network circumstances.

It is critical to put up a systematic framework for dataset selection before beginning to create such a unified dataset. Researchers will be able to compare and contrast current datasets according to how well they fit certain study goals and target settings with the help of this framework. When researchers think about things like the variety and importance of attack types, the realistic nature of network traffic, the availability of labelled data, and the dataset's size and complexity, they can make better decisions about which datasets to use and how to combine them.

In addition to fostering a more consistent and uniform approach to IDS research, the establishment of a systematic framework for dataset selection will make it easier to create a unified dataset. Researchers will be able to build upon each other's work more effectively thanks to the framework, which provides a uniform set of criteria and procedures for dataset assessment. This will enable increased comparability and reproducibility of study results.

The next step, after establishing the dataset selection framework, is to create the unified dataset. Researchers, business associates, and cybersecurity professionals should work together throughout this process to guarantee that the dataset includes the latest and most relevant attack types, network protocols, and real-world situations. This single dataset has to be built to be scalable, flexible, and easy to maintain so it can change and adapt to new network conditions and threats.

The availability of a unified, comprehensive IDS dataset, supported by a systematic dataset selection framework, will have a transformative impact on the field of intrusion detection research. It will provide researchers with a powerful tool to evaluate and compare their IDS models, validate their findings across diverse network settings, and develop more robust and effective intrusion detection solutions. Moreover, it will foster greater collaboration and knowledge sharing among researchers, accelerating the pace of innovation and advancing the state of the art in intrusion detection.

V. CONCLUSION

Amidst the constantly changing field of cybersecurity, the creation of efficient intrusion detection systems (IDS) serves as a strong defense against the continuous onslaught of cyber attacks. Yet, the path towards accomplishing this vital goal is filled with difficulties, with the primary issue being the demanding process of choosing the best suitable dataset for assessing and confirming IDS models. The fragmented, heterogeneous, and limited nature of IDS datasets poses a significant obstacle for researchers, impeding the progress of IDS research and preventing the fulfillment of its full potential.

In order to overcome this challenging barrier, we suggest an innovative strategy that aims to completely reshape the fundamental principles of IDS research. We strongly advocate for the creation of a single, all-encompassing IDS dataset, which will serve as a guiding light in the middle of the present disorderly situation. This dataset, created by a collaboration of researchers, industry partners, and cybersecurity experts, will be a comprehensive resource that includes contemporary attack types, realistic network traffic patterns, and real-world intrusion detection scenarios.

However, the formation of this consolidated dataset is just a single component of the problem. In order to fully harness its capabilities and guarantee its efficient application, it is essential to first build a structured framework for selecting datasets. This framework serves as a beacon for researchers, offering a systematic method to assess and compare current datasets. It enables informed decision-making and encourages a more consistent and uniform approach to IDS research.

By combining the unified dataset with the dataset selection methodology, we will bring about a new age of intrusion detection research. This period will be characterized by high levels of cooperation, comparability, and repeatability. This innovative method will enable researchers to create more resilient, efficient, and widely applicable Intrusion Detection System (IDS) models, thereby strengthening the defenses of cybersecurity and protecting networks from the always changing realm of cyber threats.

Ultimately, the key to achieving successful intrusion detection does not lay in separate and incomplete datasets, but rather in the collective and cooperative efforts of the cybersecurity community. By adopting a single IDS dataset and a systematic dataset selection process, we may overcome the existing limits and move towards a future where networks are more secure, robust, and immune to the actions of hostile individuals.

Although this comprehensive analysis of cybersecurity datasets offers helpful insights and presents a methodical approach for selecting datasets and creating a unified IDS dataset, it is important to recognize the limits and difficulties associated with this undertaking. The ever-changing and developing nature of cyber threats is a major obstacle in building a comprehensive dataset that stays up-to-date and accurately represents the situation over time. With the emergence of new attack routes and methodologies, the unified dataset must be continuously updated and expanded to ensure its ongoing efficacy. This requires a continuous endeavor and cooperation among researchers, industry partners, and cybersecurity professionals to guarantee the dataset's up-to-dateness and flexibility. Furthermore, the process of consolidating and standardizing datasets may encounter challenges pertaining to data privacy, adherence to legal regulations, and protection of intellectual property rights. This is because merging datasets from many origins and countries may give rise to intricate legal and ethical concerns. To tackle these challenges, it is necessary to carefully analyze and create strong frameworks for managing data. Furthermore, the actual execution of the suggested systematic dataset selection methodology may face difficulties associated with the presence, reachability, and excellence of current datasets. Researchers may meet challenges in accessing certain datasets or experience variations in data formats, labeling, and documentation, which might impede the efficient use of the framework. The creation of a single IDS dataset requires substantial computing resources, infrastructure, and skills to effectively manage the extensive data processing, storage, and analysis required. Ensuring the scalability, performance, and security of the dataset will be a crucial obstacle that necessitates innovative technology solutions and strong data management techniques.

Despite these limits and constraints, our research project has enormous potential for transforming the landscape of cybersecurity research and practice. By presenting a methodical process for dataset selection and arguing for the creation of a unified IDS dataset, we establish the groundwork for a more collaborative, effective, and adaptable approach to cybersecurity. Our efforts to confront the constraints and problems front on illustrate our determination to push the boundaries of what is feasible in this crucial subject. We think that our continuing research, innovation, and cooperation will make a substantial contribution to the development of a more secure and resilient digital future. As we manage the intricacies and challenges of this project, we remain committed to our purpose of empowering researchers, improving the accuracy and efficiency of intrusion detection systems, and ultimately strengthening industries' cybersecurity postures throughout the globe. The path ahead may be difficult, but the potential significance of our study makes it a worthwhile endeavor.

REFERENCES

[1] Y. Li and Q. Liu, "A comprehensive review study of cyber-attacks and cyber security; emerging trends and recent developments," *Energy Rep.*, vol. 7, pp. 8176–8186, Nov. 2021, doi: 10.1016/j.egyr.2021.08.126.

[2] A. H. Lashkari, G. Kaur, and A. Rahali, "DIDarknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning," in *Proc. 10th Int. Conf. Commun. Netw. Secur.*, Nov. 2020, pp. 1–14.

[3] S. Alzughaibi and S. El Khediri, "A cloud intrusion detection systems based on DNN using backpropagation and PSO on the CSE-CIC-IDS2018 dataset," *Appl. Sci.*, vol. 13, no. 4, p. 2276, Feb. 2023.

[4] V. Jyothsna, V. V. Rama Prasad, and K. Munivara Prasad, "A review of anomaly based intrusion detection systems," *Int. J. Comput. Appl.*, vol. 28, no. 7, pp. 26–35, Aug. 2011, doi: 10.5120/3399-4730.

[5] P. Aggarwal and S. K. Sharma, "Analysis of KDD dataset attributes-class wise for intrusion detection," *Proc. Comput. Sci.*, vol. 57, pp. 842–851, Aug. 2015, doi: 10.1016/j.procs.2015.07.490.

[6] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, Dec. 2019, doi: 10.1186/s42400-019-0038-7.

[7] H. Hindy, D. Brosset, E. Bayne, A. K. Seeam, C. Tachtatzis, R. Atkinson, and X. Bellekens, "A taxonomy of network threats and the effect of current datasets on intrusion detection systems," *IEEE Access*, vol. 8, pp. 104650–104675, 2020, doi: 10.1109/ACCESS.2020.3000179.

[8] C. G. Cordero, E. Vasilomanolakis, N. Milanov, C. Koch, D. Hausheer, and M. Mühlhäuser, "ID2T: A DIY dataset creation toolkit for intrusion detection systems," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Sep. 2015, pp. 739–740, doi: 10.1109/CNS.2015.7346912.

[9] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Comput. Secur.*, vol. 86, pp. 147–167, Sep. 2019, doi: 10.1016/j.cose.2019.06.005.

[10] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy*, 2018, pp. 108–116, doi: 10.5220/0006639801080116.

[11] A. Gharib, I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "An evaluation framework for intrusion detection dataset," in *Proc. Int. Conf. Inf. Sci. Secur. (ICISS)*, 2016, pp. 1–6.

[12] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *Social Netw. Comput. Sci.*, vol. 2, no. 6, Nov. 2021, Art. no. 420, doi: 10.1007/s42979-021-00815-1.

[13] T. Sowmya and E. A. Mary Anita, "A comprehensive review of AI based intrusion detection system," *Meas., Sensors*, vol. 28, Aug. 2023, Art. no. 100827, doi: 10.1016/j.measen.2023.100827.

[14] L. Kisselburgh and J. Beever, "The ethics of privacy in research and design: Principles, practices, and potential," in *Modern Socio-Technical Perspectives on Privacy*. Cham, Switzerland: Springer, 2022.

[15] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Comput. Intell. Secur. Defense Appl.*, Jul. 2009, pp. 1–6.

[16] I. Ortet Lopes, D. Zou, F. A. Ruambo, S. Akbar, and B. Yuan, "Towards effective detection of recent DDoS attacks: A deep learning approach," *Secur. Commun. Netw.*, vol. 2021, pp. 1–14, Nov. 2021, doi: 10.1155/2021/5710028.

[17] Z. Yang, X. Liu, T. Li, D. Wu, J. Wang, Y. Zhao, and H. Han, "A systematic literature review of methods and datasets for anomaly-based network intrusion detection," *Comput. Secur.*, vol. 116, May 2022, Art. no. 102675, doi: 10.1016/j.cose.2022.102675.

[18] A. Pinto, L.-C. Herrera, Y. Donoso, and J. A. Gutierrez, "Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure," *Sensors*, vol. 23, no. 5, p. 2415, Feb. 2023, doi: 10.3390/s23052415.

[19] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Comput. Secur.*, vol. 31, no. 3, pp. 357–374, May 2012.

[20] M. Ghurab, G. Gaphari, F. Alshami, R. Alshamy, and S. Othman, "A detailed analysis of benchmark datasets for network intrusion detection system," *Asian J. Res. Comput. Sci.*, vol. 7, no. 4, pp. 14–33, Apr. 2021.

[21] I. F. Kilincer, F. Ertam, and A. Sengur, "Machine learning methods for cyber security intrusion detection: Datasets and comparative study," *Comput. Netw.*, vol. 188, Apr. 2021, Art. no. 107840, doi: 10.1016/j.comnet.2021.107840.

[22] M. Malowidzki, P. Berezinski, and M. Mazur, "Network intrusion detection: Half a kingdom for a good dataset," in *Proc. NATO STO SAS-139 Workshop*, 2015, pp. 1–18.

[23] M. Bacevicius and A. Paulauskaite-Taraseviciene, "Machine learning algorithms for raw and unbalanced intrusion detection data in a multi-class classification problem," *Appl. Sci.*, vol. 13, no. 12, p. 7328, Jun. 2023, doi: 10.3390/app13127328.

[24] M. H. M. Yusof, A. A. Almohammedi, V. Shepelev, and O. Ahmed, "Visualizing realistic benchmarked IDS dataset: CIRA-CIC-DoHBrw-2020," *IEEE Access*, vol. 10, pp. 94624–94642, 2020, doi: 10.1109/ACCESS.2022.3204690.

[25] A. Momand, S. U. Jan, and N. Ramzan, "A systematic and comprehensive survey of recent advances in intrusion detection systems using machine learning: Deep learning, datasets, and attack taxonomy," *J. Sensors*, vol. 2023, pp. 1–18, Feb. 2023, doi: 10.1155/2023/6048087.

[26] M. Ni, "A review on machine learning methods for intrusion detection system," *Appl. Comput. Eng.*, vol. 27, no. 1, pp. 57–64, Dec. 2023, doi: 10.54254/2755-2721/27/20230148.

[27] S. Asiri, Y. Xiao, S. Alzahrani, S. Li, and T. Li, "A survey of intelligent detection designs of HTML URL phishing attacks," *IEEE Access*, vol. 11, pp. 6421–6443, 2023, doi: 10.1109/ACCESS.2023.3237798.

[28] Ö. Aslan, S. S. Aktug, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, "A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions," *Electronics*, vol. 12, no. 6, p. 1333, Mar. 2023, doi: 10.3390/electronics12061333.

[29] A. M. Nerabie, M. AlKhatib, S. S. Mathew, M. E. Barachi, and F. Oroumchian, "The impact of Arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach," *Proc. Comput. Sci.*, vol. 184, pp. 148–155, Jul. 2021, doi: 10.1016/j.procs.2021.03.026.

[30] M. Safaei Pour, C. Nader, K. Friday, and E. Bou-Harb, "A comprehensive survey of recent Internet measurement techniques for cyber security," *Comput. Secur.*, vol. 128, May 2023, Art. no. 103123, doi: 10.1016/j.cose.2023.103123.

[31] L. D'hooge, M. Verkerken, T. Wauters, F. De Turck, and B. Volckaert, "Investigating generalized performance of data-constrained supervised machine learning models on novel, related samples in intrusion detection," *Sensors*, vol. 23, no. 4, p. 1846, Feb. 2023, doi: 10.3390/s23041846.

[32] A. Thakkar and R. Lohiya, "A review of the advancement in intrusion detection datasets," *Proc. Comput. Sci.*, vol. 167, pp. 636–645, Nov. 2020, doi: 10.1016/j.procs.2020.03.330.

[33] E. Qazi, M. Faheem, and T. Zia, "HDLNIDS: Hybrid deep-learning-based network intrusion detection system," *Appl. Sci.*, vol. 13, Jun. 2023, Art. no. 4921.

[34] N. N. Mohd Yusof and N. S. Sulaiman, "Cyber attack detection dataset: A review," *J. Phys. Conf. Ser.*, vol. 2319, no. 1, Aug. 2022, Art. no. 012029, doi: 10.1088/1742-6596/2319/1/012029.

[35] A. Ferriyan, A. H. Thamrin, K. Takeda, and J. Murai, "Generating network intrusion detection dataset based on real and encrypted synthetic attack traffic," *Appl. Sci.*, vol. 11, no. 17, p. 7868, Aug. 2021, doi: 10.3390/app11177868.

[36] H. D. Moore, A. Stephens, and W. Scherer, "An understanding of the vulnerability of datasets to disparate membership inference attacks," *J. Cybersecurity Privacy*, vol. 2, pp. 882–906, Aug. 2022.

[37] A. Lara, V. Mayor, R. Estepa, A. Estepa, and J. E. Díaz-Verdejo, "Smart home anomaly-based IDS: Architecture proposal and case study," *Internet Things*, vol. 22, Jul. 2023, Art. no. 100773, doi: 10.1016/j.iot.2023.100773.

[38] M. Zipperle, F. Gottwalt, E. Chang, and T. Dillon, "Provenance-based intrusion detection systems: A survey," *ACM Comput. Surveys*, vol. 55, no. 7, pp. 1–36, Jul. 2023, doi: 10.1145/3539605.

[39] D. Protić, "Review of KDD cup '99, NSL-KDD and Kyoto 2006+ datasets," *Vojnotehnicki Glasnik*, vol. 66, no. 3, pp. 580–596, 2018.

[40] (1999). *KDD Cup 1999 Data*. [Online]. Available: https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[41] A. M. Aleesa, B. B. Zaidan, A. A. Zaidan, and N. M. Sahar, "Review of intrusion detection systems based on deep learning techniques: Coherent taxonomy, challenges, motivations, recommendations, substantial analysis and future directions," *Neural Comput. Appl.*, vol. 32, no. 14, pp. 9827–9858, Jul. 2020, doi: 10.1007/s00521-019-04557-3.

[42] Y. A. Mohamed, D. A. Salih, and A. Khanan, "An approach to improving intrusion detection system performance against low frequent attacks," *J. Adv. Inf. Technol.*, vol. 14, no. 3, pp. 472–478, 2023.

[43] A. M. Al Tobi and I. Duncan, "KDD 1999 generation faults: A review and analysis," *J. Cyber Secur. Technol.*, vol. 2, nos. 3–4, pp. 164–200, Oct. 2018, doi: 10.1080/23742917.2018.1518061.

[44] K. Adjepon-Yamoah, F. Cheng, and S. Sezer, "An investigation of deep learning for network intrusion detection systems," *Comput. Netw.*, vol. 215, Jun. 2022, Art. no. 108193.

[45] S. Kumar and A. K. Singh, "A localized algorithm for clustering in cognitive radio networks," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 33, no. 5, pp. 600–607, Jun. 2021, doi: 10.1016/j.jksuci.2018.04.004.

[46] S. Lakhina, S. Joseph, and B. Verma, "Feature reduction using principal component analysis for effective anomaly-based intrusion detection on NSL-KDD," *Int. J. Eng. Sci. Technol.*, vol. 2, no. 6, pp. 1790–1799, Aug. 2010.

[47] K. Barik, S. Misra, K. Konar, L. Fernandez-Sanz, and M. Koyuncu, "Cybersecurity deep: Approaches, attacks dataset, and comparative study," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2055399, doi: 10.1080/08839514.2022.2055399.

[48] (1999). *KDD Cup 1999 Data*. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/task.html

[49] D. A. Salih, Y. A. Mohamed, and M. Bashir, "Enhancing intrusion detection system performance against low frequent attacks using FC-ANN algorithm," *J. Eng. Sci. Technol.*, vol. 18, no. 5, pp. 2411–2431, 2023.

[50] P. Ahmadi and K. Islam, "A robust comparison of the KDDCup99 and NSL-KDD intrusion detection datasets by utilizing principle component analysis and evaluating the performance of various machine learning algorithms," *J. Scientists' Res.*, vol. 1, pp. 1–17, Nov. 2019.

[51] M. Aljabri, S. S. Aljameel, R. M. A. Mohammad, S. H. Almotiri, S. Mirza, F. M. Anis, M. Aboulnour, D. M. Alomari, D. H. Alhamed, and H. S. Altamimi, "Intelligent techniques for detecting network attacks: Review and research directions," *Sensors*, vol. 21, no. 21, p. 7070, Oct. 2021, doi: 10.3390/s21217070.

[52] R. Panigrahi and S. Borah, "A detailed analysis of CICIDS2017 dataset for designing intrusion detection systems," *Int. J. Eng. Technol.*, vol. 7, no. 3.24, pp. 479–482, 2018.

[53] K. Jeřábek, K. Hynek, T. Čejka, and O. Ryšavý, "Collection of datasets with DNS over HTTPS traffic," *Data Brief*, vol. 42, Jun. 2022, Art. no. 108310, doi: 10.1016/j.dib.2022.108310.

[54] M. Rodríguez, Á. Alesanco, L. Mehavilla, and J. García, "Evaluation of machine learning techniques for traffic flow-based intrusion detection," *Sensors*, vol. 22, no. 23, p. 9326, Nov. 2022, doi: 10.3390/s22239326.

[55] I. Sharafaldin, A. Habibi Lashkari, and A. Ghorbani, "A detailed analysis of the CICIDS2017 data set," *Adv. Intell. Syst. Comput.*, vol. 7, no. 2, pp. 479–482, 2019, doi: 10.1007/978-3-030-25109-3_9.

[56] M. Catillo, A. Del Vecchio, A. Pecchia, and U. Villano, "A case study with CICIDS2017 on the robustness of machine learning against adversarial attacks in intrusion detection," *Assoc. Comput. Machinery*, vol. 2023, pp. 1–8, Sep. 2023.

[57] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. Mil. Commun. Inf. Syst. Conf. (MilCIS)*, Nov. 2015, pp. 1–12, doi: 10.1109/milcis.2015.7348942.

[58] S. M. Kasongo and Y. Sun, "Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset," *J. Big Data*, vol. 7, no. 1, Dec. 2020, Art. no. 105, doi: 10.1186/s40537-020-00379-6.

[59] F. Bu and X. Wang, "A smart agriculture IoT system based on deep reinforcement learning," *Future Gener. Comput. Syst.*, vol. 99, pp. 500–507, Oct. 2019, doi: 10.1016/j.future.2019.04.041.

[60] Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE Access*, vol. 6, pp. 35365–35381, 2018, doi: 10.1109/ACCESS.2018.2836950. https://doi.org/10.1109/ACCESS.2018.2875782.

[61] G. Creech and J. Hu, "Generation of a new IDS test dataset: Time to retire the KDD collection," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2013, pp. 4487–4492, doi: 10.1109/WCNC.2013.6555301.

[62] K. B. Adedeji, A. M. Abu-Mahfouz, and A. M. Kurien, "DDoS attack and detection methods in Internet-enabled networks: Concept, research perspectives, and challenges," *J. Sensor Actuator Netw.*, vol. 12, no. 4, p. 51, Jul. 2023, doi: 10.3390/jsan12040051.

[63] S. Alosaimi and S. M. Almutairi, "An intrusion detection system using BoT-IoT," *Appl. Sci.*, vol. 13, no. 9, p. 5427, Apr. 2023, doi: 10.3390/app13095427.

[64] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022, doi: 10.1109/ACCESS.2022.3165809.

[65] J. M. Peterson, J. L. Leevy, and T. M. Khoshgoftaar, "A review and analysis of the bot-IoT dataset," in *Proc. IEEE Int. Conf. Service-Oriented Syst. Eng. (SOSE)*, Aug. 2021, pp. 20–27, doi: 10.1109/SOSE52839.2021.00007.

[66] A. Sarwar, M. F. Mushtaq, U. Akram, F. Rustam, A. Hamza, V. Rupapara, and S. Ullah, "IoT networks attacks detection using multi-novel features and extra tree random–voting ensemble classifier (ER-VEC)," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 12, pp. 16637–16651, Dec. 2023, doi: 10.1007/s12652-023-04666-x.

[67] H. Gebrye, "Mirai-based multi-class dataset," *IEEE Dataport*, Mar. 18, 2023, doi: 10.21227/h4ac-wr38.

[68] A. A. Alahmadi, M. Aljabri, F. Alhaidari, D. J. Alharthi, G. E. Rayani, L. A. Marghalani, O. B. Alotaibi, and S. A. Bajandouh, "DDoS attack detection in IoT-based networks using machine learning models: A survey and research directions," *Electronics*, vol. 12, no. 14, p. 3103, Jul. 2023, doi: 10.3390/electronics12143103.

[69] M. M. Alani and A. Miri, "Towards an explainable universal feature set for IoT intrusion detection," *Sensors*, vol. 22, no. 15, p. 5690, Jul. 2022, doi: 10.3390/s22155690.

[70] A. H. Lashkari, A. F. A. Kadir, H. Gonzalez, K. F. Mbah, and A. A. Ghorbani, "Towards a network-based framework for Android malware detection and characterization," in *Proc. 15th Annu. Conf. Privacy, Secur. Trust (PST)*, Aug. 2017, pp. 233–23309.

[71] M. Kim, D. Kim, C. Hwang, S. Cho, S. Han, and M. Park, "Machine-learning-based Android malware family classification using built-in and custom permissions," *Appl. Sci.*, vol. 11, no. 21, p. 10244, Nov. 2021, doi: 10.3390/app112110244.

[72] M. S. Akhtar and T. Feng, "Malware analysis and detection using machine learning algorithms," *Symmetry*, vol. 14, no. 11, p. 2304, Nov. 2022, doi: 10.3390/sym14112304.

[73] P. K. Mvula, P. Branco, G.-V. Jourdan, and H. L. Viktor, "A systematic literature review of cyber-security data repositories and performance assessment metrics for semi-supervised learning," *Discover Data*, vol. 1, no. 4, pp. 1–16, Apr. 2023, doi: 10.1007/s44248-023-00003-x.

[74] Y. Hyun, B. Huffaker, D. Andersen, E. Aben, C. Shannon, M. Luckie, and K. Claffy. (2011). *The CAIDA IPv4 Routed /24 Topology Dataset*. [Online]. Available: http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml

[75] S. Behal and K. Kumar, "Trends in validation of DDoS research," in *Proc. Int. Conf. Comput. Model. Secur.*, vol. 85, Jan. 2016, pp. 7–15.

[76] P. D. Bojović, I. Bašič ević, S. Ocovaj, and M. Popović, "A practical approach to detection of distributed denial-of-service attacks using a hybrid detection method," *Comput. Electr. Eng.*, vol. 73, pp. 84–96, Jan. 2019, doi: 10.1016/j.compeleceng.2018.11.004.

[77] NETRESEC. (2012). *U.S. National CyberWatch Mid-Atlantic Collegiate CyberDefense Competition (MACCDC)*. [Online]. Available: https://www.netresec.com/?page=MACCDC

[78] A. Alrawashdeh, I. Ahmad, N. M. Yasin, and M. Al-Khasawneh, "Deep learning techniques for intrusion detection systems: A comprehensive survey," *J. Netw. Comput. Appl.*, vol. 188, Jul. 2021, Art. no. 103094.

[79] E. Papadogiannaki and S. Ioannidis, "A survey on encrypted network traffic analysis applications, techniques, and countermeasures," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2022.

[80] A. F. Muhtadi and A. Almaarif, "Analysis of malware impact on network traffic using behavior-based detection technique," *Int. J. Adv. Data Inf. Syst.*, vol. 1, no. 1, pp. 17–25, Apr. 2020, doi: 10.25008/ijadis.v1i1.14.

[81] A. Lavin and S. Ahmad, "Evaluating real-time anomaly detection algorithms - the numenta anomaly benchmark," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 38–44.

[82] T. Schoormann, U. Hildesheim, G. Strobel, F. Möller, D. Petrik, P. Zschech, and U. Duisburg-Essen, "Artificial intelligence for sustainability—A systematic review of information systems literature," *Commun. Assoc. Inf. Syst.*, vol. 52, pp. 199–237, Aug. 2023, doi: 10.17705/1cais.05209.

[83] S. Sørbø and M. Ruocco, "Navigating the metric maze: A taxonomy of evaluation metrics for anomaly detection in time series," *Data Mining Knowl. Discovery*, vol. 1, pp. 1–42, Nov. 2023, doi: 10.1007/s10618-023-00988-8.

[84] S. Revathi and A. Malathi, "A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection," *Int. J. Eng. Res. Technol. (IJERT)*, vol. 2, no. 12, pp. 1848–1853, 2013.

[85] M. A. Hossain and M. S. Islam, "Ensuring network security with a robust intrusion detection system using ensemble-based machine learning," *Array*, vol. 19, Sep. 2023, Art. no. 100306, doi: 10.1016/j.array.2023.100306.

[86] Can. Inst. for Cybersecurity. (2017). *IDS 2017 Dataset*. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html

[87] S. García, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," *Comput. Secur.*, vol. 45, pp. 100–123, Sep. 2014, doi: 10.1016/j.cose.2014.05.011.

[88] Z. Ahmad, A. Shahid Khan, C. Wai Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 1, pp. 1–17, Jan. 2021, doi: 10.1002/ett.4150.

[89] Y. Kim, S. Hakak, and A. Ghorbani, "DDoS attack dataset (CICEV2023) against EV authentication in charging infrastructure," in *Proc. 20th Annu. Int. Conf. Privacy, Secur. Trust (PST)*, Aug. 2023, pp. 1–13.

[90] Can. Inst. for Cybersecurity. (2023). *Android Adware Dataset*. Accessed: Jan. 26, 2024. [Online]. Available: https://www.unb.ca/cic/datasets/android-adware.html

[91] Y. Lin, T. Liu, W. Liu, Z. Wang, L. Li, G. Xu, and H. Wang, "Dataset bias in Android malware detection," 2022, *arXiv:2205.15532*.

[92] I. Kovač ević, S. Groš, and K. Slovenec, "Systematic review and quantitative comparison of cyberattack scenario detection and projection," *Electronics*, vol. 9, no. 10, p. 1722, Oct. 2020.

[93] D. Samariya and A. Thakkar, "A comprehensive survey of anomaly detection algorithms," *Ann. Data Sci.*, vol. 10, no. 3, pp. 829–850, Jun. 2023.

[94] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2019, pp. 1–8.

[95] A. H. Lashkari, A. F. A. Kadir, L. Taheri, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark Android malware datasets and classification," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2018, pp. 1–7.

[96] R. A. Ahmed Farah and Y. A. Mohamed, "Adaptive immune-based system for network security," in *Proc. Int. Conf. Comput., Control, Electr., Electron. Eng. (ICCCEEE)*, Farah, Sudan, Aug. 2018, pp. 1–7, doi: 10.1109/ICCCEEE.2018.8515827.

[97] Y. Abdelgadir and A. Abdullah, "Biologically inspired model for securing hybrid mobile ad hoc networks," in *Proc. Int. Symp. Inf. Technol.*, Aug. 2008, pp. 1–14, doi: 10.1109/itsim.2008.4631673.

[98] M. A. Ahmed and Y. A. Mohamed, "Enhancing intrusion detection using statistical functions," in *Proc. Int. Conf. Comput., Control, Electr., Electron. Eng. (ICCCEEE)*, Khartoum, Sudan, Aug. 2018, pp. 1–6, doi: 10.1109/ICCCEEE.2018.8515882.

[99] Y. Mohamed and A. Abdullah, "I2MANET security logical specification framework," *The Int. Arab J. Inf. Technol.*, vol. 9, pp. 495–503, Aug. 2012.

AKBAR KHANAN (Member, IEEE) received the master's degree in computer science from Kohat University of Science and Technology, Kohat, Pakistan. He has been a dedicated full-time Faculty Member with the Department of Management Information System, College of Business Administration, A'Sharqiyah University, Ibra, Oman. Over the years, he has guided numerous graduate and undergraduate students with the College of Business Administration. He has also taken an active role in organizing various workshops, seminars, and training sessions. In addition to his academic contributions, he has a keen interest in quality auditing (QA) and has actively participated in numerous QA initiatives, including self-review auditing. Notably, he has played a significant role in shaping the future of learning. He has been instrumental in creating new bachelor's degrees in cybersecurity and internet and information technology, proposing an innovative program in data science and data analytics, and updating the Bachelor of Business Administration (Management Information System) Program. His research domains encompass a wide range, including the IoT, connected vehicles, wireless communications, networking, security issues in wireless networks, big data, cloud computing, and smart cities.

YASIR ABDELGADIR MOHAMED (Member, IEEE) received the B.Sc. degree in computer technology and the M.Sc. degree in network and computer engineering from the University of Gezira, in 2001 and 2003, respectively, and the Ph.D. degree in information technology from Universiti Teknology PETRONAS, in 2010. He is an accomplished professional. With a strong educational background and a wealth of experience in the field of information technology, he has made significant contributions to academia and research. His dedication to academic excellence is further demonstrated by his earlier academic achievements. As an Assistant Professor, he was the Head of the Information Systems, Computer Networks Department and the Statistics Department, Karary University, from 2012 to 2021, where he played a pivotal role in shaping the educational landscape. His commitment to knowledge dissemination is evident in his development and delivery of the M.Sc. courses in various Sudanese institutes. His contributions to the field of information technology extend beyond the classroom. He is a prolific researcher and has published numerous articles in reputable journals. Furthermore, he has actively participated in various conferences, where he has shared his expertise in areas, such as network security, cloud computing, software-defined networking (SDN), and the Internet of Things (IoT).

ABDUL HAKIM H. M. MOHAMED (Senior Member, IEEE) received the Ph.D. degree from the University of Liverpool, U.K., in 2016. He is the E-Learning Director and an Associate Professor with A'Sharqiyah University, Oman. Previously, he was a Lecturer with reputable U.K. higher institutions. He brings extensive experience from his past roles, such as a Lead Systems Analyst and a Senior Software Engineer in various projects. He actively contributes as a coordinator for multiple programs and serves as an external examiner. Notably, he played crucial roles in development and submissions of various program with A'Sharqiyah University, including the B.Sc. in cybersecurity, the Bachelor of Technology in Internet and Information Technology Program, the B.Sc. in Data Science and Business Analytics Program, the Bachelor of Business in Logistics and Supply Chain Management Program, and M.B.A. and B.A. in management. His research focuses on adaptive agents, HCI, and health informatics.

MOHAMED BASHIR received the M.B.A. degree in marketing management from the University of Medical Sciences and Technology, Sudan, and the Ph.D. degree in computer science from Chungbuk National University, South Korea. He is a Distinguished Assistant Professor in the management information systems (MIS) with the College of Business Administration, brings with him an illustrious career spanning two decades in pedagogy and scholarly research. He has been a linchpin in the academic community, amassing a plethora of publications in internationally recognized journals. His research proclivities encompass areas, such as machine learning, data mining, dynamic learning methodologies for data mining, and process mining. Additionally, his scholarly pursuits extend to bioinformatics, biomedicine, and knowledge-based information retrieval, always with an eye toward augmenting the efficacy and robustness of data retrieval in nascent application domains. A stalwart in academic innovation, he has been at the forefront of curriculum enhancement and development initiatives with the University of Medical Sciences and Technology. Notably, he was the Chief Architect behind the master's program in information systems, spanning three niche specializations. He was also instrumental in conceptualizing and curating curricula for an array of B.Sc., M.Sc., and Ph.D. courses, from 2012 to 2019.