IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 September 2024, accepted 1 October 2024, date of publication 3 October 2024, date of current version 17 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3473289

SURVEY

Network Intrusion Detection: An IoT and Non IoT-Related Survey

SULYMAN AGE ABDULKAREEM[1], (Student Member, IEEE), CHUAN HENG FOH[1], (Senior Member, IEEE), MOHAMMAD SHOJAFAR[1], (Senior Member, IEEE), FRANÇOIS CARREZ [1], (Senior Member, IEEE), AND KLAUS MOESSNER[2], (Senior Member, IEEE)

[1]5G/6G Innovation Centre, Institute for Communication Systems, University of Surrey, GU2 7XH Surrey, U.K.
[2]Department of Communications Engineering, TU Chemnitz, 09111 Chemnitz, Germany

Corresponding author: Sulyman Age Abdulkareem (sa0087@surrey.ac.uk)

This work was partly sponsored by Horizon 2020 Marie Skłodowska-Curie Actions under the project SwiftV2X (grant agreement ID 101008085). We would also like to acknowledge the support of the 5GIC/6GIC members.

ABSTRACT The proliferation of the Internet of Things (IoT) is occurring swiftly and is all-encompassing. The cyber attack on Dyn in 2016 brought to light the notable susceptibilities of intelligent networks. The issue of security in the realm of the Internet of Things (IoT) has emerged as a significant concern. The security of the Internet of Things (IoT) is compromised by the potential danger posed by exploiting devices connected to the Internet. The susceptibility of Things to botnets poses a significant threat to the entire Internet ecosystem (smart devices). In recent years, there has been a simultaneous evolution in the complexity and variety of security attack vectors. Therefore, it is imperative to analyse IoT methodologies to detect and alleviate emerging security breaches. The present study analyses network datasets, distinguishing between those of the Internet of Things (IoT) and those that do not, and provides a thorough overview of the findings. Our primary focus is on IoT Network Intrusion Detection (NID) studies, wherein we examine the available datasets, tools, and machine learning (ML) techniques employed in the implementation of network intrusion detection (NID). Subsequently, an evaluation, assessment, and summary of the current state-of-the-art research on IoT-related Network Intrusion Detection (NID) conducted between 2018 and 2024 is presented. This includes an analysis of the publication year, dataset, attack types, experiment results, and the advantages, disadvantages, and classifiers employed in the studies. This review emphasises research related to IoT NID that employs Supervised Machine Learning classifiers, owing to the high success rate of such classifiers in security and privacy domains. Furthermore, this survey incorporates a comprehensive analysis of research endeavours on IoT NID. Furthermore, we have identified publicly available IoT datasets that can be utilised for NID experiments, which would benefit academic and industrial research purposes. Moreover, we analyse potential prospects and future advancements. The review's findings indicate that the Internet of Things (IoT) has been substantiated by its swift proliferation in recent times, leading to even broader network coverage. This study presented conventional datasets gathered over a decade ago and current datasets published within the past decade and utilised in recent research. The survey provides a succinct overview of prevailing research trends in IoT NID for security professionals.

INDEX TERMS IoT, network intrusion detection, network dataset, machine learning, classifiers, tools.

I. INTRODUCTION

Recently the technology sector has encountered persistent difficulty in furnishing adequate security and service

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

excellence in expansive network domains. The proliferation of modern communication technologies and services and the continued expansion of interconnected network devices have engendered the emergence and intricacy of computer networks. Moreover, implementing an unrestricted communication model that can provide various services to individuals across various locations and time zones requires the integration and synchronisation of multiple technologies, thereby increasing the complexity of the network. Ensuring the efficient and dependable functioning of a contemporary and advanced network is challenging. Efficient network management aims to eliminate network susceptibilities and discourage unauthorised access.

Chen et al. [1] underscore that security breaches can manifest in diverse ways. However, their survey findings reveal that nearly half (49%) of such breaches are attributable to users' failure to adhere to fundamental security protocols when operating their networked workstations. Furthermore, the research suggests that safeguarding end-user applications that execute internet-related tasks on said workstations is challenging in the face of potential attacks. Both external and internal attacks perpetrated by insiders are potential security threats. Insider attacks can cause significant damage before detection due to the perpetrator's legitimate and extensive network access credentials and possible knowledge of network vulnerabilities. Quantifying the economic impact of security breaches poses a challenge. According to Gilmore [2], the British Office of Cyber-Security and Information Assurance has estimated that the annual cost of cybercrime to the United Kingdom (UK) is approximately £27 billion. Based on present data, a 20% increase in cyber intrusion attempts resulted in a loss of $4.8 trillion in 2020 compared to the previous year. To mitigate the potential exacerbation of losses resulting from cyber intrusion attempts, it is imperative to augment funding towards the exploration, advancement, and execution of intrusion detection systems (IDSs).

The primary function of an intrusion detection system is to accurately and promptly detect potential attacks within a system. Intrusion Detection Systems (IDSs) can identify malevolent network or workstation actions per the chosen deployment. In contrast to IDSs installed on individual workstations that protect the respective machines, IDSs implemented on networks offer safeguarding for the entire network, a more intricate undertaking. Intrusion Detection Systems (IDSs) have been developed using rules-based classifiers. The complexity of modern networks has rapidly outgrown the capabilities of rule-based design. Recent advancements in artificial intelligence (AI) have presented a promising approach for addressing the complex structure of network intrusion detection (NID) systems. Machine learning (ML) classifiers are a promising approach in artificial intelligence (AI) due to their ability to effectively handle complex systems, including NID systems. Machine learning classifiers can learn and enhance their efficacy by leveraging prior experiences. Collecting and labelling network traces for training machine learning classifiers is imperative to facilitate supervised learning. The primary aim of machine learning (ML) is to achieve the precise classification of network activities. The efficacy of the ML classifier is contingent upon the calibre and comprehensiveness of the dataset from which it learns. In the past twenty years, scholars have amassed datasets for Network Intrusion Detection (NID) training and developed various Machine Learning (ML) classifiers to acquire knowledge and detect intrusions using the datasets with some degree of success in isolation. Given that each machine learning classifier design exhibits unique strengths and limitations when deployed in diverse network environments, it is imperative to reassess these individual accomplishments and examine them comprehensively to gain insight into the current achievements, constraints, and related obstacles.

The proliferation of technology domains, including but not limited to sensors, automatic identification and tracking, embedded computing, wireless communications, broadband Internet access, and distributed services, has augmented the potential for integrating intelligent entities into our quotidian routines through the Internet. The Internet of Things (IoT) is a convergence of the Internet and smart devices that can communicate and interact. The paradigm in question has been recognised as a crucial participant within the Information and Communication Technology (ICT) enterprise, as noted by [3] in their publication on the Internet of Things. According to Statista Inc., the projected number of interconnected devices for the Internet of Things (IoT) will be 25.4 billion by 2030. According to various sources such as [4], [5], [6], [7], and [8], Cisco Systems has predicted that the Internet of Things (IoT) will result in a global revenue of $14.4 trillion and significant cost savings for businesses by the conclusion of 2022. In light of the significant advancements made in this field, we present a comprehensive overview of research studies on the Internet of Things (IoT) to examine how scholars have tackled the challenges associated with this technology, put forth potential solutions, and identified outstanding issues that require attention in the development of IoT Intrusion Detection Systems (IDS).

The present study initially presents a comprehensive summary of data compilations commonly employed in research about Network Intrusion Detection (NID) during the past twenty years. The datasets under consideration comprise both conventional ones that were gathered a decade ago and modern ones that have been published within the past decade and have been utilised in recent scholarly investigations. The following section delineates several machine learning methodologies frequently employed in non-intrusive load monitoring investigations and the corresponding software utilities utilised. Simultaneously, some survey papers found in the literature have examined the research terrain of preceding eras [9], [10], [11], [12], [13], [14], [15], [16]. Our study focuses on the initial research endeavours involving machine learning application to NID. Additionally, we present a

thorough overview of subsequent research endeavours that utilise modern datasets and advanced ML techniques, with a particular emphasis on studies related to IoT. The justification for the focus on the Internet of Things (IoT) is based on its notable proliferation in recent times, leading to a significant expansion of network coverage. Ultimately, an analysis of current research trends and projections for the future of this field of study will be presented and discussed.

Our comprehensive evaluation intends to examine the studies published in this domain between 2018 and 2022.

1) The study provides an overview of datasets frequently used in NID research over the last two decades. These include those conventional datasets collected some ten years ago and more modern ones published within the previous ten years and used in recent research works.
2) We describe some ML techniques commonly considered for NID research and the software tools involved.
3) We explore early research efforts on applying machine learning to NID and provide a comprehensive summary of subsequent research efforts using modern datasets and developing ML techniques, particularly considering IoT-related studies.
4) We discuss our views on the research trends and our take on this field of study's future.

The subsequent sections of the survey are organised as follows. Section II of the paper discusses the intrusion detection system, emphasising the network-based intrusion detection system (NID) and its various types. Section III discusses some related survey studies. Section IV analyses and categorises the NID datasets into conventional and modern datasets. Section V provides a comprehensive overview of supervised machine learning classifiers, including a summary of their respective family names and variants. Section VI of the manuscript addresses the topic of machine learning software tools and the various types that are utilised in NID experiments. Section VII provides an overview and timelines of studies related to Non-IoT NID. Section VIII of the paper presents studies related to IoT NID, including timelines and research studies, and a summary table of IoT ML-based research. Section IX presents the lessons learned, while Sections X and XI present future directions and summary. Section XII ultimately concludes the survey. Fig. 1 presents the graphical illustration of the taxonomy of this survey paper.

II. INTRUSION DETECTION SYSTEM

ICT systems and networks are repositories of crucial user data. As mentioned earlier, the data can be vulnerable to security breaches perpetrated by internal and external actors, as noted by Mukherjee et al. [17] in their work on network security. The infiltration of a networked device often proceeds gradually and inconspicuously, resulting in covert access and compromise of sensitive information. As mentioned earlier, the attack can be likened to the Yahoo data breach, which led to a financial loss of $350 million [18]. The advancement of the Internet of Things (IoT) has escalated security threats posed by attackers. As a result, it is imperative to implement effective monitoring mechanisms for systems susceptible to such vulnerabilities [19]. The continuous growth of the IoT market is attributed to technological advancements and the simplicity of producing intelligent devices. Saha et al. [20], research shows that new IoT devices are integrated into the Internet daily. The proliferation of IoT applications has resulted in their swift adoption in diverse sectors such as automotive, healthcare, manufacturing, retail, smart home systems, and space applications, as evidenced by various studies [21], [22], [23], [24], [25]. Moreover, this gives rise to a multitude of additional security issues. The 2020 IoT threat analysis conducted by Palo Alto Networks [26] reveals that a significant proportion of IoT data, about 98%, needs to be encrypted. This poses a potential threat to the security of personal data, particularly in cases where IoT devices are running outdated software, rendering them vulnerable to attacks. As per the survey findings, the primary concern associated with the Internet of Things (IoT) is the susceptibility of devices to vulnerability exploits. The prevalence of malware follows this and the adoption of poor user practices, such as the reuse of passwords.

According to Heady's architecture [27], an *intrusion* refers to a series of occurrences that threaten the Confidentiality, Integrity, or Availability (CIA) of computer system resources and services at any point in time. An intruder refers to any person or collective entity that initiates intrusion actions. The IDS security tool is designed to identify instances of unauthorised access to a computer system's services and resources. As mentioned earlier, the tool is anticipated to trigger alerts upon detection of said access. A host-based IDS is implemented on a host to oversee its services and resources' CIA. A host-based intrusion detection system (IDS) protects the host from unauthorised access [28]. A host-based intrusion detection system (IDS) is designed to monitor events related to the operating system and detect any malicious activity that could potentially compromise the host by exploiting its operating system services. Nevertheless, its scope is limited to protecting solely the host that operates the Intrusion Detection System. Thus, implementing a network-based Intrusion Detection System (IDS) is imperative to safeguard the network. In contrast to a host-based intrusion detection system, a network-based intrusion detection system is designed to observe and analyse network traffic. The system is integrated within a network and continuously observes critical network nodes for anomalous behaviour, particularly unauthorised utilisation of network services that could compromise network resources' CIA. The host and network-based Intrusion Detection Systems (IDS) utilise comparable detection techniques to identify malicious activities, which are broadly categorised as signature-based or anomaly-based detection, despite monitoring and safeguarding separate events and assets [11]. These methods can be broadly categorised as either signature-based or anomaly-

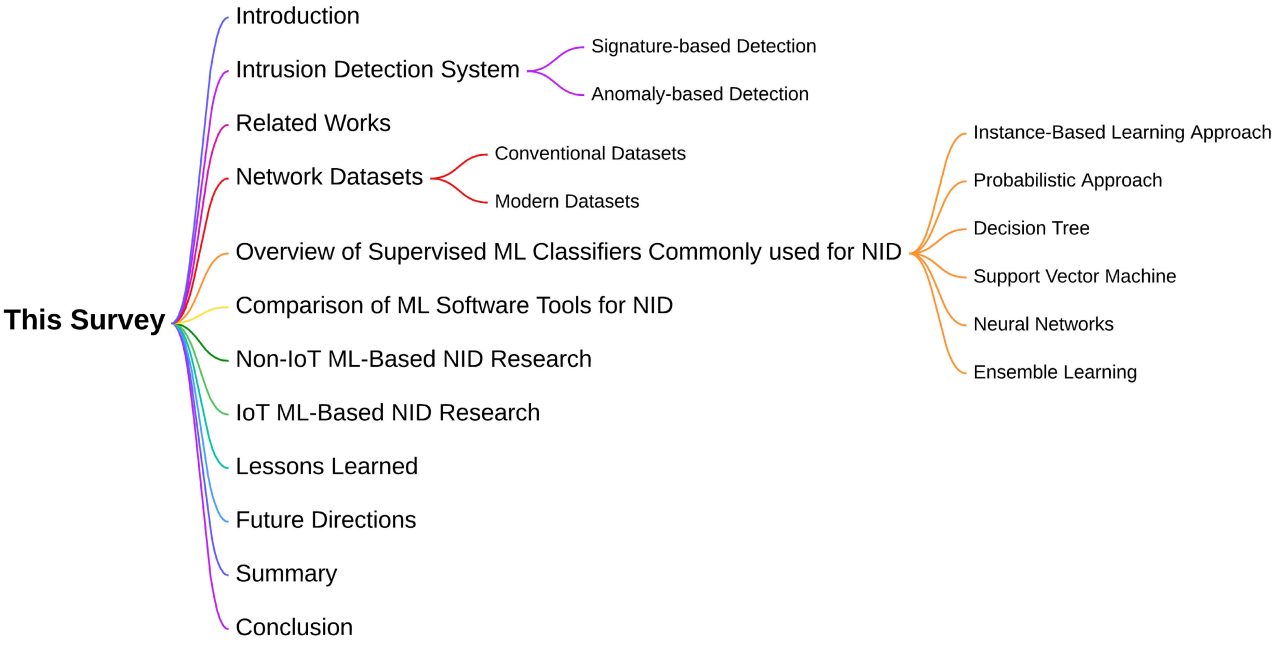

FIGURE 1. Survey paper taxonomy.

based detection. The subsequent subsections will delve into the discourse of NID and expound on the two categories of detection techniques.

A. SIGNATURE-BASED DETECTION

The notion of signature-based detection is derived from the observation that a network intrusion aimed at launching an attack often requires a series of actions to exploit the network's vulnerability. The perpetuation of an attack results in the creation of discernible traces within the network, which may manifest as a distinctive pattern [29]. The distinctiveness of the pattern renders it identifiable as the signature of the corresponding assault. Identifying a network attack can be achieved by analysing footprints left by the attack in network traces. Subsequent detection of the same attack can be facilitated by recognising the corresponding pattern in network traffic. The process of signature-based detection involves accumulating established attack patterns within a database of signatures. The Intrusion Detection System (IDS) is responsible for monitoring the network traffic that is being received and cross-referencing it with the signature database to identify any signatures that match [16]. The detection of a matching signature implies the possibility of a potential attack. The efficacy of this method in identifying known attacks has been demonstrated; however, the intricate nature of contemporary networks and the evolving patterns of attack have posed significant challenges to its effectiveness. The alteration of attack behaviour can result in a distinct pattern, thereby enabling an attacker to effectively mask its identity by modifying its behaviour to evade recognition as a known attack. In addition, it should be noted that signature-based detection mechanisms are vulnerable to zero-day attacks, which are yet to be identified [30], [31], [32].

B. ANOMALY-BASED DETECTION

Anomaly-based detection involves initially capturing and modelling the expected behaviour of a network operation. In the detection process, alarms are activated to prompt further examination in instances where the operational behaviour of the network diverges from its typical pattern [33]. Historically, conventional conduct has been simulated through domain expertise or statistical methodologies. Anomaly-based detection is advantageous over signature-based detection because it can identify novel (zero-day) attacks resulting in atypical network behaviour [11], [34]. However, with the increasing complexity of network activities, developing a manageable model to depict typical behaviour presents a more significant challenge. The advancement of machine learning technology has sparked a renewed fascination with employing machine learning methodologies for intrusion detection. Machine learning classifiers have two different applications, namely supervised and unsupervised methods. The supervised learning mode involves acquiring network traces for training, whereby these traces are categorised as either 'normal' or 'anomaly'. Upon completion of its training, the machine learning classifier is employed to classify network traffic and detect occurrences of atypical network behaviour. Furthermore, an additional categorisation can be conducted to ascertain the origin of the atypical network behaviour. *Unsupervised learning*

is a learning approach that involves the identification of regular patterns of behaviour through the training phase. The machine learning classifier that has undergone training can disclose how much the current network activity differs from the previously acquired behaviour patterns. It is possible to establish specific thresholds that can trigger warning signals in the event of network activity that deviates from the expected norm. Fig. 2 illustrates the graphical illustration of anomaly-based detection using ML. This survey study will focus on NID's supervised machine learning technique.

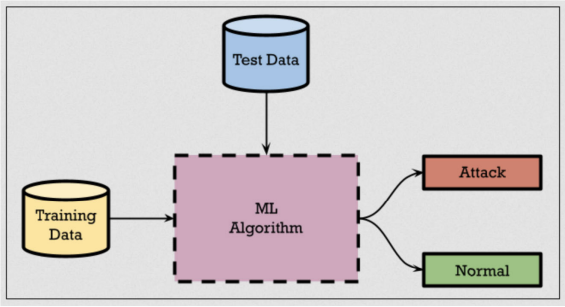

FIGURE 2. Framework for anomaly-based detection using ML.

III. RELATED WORKS

Intrusion detection has been the subject of numerous surveys in recent years. Axelsson [9] presents a taxonomy for Intrusion Detection Systems (IDS), which is subsequently employed to analyse and classify various conceptual studies. The taxonomy encompasses two components: a categorisation of the identification methodology and diverse implementation tactics of the Intrusion Detection System. Lunt [35] suggests a study on the advancements in automatic audio trail examination procedures and intrusion detection systems over the past few years. The research by Tsai et al. [10] investigates 55 scholarly papers published between 2000 and 2007. This study aims to identify the methodologies employed, the research conducted, and the potential areas for future investigation in machine learning. Liao et al. [11] seek to present an additional detailed picture for a thorough examination. The authors provide a lexicon for categorising contemporary Intrusion Detection Systems (IDSs) through their comprehensive analysis and sophisticated organisation. In their scholarly work, Agrawal and Agrawal [12] provide an extensive overview of data mining (DL) techniques for detecting anomalies. Their survey aims to provide a comprehensive understanding of the various methods available to researchers in this field to facilitate future studies.

Buczak and Guven [13] provide a comprehensive analysis of machine learning (ML) and data mining (DM) techniques utilised in cybersecurity analysis to facilitate intrusion detection (ID) in their systematic review. A concise introductory explanation accompanies each machine learning/data mining technique. The KDD99 dataset was subject to a review of intrusion detection systems (IDS) and machine learning (ML) techniques by Ozgur and Erdem [14]. The study by Ahmed et al. [15] investigates four crucial techniques for detecting anomalies: categorisation, statistics, computational modelling, and grouping. This report aims to address the gaps in the existing literature by utilising datasets for Intrusion Detection Systems (IDS). In their publication, Khraisat et al. [16] provide an overview of contemporary Intrusion Detection Systems (IDS), which involves a comprehensive examination of recent research and an evaluation of data commonly utilised for performance evaluation. The text underscores the significance of identifying research issues that can effectively address the employment of deceptive tactics by malicious actors to evade detection, thereby enhancing the security of computer systems. The study by Fernandes et al. [36] aims to evaluate the critical components of ID comprehensively. This includes assessing a foundational examination and exploring the field's most pertinent approaches, methodologies, and technologies.

Tabassum et al. surveyed to examine the latest intrusion detection systems (IDSs) for Internet of Things (IoT) systems. Their investigation focused on hybridisation and cognitive strategies and aimed to provide characterisation and analysis of these systems. The findings of their study are reported in their publication [37]. In addition, the text offers a comprehensive analysis of Internet of Things (IoT) elements, networking equipment, and potential security breaches, underscoring the criticality of Intrusion Detection Systems (IDS) in layered and protocol-based approaches. Finally, the study assesses the constraints and advantages of each approach to suggest potential novel pathways for the adoption of IDS. The authors, Chaabouni et al., investigated the contemporary design of Network Intrusion Detection Systems (NIDS) and the available datasets and free and open-source network monitoring technology. Their findings are presented in their publication [38]. This study conducts an analysis, assessment, and comparison of current NIDS methodologies in the context of the IoT ecosystem, with a focus on strategy, recognition methodology, verification tactics, addressed vulnerabilities, and technique implementations. The manuscript pertains to conventional and machine learning-based network intrusion detection systems (NIDS) methodologies and examines developing patterns. Table 1 summarises the domains covered in previous IDS surveys and those incorporated in the current investigation.

In contrast to prior literature reviews, the study represents a comprehensive survey of datasets utilised in NID research about IoT in recent years. Our analysis focuses on using the IoTs dataset in the field of research, which is a relatively recent development. The research we examined is current, spanning from 2018 to 2022. Table 1 summarises the literature review, indicating that our study solely encompassed all five criteria metrics utilised for analysing the reviewed works. The primary objective of the survey conducted by Sicato et al. [39] is to provide a comprehensive analysis of

the current Intrusion Detection Systems (IDS) for Internet of Things (IoT) environments, computer security risks, and pertinent issues and inquiries that have been explored and resolved. The study introduces a cloud-based framework focusing on a software-defined Intrusion Detection System (IDS) to establish a secure Internet of Things (IoT) environment. As per the evaluation of the implementation outcomes, the proposed framework exhibits superior identification and precision compared to conventional methodologies.

Access Control, a mechanism that restricts the viewing and utilising of resources in a computing environment, is another critical network security aspect. It pertains to the establishment of permissions for systems and users. Role-Based Access Control (RBAC), Discretionary Access Control (DAC), Mandatory Access Control (MAC), and Attribute-Based Access Control (ABAC) are all forms of access control. IDS and access control systems collaborate to safeguard an environment. IDS monitors for unauthorised access attempts and other suspicious activities, while access control restricts access to resources. The security posture of an organisation is improved by the combination of IDS and access control, which constitute a defence-in-depth strategy. This strategy prevents unauthorised access and detects potentially malicious activities. The research conducted by [40] and [41] delves more deeply into access control, a topic that is beyond the scope of our current investigation.

IV. NETWORK DATASETS

Network datasets are crucial to implementing the supervised machine learning methodology for NID. The provision of data is essential for training machine learning classifiers capable of detecting the typical behaviour of a given network. Acquiring network traces that effectively capture anomalous activities while accurately representing a network's usual behaviour is crucial, as machine learning classifiers rely on datasets for learning purposes.

Network traces contain a series of protocol packets describing network activities and services. Key features of each packet, such as protocol and port numbers, source and destination addresses, packet size, and others, are extracted to form a dataset. Each record in the dataset is a set of extracted features or metadata describing a packet. Each record also contains a label specifying its behaviour, which can either be a normal activity or a malicious action with or without further description of the malicious action. While recording each packet as a record preserves details, a malicious action or an attack often performs a sequence of actions that generates a chain of packets. Each packet may not constitute a valid attack, but the entire chain describes the attack behaviour. As a result, it is preferable to merge packets from the same chain into a single network flow record rather than utilising each packet as a separate record to create the dataset. The features of a network flow are slightly different from those of the individual packet. Many features of packet-based data, such as protocol, port numbers, and source and destination addresses, are shared with flow-based data. Flow-based data may also include additional features that cannot be represented using packet-based data. These features include the duration of the flow, the number of packets involved, and the average packet size.

Network datasets are gathered from a real-world network, a controlled test bed, or a simulated network environment. While datasets derived from real-world networks are preferable because they represent real-world network behaviours, the creation procedure can be time-consuming, and the resulting datasets are environment-specific. Collecting network datasets from a test-bed or simulated network enables much freedom in the network environment's architecture. Nonetheless, the requirement for a test bed can constrain network design, whereas simulated networks give a fully controlled environment capable of managing a network's scale and complexity. As a result, nearly all HID network datasets are derived from a simulated network environment. Kyoto 2006 [45] is an effort to create a dataset from a real-world network, and it is one of the few publicly accessible datasets derived from such a network. Despite this effort, its use in the literature for machine learning-based NID is quite limited.

The network dataset's metadata and data should be made publicly available to encourage the usage of a given network dataset for machine learning-based NID classifier development. Wilkinson et al. [46] identify four fundamental characteristics that a research dataset should adhere to: Findability, Accessibility, Interoperability, and Reusability (FAIR). According to FAIR principles, a dataset referenced in a research paper should be discoverable, and its information should be publicly accessible. Metadata and data should be accessible to various software and operating systems. The network environment, collected metadata, and data should be precisely documented for replicating and reusing. Fortunately, most network datasets for NID described in the literature comply with FAIR, except for a few ''private'' datasets that are not publicly available. This survey will explore some of the most frequently utilised publicly available datasets in historical and recent studies. We refer to datasets utilised in earlier research as ''conventional'' datasets, while those used in recent research are called ''modern'' ones.

A. CONVENTIONAL DATASETS

The datasets characterised as ''conventional'' in this survey article have existed for a period exceeding a decade. Owing to their extensive historical background, they are frequently employed as reference points in studies about NID. Although modern network design and emerging attacks are not described, the relevance of conventional attacks in ML-based NID research is maintained by using these attacks to test the ML classifier. KDD99 and NSL-KDD are among the noteworthy datasets. The summarised information can be found in Table 2. The datasets exhibit variations in size, features, and categories of intrusions

TABLE 1. Comparison between this and previous surveys: (✓: in the survey, ✗: not in the survey).

Survey	Year	IDS Types	Network Datasets	ML CLassifiers Overview	Software Tools Overview	Overview of ML-Based Studies	IoT Related Discussion
Lunt [35]	1988	✓	✗	✗	✗	✗	✗
Axelsson [9]	2000	✓	✗	✗	✗	✗	✗
Tsai *et al.* [10]	2009	✓	✓	✓	✗	✓	✗
Liao *et al.* [11]	2013	✓	✗	✓	✗	✗	✗
Agrawal and Agrawal [12]	2015	✓	✗	✓	✗	✓	✗
Buczak and Guven [13]	2016	✓	✓	✓	✗	✓	✗
Ahmed *et al.* [15]	2016	✓	✓	✓	✗	✓	✗
Ozgar and Eredem [14]	2016	✓	✓	✓	✓	✓	✗
Khraisat *et al.* [16]	2019	✓	✓	✓	✗	✓	✗
Fernandes *et al.* [36]	2019	✓	✗	✓	✗	✓	✗
Tabassum *et al.* [37]	2019	✓	✗	✗	✗	✓	✓
Chaabouni *et al.* [38]	2019	✗	✓	✗	✗	✓	✓
Sicato *et al.* [39]	2020	✗	✗	✗	✗	✓	✓
He *et al.* [42]	2023	✗	✗	✗	✗	✓	✗
Alotaibi *et al.* [43]	2023	✗	✓	✗	✗	✓	✗
Santhosh *et al.* [44]	2023	✗	✗	✗	✗	✗	✓
This Study	2024	✓	✓	✓	✓	✓	✓

incorporated. Therefore, it is crucial to meticulously assess them and choose the most suitable one that aligns with one's research investigation and implementation. Furthermore, it is crucial to acknowledge that specific datasets may possess limitations or biases. Hence, it is imperative to consider these aspects when interpreting the outcomes of experiments.

B. MODERN DATASETS

Various research institutions initiated the development of supplementary datasets that capture modern network configurations to mitigate the limitations of conventional datasets. These supplementary datasets encompassed Internet of Things (IoT) traces and nascent cyber attacks. As mentioned earlier, the datasets portray current network traffic scenarios and a broad spectrum of intrusion attacks while also remedying the limitations of initial datasets. The subsequent sections provide an overview of diverse contemporary datasets commonly employed while adhering to the FAIR principles. The datasets in question are of recent origin and have been specifically crafted to capture modern network traffic patterns and cyber attacks, emphasising those unique to IoT networks. Table 3 summarises the modern datasets, including the dataset names, authors and a brief description detailing the collection environment and the number of features.

Given the variations in size, attributes, and types of attacks and devices encompassed in these datasets, conducting a comprehensive analysis and selecting the most pertinent ones for one's research scope and implementation is imperative. As stated earlier in the previous subsection, remembering that specific datasets may have limitations or biases is imperative. Hence, it is essential to consider these factors while assessing the results of experiments.

V. OVERVIEW OF SUPERVISED ML CLASSIFIERS COMMONLY USED FOR NID

A supervised ML classifier is used to learn the relationship between a set of input instances and the output classes of different features [60]. Its objective is to establish the belonging of an input instance to one of the output classes. In the context of NID, each input is a feature vector describing the features of either a packet for packet-based datasets or a flow for flow-based datasets. The output classes always include a normal class and a range of attack types, often grouped into several categories [61]. The ML classifier learns the relationship between the input features and the corresponding class during the training by fitting its model to achieve the mapping of belonging. A trained ML classifier is then tested using a different set of data not used during the training, and its accuracy is measured, along with other performance metrics.

Since DARPA released the initial network dataset, multiple machine learning classifiers have been evaluated in the literature for their performance and applicability to NID [62]. Specific classifiers are designed for binary classification, which entails classifying a network action as normal or attack. Other classifiers operate as multiclass classifiers, distinguishing between normal and attack instances and classifying each detected attack instance according to its type. Machine learning classifiers may apply a specific model to suit the data during training. The approach utilised in the model is frequently used to classify machine learning classifiers into multiple family groupings. Decision trees (DT), rule-based learning approaches, support vector machines (SVM), neural networks (NN), and ensemble learning approaches like bagging and boosting are some of the notable ML families employed in NID. This section will revisit these machine learning classifier families and quickly outline their learning

TABLE 2. Summary of conventional datasets.

Dataset	Authors	Description
DARPA 1998	Lippmann *et al.* [47] in 1998	The collection of network traffic has 2.5 million instances that illustrate various forms of attacks, each characterised by 35 features derived from network packet headers.
KDD 99	Stolfo *et al.* [48] in 1999	An attack network traffic dataset comprising 5 million instances with 41 features derived from network packet headers.
Kyoto 2006+	Song *et al.* [49] in 2006	The network traffic dataset comprises 9 million instances of various attacks, each characterised by ten features derived from network packet headers.
NSL-KDD	Tavallaee *et al.* [50] in 2009	An extensive network traffic dataset comprising 4 million examples of various forms of attacks, with 41 features derived from network packet headers.
CTU-13	Garcia *et al.* [51] in 2011	The CTU-13 constitutes a collection of botnet traffic collected at CTU University in the Czech Republic in 2011. The objective of the dataset was to obtain a comprehensive collection of authentic botnet traffic interspersed with normal and background traffic. The CTU-13 dataset comprises thirteen botnet sample captures referred to as scenarios.
MACCDC	MACCDC [52] between 2010-2012	The dataset consists of network traffic logs obtained from the Mid-Atlantic Collegiate Cyber Defence Competition and encompasses a range of attack scenarios.
ADFA-LD12	Creech and Hu [53] in 2013	ADFA-LD12 is specifically intended for anomaly-based systems, not signature recognition IDS. Each of the three data groups comprises unprocessed system call traces.

techniques and significant family variations. Table 4 gives an overview of the ML classifier family and their respective variants.

A. INSTANCE-BASED LEARNING APPROACH

In instance-based learning, the classification of an instance is generally performed based on some measurement of the similarity of trained data. The training is mainly a process of memorizing and storing the trained data in a specific structure for efficient retrieval [63]. During the classification process, the input instance is compared with each instance in the trained data to find the most similar individual or a set of instances from the trained data to infer its class belonging. Due to the lack of active learning during the training, this approach is called ''lazy'' instead of ''eager'' learning. While this approach is widely used in NID studies because of its simplicity, its classification performance is often inferior to the eager learning approach. The k-Nearest Neighbour (kNN) is a commonly used instance-based learner, and IBk implements kNN in WEKA with various customization.

B. PROBABILISTIC APPROACH

Sometimes referred to as the statistical approach, the probabilistic approach applies a probability framework to fit its model during training. The main characteristic of this approach is the modelling of uncertainty based on a sufficient amount of observed data labelled with appropriate classes, and the employed probability model is used to infer the classes of unobserved data [64]. Some significant variants in this family applied in NID include Naive Bayes (NB), BayesNet (BN), Logistic Regression (LR), Simple Logistic, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Averaged One Dependence Estimator (AODE).

C. DECISION TREE (DT)

In DT ML learners, the fitting model is based on decision-making branching out in a tree-like structure to find the most appropriate belonging class for input through a series of decision-making. Generally, a decision tree has four essential elements: root node, decision node, leaf node, and branch. The model illustrates all viable outcomes of the decision using the branching method. If-then rules represent the learned tree, with the decision node specifying a test over some data features.

Due to the vastly available techniques in making a decision, there are many variants in the DT ML family. For example, Iterative Dichotomiser 3 (ID3) utilizes information gain and entropy in selecting the optimal feature when classifying inputs [65]. Information gain evaluates how well a given feature separates the training samples based on their target classification, while entropy specifies the impurity of an arbitrary collection of samples. Another variant known as Classification and Regression Trees (CART) utilizes the Gini Index when generating decisions that perform well when dealing with data with missing values. Other variants in the DT ML family demonstrated for NID include C4.5 (or its Java implementation J48) and C5.0, extensions of ID3, and other approaches such as Decision Table, ZeroR, PART (Projective Adaptive Resonance Theory) and OneR. DT is good for multiclass classification problems. However, it has the pitfall of over-fitting when evaluated with noisy data, consequently reducing its accuracy performance.

D. SUPPORT VECTOR MACHINE (SVM)

SVM is a supervised learning method used in ML for classification. In performing its classification task, it constructs a hyper-plane or a set of hyper-planes in a high dimensional space that segments input data from each other. To achieve this, the model manipulates input data using kernel functions to map the data to a trans-

TABLE 3. Summary of some modern datasets.

Dataset	Authors	Description
UNSW-NB15	Moustafa and Slay *et al.* [54] in 2015	The collection comprises 2.5 million network-based attacks, each with 47 features derived from network packet headers.
NGIDS-DS	Haider *et al.* [55] in 2016	This packet-based dataset contains five days of data collection on two networks. Network 1 produced normal and attack traffic, while Network 2 served as these organisations' collective victim network or sensitive cyberinfrastructure.
VPN-nonVPN (IS-CXVPN2016)	Draper-Gil *et al.* [56] in 2016	A network traffic dataset for different types of VPN and non-VPN traffic, containing 10 million instances with 41 features extracted from network packet headers.
TRAbID	Viegas *et al.* in 2017 [57]	The authors introduced a novel approach for developing intrusion databases that accurately represent the intended features of such datasets. The data are collected in a controlled setting that replicates the typical activity of users and attackers on the network.
CICIDS2017	Sharafaldin *et al.* in 2017 [58]	The dataset is provided in packet or bidirectional flow formats. It captures data from a simulated network environment for five days using CICFlowMeter software. CICIDS2017 is another effort that addresses the shortcomings of conventional datasets by producing a dataset that can represent the modern-day network.
CSE-CIC-IDS2018	Shiravi *et al.* in 2018 [58]	This dataset includes network traffic logs from various sources, including web and email traffic, and a diverse set of attack scenarios.
CICDDoS2019	Sharafaldin *et al.* in 2019 [59]	The dataset used CICFlowMeter-V3 for traffic generation in a simulated network environment, and the dataset contains two days of packet records carrying 87 features. The simulated network comprises two distinct categories of networks: the victim and the attack network.

formed space to achieve data segmentation via hyperplanes. The underlying kernel function characterizes SVM, and some commonly used kernel functions include linear, polynomial, Gaussian, and Radial Basis Function (RBF). According to Buczak and Guven [13], SVM can yield excellent accuracy when the classification task involves two classes, as it has a simple decision boundary that reduces over-fitting.

E. NEURAL NETWORKS (NNs)

NNs are artificial neurons inspired by the biological neural system, which functions similarly to the human brain [66]. A NN often consists of three main elements: an input layer, one or more hidden layers, and an output layer. Each layer contains several neurons connected to other neurons in the next layer with some weights. The connection is carried forward from one layer to the next until it reaches the final output layer. Recurring connections from a layer back to some previous layer may be possible. Input features are fed into the input layer, producing outputs at the output layer. Each output neuron reflects the probability that the input instance corresponds to each class. The backpropagation classifier is often used during training to fit the input features to the output class by adjusting the connection weights. Due to the complexity of the structure, a sizable dataset is often necessary to train the NN classifier accurately.

The flexibility of NN design has led to many NN variants with different characteristics and functions. Some NN variants that have been applied in NID research are Feed-Forward Neural Networks (FFNN), sometimes known as Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), and Radial Basis Function Networks (RBFN).

F. ENSEMBLE LEARNING

The ensemble learning approach uses multiple learning classifiers to perform better than a single learning classifier. Each learner or classifier in an ensemble learner can be a weak learner without the ability to deliver an accurate classification outcome; combining these weak learners can become a single strong learner and produce a good outcome [67].

TABLE 4. Summary of commonly used ML classifiers for NID.

ML Classifier Family	Variants
Instance-based	kNN (or IBk)
Probabilistic	NB, NB Updatable, NB Multinomial, BN, LR, Logistic, Simple Logistic, LDA, QDA, AODE, Online NB, Online AODE
Decision Tree (DT)	ID3, CART, C4.5 (or J48), J48 Graft, C5.0, REPTree, Enhanced J48, Decision Table, ZeroR, OneR, JRip, PART
Support Vector Machine (SVM)	SVM-Linear, SVM-Poly, SVM-Gaussian, SVM-RBF, Chi-SVM
Neural Networks (NN)	FFNN (or MLP), RNN, RBFN, Voted Perceptron
Bagging	RF, RT, Extreme Tree Classifier, Random Committee
Boosting	AdaBoost, GB, XGBoost

Ensemble learners can be either homogeneous or heterogeneous. Homogeneous ensemble learners employ the same weak learning model but train each model on a subset of the dataset. On the other hand, heterogeneous ensemble learners apply different weak learning models with all models training using the same dataset. For homogeneous ensemble learners,

bootstrap aggregating (bagging) and boosting are methods used to combine weak learners to form strong learners [68]. For heterogeneous ensemble learners, stacking is often used to combine the outcomes of different models.

Bagging, boosting and stacking are different ways to aggregate outcomes from weak learners [69]. In bagging, each weak learner is trained in parallel using various sections of the dataset, and the results of weak learners are consolidated by voting. Random Forest (RF), Random Tree (RT), and Extreme Tree Classifier are three bagging ensemble approaches often employed in machine learning-based HID. In boosting, weak learners are trained progressively, each weak learner attempting to compensate for a previous weak learner's classification error, increasing classification accuracy. Examples of the popular boosting technique include AdaBoost, Gradient Boost (GB) and eXtreme Gradient Boost (XGBoost).

Stacking is a component of the heterogeneous ensemble learning paradigm. It encompasses two tiers of learning: base learning and meta-learning. Base learning leverages several weak learners trained using the same input data set to generate classification results. In meta-learning, the optimal combination of results from the weak learners is learned to provide the ultimate output. Table 4 summarized the commonly used ML classifier for NID.

VI. COMPARISON OF ML SOFTWARE TOOLS FOR NID

Numerous software tools have been utilised for ML-based NID research throughout the years. These tools are essential to training and testing ML classifiers, allowing other researchers to reproduce the results under the same setup. The ease of using a tool for ML training and testing often directly contributes to its popularity among researchers. WEKA, Python, R, and MATLAB are widely used tools. According to Ozgur and Erdem's 2016 review [14], WEKA is the most extensively utilised machine learning (ML) classification tool in the intrusion detection domain. While its popularity remains at the top, Python's position is now challenged, which is particularly powerful for ML and data science problems. Table 5 gives more insight into the ML software tools used for NID and how they differ. The presented table is not exhaustive, as numerous other machine learning software tools can be utilised for NID. Nonetheless, the tools listed in the table are among the most widely used.

TABLE 5. Comparison of some ML software tools for NID.

Tool Name	Developer	Language	License	Supported OS	Framework Type
Tensor Flow	Google	Python	Apache 2.0	Windows, Linux, & MacOS	Deep Learning
PyTorch	Facebook	Python	BSD	Windows, Linux, & MacOS	Deep Learning
Keras Flow	Francois Chollet	Python	MIT	Windows, Linux, & MacOS	Deep Learning
Scikit-learn	Open-source community	Python	New BSD	Windows, Linux, & MacOS	Machine Learning
H2O.ai	H2O.ai	Java, Python & R	Apache 2.0	Windows, Linux, & MacOS	Machine Learning
Apache Spark MLlib	Apache Software Foundation	Java, Scala, Python & R	Apache 2.0	Windows, Linux, & MacOS	Machine Learning
WEKA	University of Waikato	Java	GPL	Windows, Linux, & MacOS	Machine Learning
Rapid Miner	Rapid Miner	Java	AGPL Commercial	Windows, Linux, & MacOS	Machine Learning
MATLAB	Math Works	MATLAB	Proprietary	Windows, Linux, & MacOS	Machine Learning
Python	Open-source community	Python	Open-source	Windows, Linux, & MacOS	Machine Learning & Deep Learning

VII. NON-IoT ML-BASED NID RESEARCH

The DARPA dataset was created in 1998 and was updated in 1999. The data collection included raw TCP/IP packet traces from legitimate and malicious networks. They built the KDD99 dataset in 1999 by extracting essential packet features from the DARPA dataset to make data manipulation easier. Initially, intrusions were identified by comparing ongoing network events to pre-defined standard norms using statistical analysis. Due to difficulty spotting attacks, academics suggested alternate ways, such as machine learning. The KDD99 dataset allows researchers to train and evaluate machine learning classifiers readily. kNN and SVM were formerly the mainstays of NID research. During the early stages of the research, only a few ensemble learners were seen [10]. Apart from optimising the machine learning classifier, researchers started using feature selection techniques to remove redundant features from the dataset, resulting in improved classification performance. During this time span, [9], [10], [11] surveyed NID research. Over the last decade, considerable research has been done on NID employing KDD99 machine learning classifiers. The KDD99 dataset has been found to have some significant shortcomings, such as an overabundance of redundant recordings of similar packet properties, a lack of data for particular classes in the training set, a disproportionate ratio of normal-to-attack instances, and skewed performance results, as noted in previous research [50], [86]. In 2009, Tavallaee et al. [50] introduced the NSL-KDD dataset as an enhanced version to tackle the concerns mentioned earlier. Fig. 3 presents and illustrates the Non-IoT IDS Architecture.

The NSL-KDD dataset received widespread acceptance. Its upgrade ensured that machine learning classifiers were trained effectively. As a consequence, machine learning

TABLE 6. Summary of IoT datasets.

Dataset	Authors	Description
N-BaIoT	Meidan *et al.* in 2017 [70]	The collection captures network data from nine commercially available IoT devices that have been proven to be under network intrusion attacks. This dataset contains 115 statistical features that could be negatively affected by a hostile attack.
DS2OS	Pahl and Aubet in 2018 [71]	The dataset consists of generated data collected within a virtual IoT environment based on DS2OS. The system architecture of the IoT environment consists of a set of microservices that communicate via the message-queuing telemetry transport (MQTT) protocol. The dataset consists of thirteen features established by monitoring the connections between seven different virtual state layer (VSL) service types.
WUSTL-IIOT-2018	Teixeira *et al.* [72] in 2018	This dataset was created using a testbed for the Supervisory Control and Data Acquisition (SCADA) system. The testbed comprises a control system for a water storage tank and a stage at which water is treated and distributed. Highly advanced cyber-attacks were carried out on the testbed. Network traffic was intercepted during the attacks, and features were collected from the traffic to generate the dataset.
Bot-IoT	Koroniotis *et al.* in 2018 [73]	The dataset includes IoT and non-IoT traffic and traffic utilised by botnets. A realistic testbed generates this dataset, and its features are appropriately annotated. Therefore, new features were created to boost the ML classifier model's prediction performance. Labelled features define an attack flow, its categories, and subcategories.
Kitsune	Mirsky *et al.* in 2018 [74]	The dataset is a collection of nine network attack datasets captured from an IP-based commercial surveillance system and a network of IoT devices. The dataset contains millions of network packets and different cyberattacks. All collected records were labelled as either Normal or Attack.
WUSTL-IIoT	Zolanvari *et al.* [75] in 2019	The dataset was gathered via a testbed established at Washington University in St. Louis to conduct practical IIoT operations, specifically targeting cyber threats that are more pertinent to IIoT systems.
IoT Network Intrusion Dataset	Kang *et al.* [76] in 2019	A dataset was generated with two conventional smart home devices: an SKT NUGU (NU 100) and an EZVIZ Wi-Fi Camera (C2C Mini O Plus 1080P), together with a small number of personal computers and mobile phones. These devices were interconnected on a single wireless network, which was used to replicate different attacks using Nmap.
IoT-23	Parmisano *et al.* [77] in 2020	This collection comprises 20 instances of malware executed on IoT devices and three benign IoT device traffic instances. Initially released in January 2020, the publication included captures from 2018 to 2019. The IoT network traffic was recorded in the Stratosphere Laboratory, AIC group, Frontiers of Electronics Laboratory, CTU University, Czech Republic.
TON_IoT	Alsaedi *et al.* [78] in 2020	The collection comprises several data sources obtained from Telemetry datasets of IoT and IIoT sensors, Windows 7 and 10 operating system datasets, Ubuntu 14 and 18, and TLS and Network traffic datasets.
MedBIoT	Guerra *et al.* [79] in 2020	The dataset is derived from a network of moderate size, consisting of 83 devices. Various gadgets, such as smart locks, switches, fans, and light bulbs, integrate actual and simulated components from different categories.
MQTT-IoT-IDS2020	Hindy *et al.* [80] in 2020	The dataset is created using a simulated MQTT network structure. The network includes twelve sensors, a broker, a simulated camera, and an attacker.
X-IIoTID	Al-Hawawreh *et al.* [81] in 2021	The dataset is a meticulously designed simulation of current perpetrators' strategies, methods, and processes, as well as the actual operations of IIoT systems. This includes devices in industrial control loops, edge, mobile, and cloud traffic and activities, the behaviours of their new connectivity protocols and services, the various communication patterns, and the high volume network traffic and events of the systems.
CIC IoT Dataset 2023	Dadkhah *et al.* [82] in 2022	This dataset is a state-of-the-art data collection for intelligent identification and intrusion detection of sixty IoT devices. The operation of these devices is documented to utilise a range of protocols, such as IEEE 802.11, Zigbee-based, and Z-Wave. It consists of several stages of individual devices and a wide range of events representing the simulated network dynamics of a smart home.
Edge-IIoT	Ferrag *et al.* [83] in 2022	This dataset consists of data from many replicated edge-IIoT systems, including various attacks and normal activities. The collection comprises data from several IIoT devices, encompassing sensors, edge nodes, and cloud servers.
CICEV2023	Kim *et al.* [84] in 2023	The dataset was generated to enhance the analysis of EV charging systems and offer training and testing capability for classifiers designed to detect DDoS attacks. To generate this dataset, the authors created a simulator to replicate several electric vehicles (EVs), charging stations (CSs), and a grid station (GS) of a charging infrastructure network. They also incorporated four attack scenarios.
CICIoV2024	Neto *et al.* [85] in 2024	CICIoV2024 is a comprehensive security dataset designed to fill the Internet of Vehicles (IoVs) cybersecurity gap. The dataset presents a comprehensive analysis of intra-vehicular communications obtained through rigorous testing on the Electronic Control Units (ECUs) of a 2019 Ford vehicle. This dataset is a standard against which enhanced cybersecurity solutions can be developed in the Internet of Vehicles (IoV).

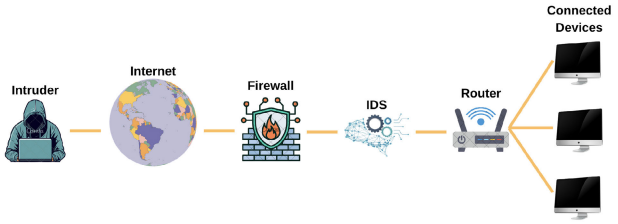

FIGURE 3. Non-IoT IDS Architecture.

classifiers enhance detection capabilities. The encouraging results sparked more studies into machine learning classifiers. Among the 149 ML-based NID publications reviewed by Atilla and Hamit in [14], 9, 29, and 133 used DARPA, KDD99, or NSL-KDD datasets, with some using multiple datasets, indicating researchers preferred NSL-KDD as a network dataset. The survey also found that most studies employed SVM and DT machine learning techniques. Several ML methods have been explored. However, none emerged as the top-performing ML classifier for NID. Critical survey papers about NID research from this time include [12], [13], [14], and [15]. The study landscape has transformed since the first network dataset was created in 1999. By 2015, computers could handle increasingly complicated ML classifiers. ML classifiers have advanced, with various novel architectures proposed. This made them a good competitor for numerous complex problems. Conventional datasets, such as DARPA network traces, were questioned due to the increasing complexity and sophistication of network settings and services and the increasing number of modern attacks. Modern datasets have been in the literature since 2015, allowing researchers to modernise the study of ML-based NID research with network traffic and attacks that better reflect how modern networks and services are set up and work [54], [59]. The adoption of modern datasets at first could have been faster because modern datasets needed time to build up a good reputation. UNSW-NB15 [54] was a prominent early modern dataset. It was first published in 2015 and has been utilised in several research publications.

Despite the availability of current datasets in 2016, most ML-based NID research still employed NSL-KDD. However, it was evident that KDD99 usage had plummeted. Interestingly, most of the research papers in ML classifiers incorporated the C4.5 DT classifier (or J48, which is the Java implementation of C4.5). It was unsurprising because C4.5 has a solid reputation in the data mining community. Wu et al. [87] recognised C4.5 as one of the top 10 most influential classifiers in the data mining community in 2006. In 2017, a new modern dataset called CICID2017 was released, becoming one of the most widely used modern datasets. In 2018, research such [58], [88], [89], [90], [91], and [92] used modern datasets for ML-based NID. Nevertheless, some researchers still used conventional datasets. Some of such studies are [93], [94], [95], and [96]. Furthermore, CICID2017 was a top choice for researchers with more appearances in the publications. Researchers also began exploring other software tools for their work, most notably Python used by authors in [58] and [95] and R used by authors in [91] and [92].

The applicability of machine learning classifiers has increased since 2019. Particularly more appearance of ensemble learning classifiers was observed [59], [97], [98], [99], [100], [101], [102], [103]. In 2019, the switch from WEKA to Python was evident among researchers. Many ML tools and modules are available for Python, making ML experiments, including NID experiments, versatile and easy. In 2019, researchers used more modern datasets than conventional ones, continuing the trend. The number of ML-based NID research publications in 2020 and beyond has surpassed previous years. The availability of modern datasets and the rapid development of software tools and libraries for ML research contributed to the increased interest. Many published articles evaluated both modern and conventional datasets for comparative purposes. An investigation into datasets utilised in research publications between 2015 and 2022 revealed that modern datasets are growing at a similar rate to conventional datasets. While conventional datasets will remain important for comparison, we expect academics to use modern datasets as this field advances progressively.

The ensemble learning model became popular among all ML families in ML classifiers. Some notable ML-based NID research works published in 2020 and beyond are [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], and [123]. Based on the reported evaluation results, these studies conclude that ensemble learners perform better classification than other machine learning techniques.

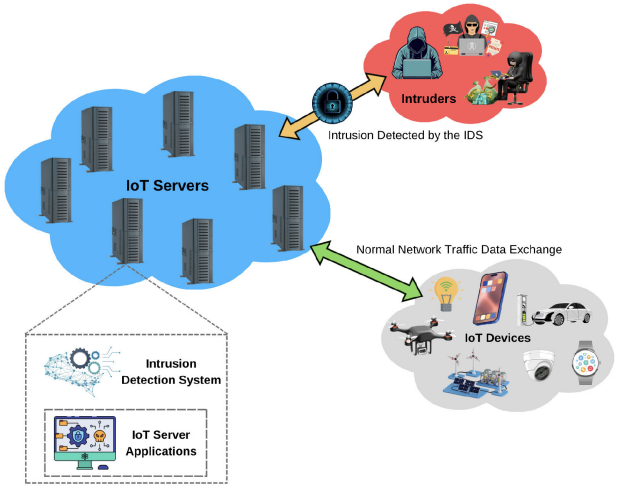

FIGURE 4. IoT-IDS architecture.

VIII. IoT ML-BASED NID RESEARCH

The Internet of Things (IoT) is a comprehensive term encompassing diverse applications that result from the integration of smart devices with the Internet. Such applications

include indispensable devices for smart households and advanced equipment for industrial facilities. Despite the unique objectives of IoT applications, there are commonalities among them. The relevant literature discusses the security architectures and deployments of IoT NIDS from their inception to current advancements. Our study exclusively incorporates research that employs machine learning, as this is the primary area of interest. Fig. 4 presents an illustration of IoT-IDS Architecture.

Although the world's first IoT device was invented in the early 1980s at Carnegie Melon University USA, the first publicly published IoT dataset only appeared about four decades later [124]. In 2017, Meidan et al. [70] created the first IoT dataset, dubbed N-BaIoT, which collects network traffic behaviour snapshots from exploited real IoT devices. Its initial acceptance by researchers was slow initially. However, it has been used in several studies in recent years, signifying wider adoption by researchers. Some notable early studies that utilised this dataset to evaluate the performance of different ML classifiers are [70], [125], and [126]. A year later, in 2018, Koroniotis et al. [73] offered a dataset named Bot-IoT that includes both real and simulated IoT network traffic and numerous attacks. A realistic testbed environment is offered to resolve the current dataset's shortcomings in capturing complete network information, correct labelling, and recent and varied attack variety. Like the previous dataset, adopting the BoT-IoT dataset in studies in its early days of being publicly available was negligible as the previous was more popular at its publication. This has, however, evolved through time. Some of the first few studies that utilised it are [127], [128], [129]. Also, in the same year, Mirsky et al. [74] published another IoT dataset named *Kitsune*, which gained much popularity amongst IoT network security researchers as it possessed different variants of Botnet attacks. This made it suitable for the study of different IoT attacks and the development of ML classifiers to detect these various attacks. Although Pahl and Aubet [71] published another IoT dataset called *DS2OS* in 2018. However, its utility for NID research has been comparatively low compared to the preceding three datasets. This is because it comprises only features that identify anomalies in IoT traffic frequency and microservices' communication baseline models. Its limitation makes it impossible to apply the dataset to IoT research works differently from the data features contained in the DS2OS dataset. However, with the two previously stated IoT datasets available in the research domain, researchers could design and evaluate several machine learning classifiers on the datasets to cover a broader scope. This, in turn, directly contributed to more ML classifier-related IoT-related research publications that further amplified their popularity.

In adding to the pool of publicly available IoT datasets, Kang et al. [76] developed and made available to the public *IoT Network Intrusion Dataset*. However, like the DS2OS dataset, its NID research utilisation has been minimal. We attribute this to two reasons. The main one is that the dataset is in the IEEE repository, making it inaccessible to researchers who are not IEEE members. Secondly, compared to N-BaIoT, BoT-IoT and Kitsune, the dataset size is small as its collection environment comprises few connected devices, leading to low adoption in this research domain. 2020 saw the release of two more IoT datasets. These datasets were made to include new IoT attacks and represent the typical network traces of the IoT environment. The datasets are *MedBIoT* [79] and *IoT-23* [77]. Even though both datasets were created from data collected from real connected IoT devices, using both is still very minimal for NID research. We believe these two datasets will garner traction in this research area in the following years and assist researchers in learning more about emerging IoT attacks and creating more effective and efficient ML classifiers to identify these attacks. More recently, in 2021 and 2022, *WUSTL-IIoT* [75] and *Edge-IIoTset* [83] are datasets that were created based on Industrial Internet of Things network traces of the current IoT environment. Both datasets were created using testbeds to replicate the actual IIoT environment and include the most recent attacks in the environment. These datasets were created to be distinct from the other publicly available IoT datasets to address the NID research area's non-availability of the IIoT datasets. Both datasets are publicly available and can be used for ML-related NID research.

Consequently, with the increase in available IoT network datasets, researchers in IoT network security are constantly exploring new ways of using machine learning to protect the network. The following section summarises past work on ML-based IDS in IoT networks, and Table 10 gives a summary comparison.

Bahşi et al. [126] applied feature selection on the N-BaIoT dataset to reduce the number of features used to detect IoT bots in their work. They begin by comparing the Decision Tree and k-NN performance on a dataset with varied sizes of top features (2, 3, and 10) as determined by Fisher's score. The Decision Tree achieved the highest accuracy with three top features, and only the Decision Tree is used for further assessments in their work. Other follow-up experiments using the IoT dataset included discriminatory power analysis of feature categories to identify the features that contribute the most to classification, device-based modelling by handling data from individual devices separately, and the effect of dataset balancing on the classifier accuracy. A multiclass classifier based on a shallow approach, a decision tree, demonstrated that a multiclass classifier could achieve remarkably high accuracy rates and provide interpretable results with fewer features. Although the IoT dataset used in the study is publicly available, the source code to help researchers recreate the exact experiment is unavailable. This makes it difficult for researchers to recreate the published work if they desire.

Soe et al. [130] proposed a detection system that uses Artificial Neural Network (ANN) to detect DDoS kinds of

attacks in a public IoT dataset. They employed Bot-IoT, a modern dataset for botnet-based attacks, to detect DDoS attacks. However, a critical issue (data imbalance) had to be solved since the dataset contains a small number of benign (normal) data and a significant number of attack data. They solved the imbalanced data problem by implementing the SMOTE (Synthetic Minority Oversampling Technique) machine learning-based DDoS detection system. Their findings suggest that the proposed technique can detect DDoS attacks in an IoT environment. Whilst ANN succeeded well in their work, their results were limited to the DDoS attack category and not to the whole dataset, including other attack categories.

In their publication, Anthi et al. [131] proposed a three-layer intrusion detection system (IDS) that employs a supervised technique to identify a variety of common network-based cyber-attacks on IoT networks. Three primary functions comprise the system: categorise and profile each connected IoT device's usual behaviour, detect malicious packets on the network when an attack occurs, and classify the attack. The system is examined in the context of a smart home testbed comprised of eight widely accessible commercial products. The efficacy of the proposed IDS architecture is evaluated by implementing 12 attacks from four primary network-based attack categories: Denial of Service (DoS), Man-In-The-Middle (MITM)/spoofing, Reconnaissance, and Replay. In addition, the system is evaluated against four scenarios, specifically multistage attacks that include intricate sequences of events. Overall, Weka's implementation of the J48 decision tree classifier with pruning achieved the best results, with F-measures of 99.7%, 97.0%, and 99.0% for each experiment, respectively. Demonstrating the capability of the proposed architecture to distinguish between IoT devices on the network autonomously, ascertain the nature of network activity (malicious or benign), and identify the specific attack that was successfully carried out on each device connected to the network.

An optimised machine learning-based framework for detecting attacks on IoT devices was proposed by Injadat et al. [132]. This framework combines a Bayesian optimisation Gaussian Process (BO-GP) classifier with a Decision tree (DT) classifier. A performance evaluation of the proposed system is conducted using the Bot-IoT dataset. The experimental results indicate that the optimised framework achieves high accuracy, precision, recall, and F-score in detecting botnet attacks in IoT settings. This suggests that the framework is adequate and robust in detecting such attacks.

Recently, the IoT Security research community has been actively creating models that use machine learning classifiers to detect anomalous, intrusion, and cyber-attack data and analyse IoT security. However, the crucial and foremost issue that has not been thoroughly investigated is how to choose an effective machine learning classifier when there are several machine learning classifiers for cyber-attack detection systems for IoT security. In their publication, Shafiq et al. [133] provided a novel framework model and a hybrid method to handle this issue. To begin, the machine learning classifier uses the BoT-IoT dataset and 44 of its features. Then, five separate ML classifiers are chosen to identify illegitimate and abnormal traffic and the most commonly used metrics to evaluate ML classifier performance. A bijective soft set technique and classifier are used to evaluate which machine learning classifier is the most effective and can identify IoT anomalies and intrusion traffic. Then, they applied the proposed classifier using the bijective soft set technique. The empirical results illustrate that the proposed model and classifier utilise an ML classifier (Naive Bayes) from a diverse set of ML classifiers to successfully identify abnormalities and breaches in IoT networks.

The authors Alqahtani et al. [134] introduced a strategy that integrates a Fisher-score-based feature selection technique with a genetic-based extreme gradient boosting (GXGBoost) model to identify the most important features and detect IoT botnet attacks. Fisher's score is a filter-based feature selection method employed to discover important features and exclude irrelevant ones by reducing the gap between intra-classes and increasing the distance between intra-classes. In contrast, GXGBoost is a highly effective and precise approach for categorising IoT botnet attacks. A series of tests were conducted using a dataset of publicly accessible botnet (N-BaIoT) IoT devices. An analysis of holdout and cross-validation results showed that the suggested method achieved a high detection rate using just three out of the 115 data traffic features. This strategy significantly improved the overall performance of the IoT botnet attack detection procedure.

The study by Guerra-Manzanares et al. [79] aimed to tackle the scarcity and restricted scale of publicly accessible datasets for IoT networks. Their research addresses these issues by offering a comprehensive IoT data set (MedBIoT) that encompasses normal and malicious botnet network activity inside a medium-sized IoT network architecture of 83 IoT devices. Three well-recognised botnet malware families, Mirai, BashLite, and Torii, are implemented, and statistics on botnet infection, propagation, and communication with command and control stages are collected. Three (Random Forest, Decision Tree and k-NN) ML classification models are run on the acquired data for binary and multiclass classification to demonstrate the suitability and reliability of the generated dataset for ML-based botnet detection IDS testing, design and deployment. Random Forest outperformed the two other models in both classification scenarios as it can discriminate the labels more accurately with 95.32% and 97.66% accuracy, respectively. This paper's IoT dataset is publicly available. However, the source code to help researchers recreate the exact experiment is not publicly available. This makes it difficult for researchers to recreate the published work if they desire.

Toutsop et al. [135] analysed an attack dataset of real-world IoT devices, such as smart cameras, laptops, and smart-

phones, obtained from the Hacking and Countermeasure Research Lab (HCRL). This is due to previous studies revealing a 67% increase in security breaches over the past five years and a 95% vulnerability of HTTP servers to Man-in-the-middle (MIM) attacks. Their approach to this problem involved presenting a model incorporating Random Forest, Logistic Regression, and Decision Tree methodologies. Data analysis indicates that the overall detection rate ranges from 98 to 100%, surpassing traditional IDS.

In the work conducted by Latif et al. [136], ML methods are employed to identify several types of cybersecurity attacks, such as denial of service (DoS), malicious operation, malicious control, data type probing, surveillance, scan, and incorrect configuration, on the DS2OS IoT dataset. This article introduces a novel prediction model based on a lightweight random neural network (RaNN) for predicting the attacks above. An analysis was conducted to assess the performance of the conventional artificial neural network (ANN), support vector machine (SVM), and decision tree (DT) using widely used evaluation criteria, including accuracy, precision, recall, and F1 score. The results reveal that the RaNN model in this study attains an accuracy of 99.20% when trained at a rate of 0.01 and predictions made within 34.51 milliseconds. The precision, recall, and F1 score achieved comparative values of 99.11%, 99.13%, and 99.20%, respectively. Compared to the state-of-the-art ML classifiers for IoT security, the suggested technique improves the accuracy of attack detection by an average of 5.65%.

Using the Bot-IoT dataset, Leevy et al., [137] developed a prediction model to detect information theft attacks. This work's contribution is characterised by the innovative use of eight classifiers and two performance indicators to identify information theft traffic. Before this study, the Information Theft attack category of the dataset had never been the primary subject of a research investigation. A total of four ensembles (CatBoost, Light-GBM, XGBoost, and Random Forest) and four non-ensembles (Decision Tree, Logistic Regression, Naive Bayes, and a Multilayer Perceptron (MLP)) were evaluated using the dataset. Classifiers are evaluated using the metrics of Area Under the Receiver Operating Characteristic Curve (AUC) and Area Under the Precision-Recall Curve (AUPRC). To obtain the optimal classifier(s), they employ cross-validation to train and evaluate Bot-IoT examples, including regular and information theft traffic. Their analyses indicate that the most successful models are ensemble classifiers, namely CatBoost, LightGBM, and XGBoost. The IoT dataset discussed in this paper is accessible to the public. However, the source code necessary for empirical replication of the precise experiment is not accessible to the public. Therefore, this poses a challenge for researchers who wish to replicate the reported findings.

Rajesh et al. [138] introduced a new model called Hybrid Ensemble Learning Enabled Sigmoid-Cosine Integrated Pigeon Inspired Feature Selection Based Intrusion Detection (HSPFSID) for detecting and mitigating malicious activities in Tactile Internet-driven Consumer Healthcare IoT systems. The study investigated the security issues in healthcare IoT by assessing and contrasting the efficacy of the suggested model on four separate datasets (IoT bot, NSL-KDD, CICIDS2017, and KDD99). The results exhibited its superior performance compared to current methods and its potential to enhance cybersecurity measures for healthcare IoT systems. The study also examined the difficulties and progress in Tactile Internet and its implementations in the healthcare-oriented consumer sector, underscoring the requirement of providing patients with accurate and safe therapy. Furthermore, the study prioritised developing and assessing the proposed HSPFSID model, its feature selection algorithm, and its capacity to effectively identify and reduce risks in healthcare IoT settings.

In their study, Ullah and Mahmoud [139] utilised an ML methodology to identify IoT devices linked to the network by analysing the transmitted and received network traffic. They altered the IoT23 Pcap files to conduct experiments on a smart home network and generate traffic measurements. Furthermore, they devised a systematic approach for detecting IoT devices by analysing network data. Feature extraction from the complete, reduced, and flow-based datasets is used to evaluate the proposed model. This work uses flow-based features to identify IoT devices linked to the network. The proposed methodology attains a perfect accuracy, precision, recall, and F1 score of 100% for all dataset features. The assessment conducted using their dataset demonstrates that their proposed model can accurately categorise IoT devices.

Abbasi et al. [140], highlight that a critical challenge in the IoT is identifying and preventing unauthorised access by attackers to the network and devices. Furthermore, they said that traditional IDS lack responsiveness or, at the absolute least, are ineffective when employed in the IoT. This study focuses on the deployment of ML classifiers for anomaly detection. The authors provide two methods for extracting features and performing classification. Feature extraction and classification in the first approach are performed using Logistic Regression (LR), whereas the second approach utilises an Artificial Neural Network (ANN). The performance of the suggested technique is evaluated using the N-BaIoT dataset, which comprises data samples from nine IoT devices and incorporates several attacks. In contrast to the four deep learning (DL) methods, logistic regression demonstrates superior efficiency and attains the highest classification accuracy of 99.98%.

In their study, Gandhi and Li [141] highlighted the enormous threat posed by botnets to IoT devices. They attributed this threat to these devices' inadequate default security settings and the limited security awareness among end-users. The researchers also observed the default number of open ports and the unchanged default user credentials. Many detecting systems have been devised to counteract the increasing prevalence of botnet attacks. Most of these studies

concentrated on a particular method or botnet dataset, leading to a need for comprehensive comparisons of several ML and DL methods to this problem employing a variety of IoT datasets. To tackle this problem, the researchers assessed the effectiveness of five ML and two DL classifiers by using four recently released datasets of IoT botnets extracted from actual and virtual IoT devices infiltrated by the Mirai malware. The experimental results indicate that Random Forest achieves the highest detection accuracy and the shortest required testing time for all four datasets.

The study by Alsalman, [142] sought to tackle the cybersecurity issues on Internet of Medical Things (IoMT) healthcare networks by developing sophisticated IDS utilising ML methods. Furthermore, it aims to investigate the incorporation of blockchain technology into IoMT healthcare to guarantee data validity, regulate access, and enable auditing. The objective of the work was to create a system that can detect deviations from anticipated conduct in IoMT devices and users. This system would quickly identify such anomalous actions to improve network security. In addition, the study concentrated on creating dynamic real-time reaction and mitigation techniques for cybersecurity problems in IoMT healthcare networks. This involved automating responses, alerts, and threat containment to enhance security. The study's primary objective was to provide original contributions to healthcare cybersecurity by tackling the unique obstacles encountered and improving the rapid identification and responses to threats.

Chunduri et al. [143] emphasised that botnets provide a significant and widespread threat to cyber-physical devices globally, exhibiting rapid and varied evolution with extensive scalability. Furthermore, they emphasised that one of the variations focuses on the IoT ecosystem, which includes a wide range of devices such as sensors, actuators, and other intelligent devices. Furthermore, it is noted that contemporary botnet threats possess several capabilities instead of solely focusing on sending DDoS attacks to devices. Their study included two IoT botnet datasets, IoT-23 and MedBIoT, which consist of contemporary attacks to carry out multiclass classification. The researchers analysed six different types of IoT botnet assaults from both datasets and classified them into three categories. Furthermore, to assess the resilience of the four machine learning classifiers employed in the research, test samples were created using conditional generative adversarial networks (CTGAN). The Random Forest model achieved the highest accuracy score. This study is restricted to analysing four different types of IoT botnets and focuses on only twelve specific features.

Kumar et al. [144] introduced a new distributed IDS that utilises fog computing to identify DDoS events targeting a mining pool in a blockchain-enabled IoT network. The performance of the proposed IDS is evaluated on distributed fog nodes using a Random Forest (RF) classifier and an efficient gradient tree boosting technique (XGBoost). The effectiveness of the suggested method is assessed using BoT-IoT, which incorporates the latest cybersecurity risks identified in blockchain-enabled IoT networks. Empirical results indicate that XGBoost exhibited superior performance in detecting binary attacks, while RF showed superior performance in detecting multiclass attacks. Training and testing distributed fog nodes using Random Forest (RF) typically require less time than XGBoost.

In their study, Mohy et al. [145] developed and verified an IDS model for IIoT security. The proposed approach combines Isolation Forest (IF) for identifying anomalies, Pearson's Correlation Coefficient (PCC) for selecting features, and Random Forest (RF) for the binary classification of normal packets and intrusions. The study seeks to empirically establish the efficacy of the suggested model in mitigating the security risks linked to IIoT and to highlight its superior performance compared to prior similar notions. Furthermore, the work aims to enhance IDS for IIoT security using feature engineering and machine learning methodologies.

The study by Maz et al. [146] aimed to improve the identification of keylogging attacks in IoT networks by creating a Majority Rating Ensemble Classifier. This methodology integrates the forecasts of three deep learning models (CNN, RNN, and LSTM) to enhance the precision and resilience of intrusion detection programs. This study examined the difficulties of feature selection, false positive reduction, and basic heuristics' constraints in detecting keylogging attacks. Furthermore, its objective was to enhance cybersecurity by offering an innovative method for identifying keylogging attacks on IoT devices and improving the precision of threat detection. Moreover, the study aimed to showcase the efficacy of ensemble classifiers in dealing with the changing dynamics of IoT security and to offer valuable knowledge and approaches for future research and real-world applications in cybersecurity.

The research by Saheed et al. [147] introduced a new method for identifying intrusion threats in IoT networks. This method utilises ensemble learning models optimised by a Grey Wolf Optimiser (GWO). The aim is to tackle the increasing security vulnerabilities linked to the rising prominence of IoT applications, namely, IDS. The work aims to create a voting GWO ensemble model that integrates a traffic analyser and a classification phase engine to enhance accuracy and detection rate while reducing false alarms in identifying IoT cyberattacks. Furthermore, the paper highlights the need to employ authentic datasets that accurately reflect actual attack situations in the IoT to assess the effectiveness of the described model.

The authors Zhu and Liu [148] introduced an innovative approach to intrusion detection in IoT networks. This method utilises subspace clustering and ensemble learning methodology to attain exceptional performance and resilience. This work aims to tackle the difficulties posed by varied, intricate, and ever-changing data in IoT networks and efficiently and precisely identify both prevalent and novel attacks. The described approach combines subspace

clustering with ensemble learning, utilising three synergy techniques: Clustering Results as Features (CRF), Two-Level Decision Making (TDM), and Iterative Feedback Loop (IFL). The work proposes comprehensive experiments on a publicly available dataset to assess, verify, and compare the suggested approach with current methods. The results indicate that the suggested approach can efficiently identify prevalent and novel attacks in IoT networks, achieving high accuracy and a low percentage of false positives.

By providing an enhanced IDS, Jayalatchumy et al. [149] tackled the issues of network intrusion detection in the IoT framework. The main goals of this study are to solve data imbalance by using data-denoising techniques, to optimise feature selection using an enhanced Crow Search Algorithm (CSA), and to implement an ensemble classifier for multiclass classification in two denoised datasets, namely the NSL-KDD and UNSW-NB15 datasets. Furthermore, the study evaluated the effectiveness of the suggested method by employing several performance measures such as accuracy, F1-score, false positive rate, recall, and precision rate. The primary objective of this work was to augment the efficiency of network intrusion detection in IoT networks through the development and validation of an enhanced IDS system.

The study conducted by Inuwa and Das [150] used several ML techniques to identify cyber abnormalities in IoT systems and evaluate the effectiveness of these techniques. It includes a comparative investigation of ML approaches, including SVM, ANN, DT, LR, and kNN, to provide insights into their contributions to categorising cyber-attacks in IoT systems. The main goal of their article is to provide invaluable knowledge to cybersecurity professionals, directing the creation of strong protection measures for the IoT ecosystem. Their results have the potential to make a substantial contribution to the development of cybersecurity practices and strengthen IoT settings against possible threats.

The study conducted by Talukder et al. [151] introduced a novel intrusion detection method for Wireless Sensor Networks (WSNs) that combines ML approaches with the Synthetic Minority Oversampling Technique Tomek Link (SMOTE-Tomek) algorithm. The objective of the work was to tackle the difficulties encountered by WSNs, including low detection rates, computational burden, and false alarms, by creating a resilient IDS that can acquire knowledge about intricate patterns and abnormalities from WSN data. Furthermore, the research aims to offer proactive and adaptable security protocols for WSNs by precisely detecting and reducing unauthorised access in real time, thus protecting the integrity and confidentiality of transmitted data. The study also sought to boost intrusion detection in WSNs by providing a new ML-based method that tackles the issues of imbalanced datasets and improves the security of WSNs.

Maghrabi's work, [152], introduced an automated NIDS designed for IoT settings, specifically emphasising improving security controls. It utilised the Random Forest classifier and suggested a filter-based method incorporating Pearson's correlation coefficient (PCC) and data balancing strategies to tackle imbalances in class distributions in the UNSW-NB15 dataset. The study's objective was to increase the rates of intrusion detection, decrease the occurrence of false alarms, and boost the overall accuracy of intrusion detection in IoT networks. Moreover, the study assessed the performance of the suggested model by comparing it to the state-of-the-art method while proposing future research directions in IoT deployments.

Chander et al. [153] introduced an Enhanced Pelican Optimisation Algorithm with an Ensemble Voting-based Anomaly Detection (EPOA-EVAD) method for ensuring security in IIoT. This study aimed to mitigate the critical requirement for anomaly detection in IIoT settings by incorporating several sophisticated methods like ensemble-based voting, handling of class imbalance data, and hyperparameter optimisation using the seagull optimisation (SGO) algorithm. The work aimed to improve the precision and effectiveness of anomaly detection in dynamic IIoT environments. This would ultimately contribute to the progress of anomaly detection (AD) methods that are essential for preserving the integrity and efficiency of industrial processes in the era of IIoT.

In general, datasets related to the Internet of Things pose distinct challenges and prospects in contrast to datasets not associated with IoT. This is evident based on the comparison between both datasets in Table 7.

Generally, IoT classifiers should be capable of dealing with the massive amount, velocity, and diversity of data created by IoT devices, along with the requirement for real-time computation and particular preparation procedures. Safety and confidentiality are also vital issues when dealing with IoT data. There are several reasons why some machine learning methods may work well on some network intrusion detection datasets and not on others. Here are a few possible reasons, as shown in Table 8. In addition, we present in Table 9 some factors that influence the functionality of ML classifiers on ID datasets.

In Table 10, we gave a brief overview of some notable research works that utilised IoT datasets in their studies. We believe that researchers looking to embark on IoT-related studies can scan through the table to see some of the datasets, classifiers, classification types and experiment results that have been reported and published. This serves as a guide towards re-validating the experiments or approaching the research from a new dimension.

IX. LESSONS LEARNED

This survey covers the datasets used in NID research over the past two decades. This study analysed traditional datasets from over a decade ago and contemporary datasets published within the previous decade and used in recent research efforts. The investigation prioritises IoT datasets. Our study found that while many datasets are publicly available, the variety of IoT datasets is limited. The delayed attention to IoT NID research can be ascribed to the fact that the initial public IoT dataset was not made available until late 2017.

TABLE 7. Comparison between Non-IoT and IoT datasets for NID.

Aspect	Non-IoT Dataset	IoT Dataset
Traffic behaviour	Humans primarily generate data through web browsing, emailing, and social media activity.	Sensors, machines, and IoT devices generate data, including wearables, smart home appliances, and industrial equipment. This data is typically generated at a high frequency and in large volumes.
Used features	Often include user identification, location, device type, and browser information.	Includes sensor data such as temperature, humidity, pressure, and motion and device-specific information such as battery level, firmware version, and signal strength.
Data types	Typically contain structured data in tables, with each row representing a unique data point.	Datasets can contain structured and unstructured data, including media data, sensor readings, and wearable data, just to name a few.
Data volume	Datasets are typically moderate, with tens or hundreds of thousands of data points.	Datasets can range from tens of thousands to billions of data points, depending on the number of devices and sensors involved.
Data velocity	Datasets are generated at relatively low velocity, with new data points typically added over days or weeks.	Datasets are generated at a high velocity, with new data points added continuously or at a high frequency.
Data Variety	Datasets are typically limited in variety, with data points representing a relatively small number of actions or interactions.	Datasets can be highly varied, with data points representing a wide range of sensor readings, device statuses, and other parameters.
Data quality	The datasets are typically high quality, with relatively few missing or incorrect data points.	The datasets can suffer from quality issues due to sensor malfunction, signal interference, or other factors. Data cleaning and preprocessing are often necessary to address these issues.

We also discussed non-intrusive load monitoring software and machine learning methods. Our analysis found that academics prefer individual classifiers over ensemble techniques for NIDS development despite the availability of more advanced machine learning methods. Researchers have switched from WEKA to Python for NID studies, and the broad range of Python libraries has influenced this change.

Additionally, this study examines some initial NID machine learning research. A comprehensive review of IoT-related research employing advanced machine learning methods and modern datasets follows. The rationale for the recent emphasis on the IoT is its significant proliferation, which has led to an even greater expansion of its network coverage. Despite the comparatively small number of IoT research studies compared to conventional datasets, our literature analysis suggests a promising future. Since the release of the IoT dataset, statistical data from research outputs have supported this result. NID research in the IoT has great promise, given its current pace. However, this potential can only be maximised by adding datasets to current ones, enhancing IoT research.

X. FUTURE DIRECTIONS

The DARPA network data, a pivotal dataset since 1998, has significantly advanced ML-based NID research. Its influence has recently extended to IoT, which focuses on developing classifiers that can accurately identify attacks in modern networks. The creation of IoT-related ML classifiers and datasets has surged in recent years. However, the current state-of-the-art is still far from the desired outcome, with several barriers to overcome. Addressing the following issues is crucial to propel this field's research forward.

The IoT has witnessed incremental growth across various industries, including military, healthcare, industrial control, logistics, and smart environments. These sectors rely on the network to transmit sensitive data, necessitating advanced security solutions. A single rogue node in a network can lead to data leaks and various issues. The limited resources of IoT devices further complicate the development of advanced detection classifiers. The recent use of ML to create effective security solutions that leverage prior experiences and handle IoT device features is a promising development. However, the research on ML-based NID, which relies on ML classifiers, has been fragmented. Many ML classifiers have been attempted and done separately, with their hyper-parameter reduction and choice of features being mostly ad hoc. This makes it challenging to compare performance outcomes across research studies and agree on the optimum ML strategy for NID after years of research.

The convergence of ML techniques towards ensemble learning classifiers (ELC) in IoT research papers is evident in Table 10, as they outperform single ML learners in network attack classification. Consistent approaches to compare results and explain superiority should be used with other new ML techniques to further the research. Researchers conducting IoT-related ML research can use Table 10 for an overview and guide featuring state-of-the-art studies. We summarised various IoT-related studies in Section VII to better comprehend the research.

ML classifier performance depends on dataset quality; thus, it is an important design aspect. ML-based NID research should leverage real-world networks to fully capture network and user activities. However, some publicly available IoT datasets in the literature are real-world datasets. The simulated ones are developed using model-based network traffic simulation. This raises questions about how effectively datasets represent real-world networks and attack actions. Network technology, services, and attack techniques evolve quickly in the diverse IoT ecosystem, making dataset preservation difficult. This opens opportunities for research on modelling user and network behaviours to develop simulation platforms and productive tools that can quickly produce more representative network datasets. The publicly available datasets must be updated periodically with new IoT network attacks to train ML classifiers on new attack variants and

TABLE 8. Comparison between Non-IoT and IoT ML classifiers for NID.

Aspect	Non-IoT ML Classifiers	IoT ML Classifiers
Deployment	These datasets are typically processed with ML classifiers on centralised servers or cloud platforms.	Classifiers may be processed on edge devices such as sensors, gateways, and smart cameras. This requires lightweight classifiers to handle these devices' limited processing power and storage capacity.
Interpretability	These ML classifiers can often be interpreted and explained using feature importance analysis and model visualisation techniques.	The ML classifiers may be more complex and difficult to interpret due to many features and the need for specialised preprocessing techniques. Explainability is important for safety-critical applications such as autonomous vehicles and medical devices.
Resource consumption	ML classifiers can consume significant amounts of processing power, memory, and storage depending on the dataset's size and the model's complexity.	ML classifiers must be optimised for resource consumption due to the limited resources available on edge devices. This requires specialised techniques like classifiers' hyperparameter optimisation and dataset feature dimensionality reduction.

TABLE 9. Factors that influence the functionality of ML classifiers on ID datasets.

Factors	Discussion
Data Features	The performance of machine learning classifiers can be influenced by various dataset features, including but not limited to their size, quality, diversity, and class balance. In cases where the dataset is limited in size or exhibits imbalanced class distribution, certain models may encounter difficulty achieving optimal generalisation and accuracy.
Model Complexity	The capacity of a classifier to learn insights from the data can be influenced by the intricacy of its architecture and hyperparameters. If the classifier is too simple, it may not capture the complexity of the underlying patterns. In contrast, excessive complexity may lead to overfitting the training data and limited generalizability to new data.
Feature Engineering	The performance of classifiers can be significantly influenced by the quality and relevance of the features employed to represent the data. If the constituents lack informative value or need to encompass pertinent data about the problem, the classifiers may encounter difficulties learning meaningful patterns from the dataset.
Classifier Suitability	Classifiers exhibit varying strengths and limitations, with some being more appropriate for specific data types or problem domains than others. Selecting a suitable classifier that aligns with the problem and data features is paramount.

evaluate their detection efficacy and efficiency. Additionally, we have included a summary table and addressed public IoT datasets in Table 6 and Section VI to help readers comprehend their categories and attack methods.

An IDS developed using one method is insufficient for today's heterogeneous and advanced IoT network. Advanced mechanisms using multiple strategies may be the best option. Modern research suggests that ELCs better detect intrusions than single machine learning classifiers. Explainable artificial intelligence (XAI) may help develop effective and efficient machine learning classifiers in addition to preprocessing and feature selection methods. Combining XAI with ML classifiers allows people to understand classification results, unlike the "black box" approach of using only classifiers, where even classifier developers cannot explain how an ML classifier classified an event. This will enable researchers to study how IoT properties help classify network records in ELCs. During preprocessing, redundant features will be discovered and eliminated, lowering the number of training features and improving detection and optimisation.

The validity of ML-based NID for IoT research outputs with baseline datasets is also essential. Previous studies [98], [102], [114], [156] validated ML classifier performance using traditional datasets. We recommend that IoT researchers do the same. That is, using a benchmark IoT dataset to evaluate any proposed ML classifier. This will ensure that research ML classifiers are cross-validated using multiple datasets to verify classification efficiency and efficacy.

In some circumstances, most ML-related NID studies, including IoT, used publicly available datasets, ML classifiers, and feature dimensionality reduction approaches. According to our observations, researchers do not publish their experiment source code for other researchers to evaluate and replicate their work for validation. This may lead to unmatched outcomes when repeating the experiment because the researchers lack the same processes as the first publication authors. Providing the source codes can help other researchers understand the preprocessing processes needed to attain the same results as the published articles.

XI. SUMMARY

This article introduces the IoT and its heterogeneity, enormous scale, and dynamic behaviour, which present new NID challenges. The paper discusses typical IDS constraints and the benefits of using ML to overcome them. IDS ML classifiers, including supervised learning, are also discussed.

TABLE 10. Summary of some notable IoT ML-based NID research.

Ref	Pub. Year	Dataset	Attack Types	Experiment Results	Pros	Cons	Classifier
Soe *et al.* [130]	2019	Bot-IoT	DDoS	0% for FPR, and 100% for TPR, Precision, Recall & F-Measure	High detection accuracy with 0 FPR.	The study utilized the DDoS attack category of the dataset.	ANN
Anthi *et al.* [131]	2019	Private IoT dataset	DoS, DDoS, MITM, Spoofing, Insecure Firmware & Data Leakage	F-measure of 99.7%, 97.0%, and 99.0% and test times of 0.1, 0.4, and 0.2s for 3 experiments.	High F-measure for all the experiment variations.	The experiment dataset is not publicly available for further re-validation.	J48
Injadat *et al.* [132]	2020	Bot-IoT	Botnets	99.99%, 99%, 100%, 100% for Accuracy, Precision, Recall & F-score	High classification performance based on metric scores.	Only binary classification is considered in the study.	BO-GP & DT
Alqahtani *et al.* [134]	2020	N-BaIoT	Bashlite, & Mirai	99.96% Accuracy for binary and multiclass classification	Same detection accuracy for both classification categories.	The study did not confirm if XGBoost parameter values reach the global optimal.	Fisher Score + GXGBoost Model
Guerra-Manzanares *et al.* [79]	2020	MedBIoT	Mirai, BashLite & Torii	95.32% & 97.66% Accuracy for binary and multiclass classification.	High detection accuracy and low complexity	Lack of parameter tuning details.	RF
Toutsop *et al.* [135]	2020	IoT Network Intrusion Dataset	Man-in-the-middle (MIM)	100% Accuracy for DT & RF	High detection accuracy for different dataset size variations.	Only a single attack category of the dataset is considered in the study.	DT & RF
Latif *et al.* [136]	2021	DS2OS	DoS, Malicious Control, Data Type Probing, Scan, Malicious Operation, Wrong Setup & Spying	99.20% accuracy at a learning rate of 0.01	High detection accuracy with fast test time.	The dataset only possesses anomalies in IoT traffic frequency and the microservice communication baseline model.	RaNN
Leevy *et al.* [137]	2021	Bot-IoT	Information Theft	98.785% for AUC & 99.463% for AUPRC	High detection performance	Only a single attack category of the dataset is considered in the study.	LightGBM
Ullah and Mahmoud [139]	2021	IoT-23 extract	Device classification	100% Accuracy	High classification of IoT devices.	No attack is considered in the study.	DT
Abbasi *et al.* [140]	2021	N-BaIoT	Botnets	99.98% Accuracy	Lightweight classifier with High detection performance	Only a single attack category of the dataset is considered in the study.	LR
Chunduri *et al.* [143]	2021	IoT-23 & MedBIoT	Botnets variants	99.88% Accuracy	High detection accuracy for both datasets.	Only two attack categories were considered in the study.	RF
Kumar *et al.* [144]	2022	BoT-IoT	DoS, DDoS, Reconnaissance & Information Theft	99.99% for binary and multiclass classification.	High detection performance	The subcategory of the attacks was not considered in the study.	RF & XGBoost
Mohy *et al.* [154]	2023	BoT-IoT	Binary	99.99% Accuracy	High detection performance	Only one classifier was evaluated in the paper.	kNN + PCA
Thakkar *et al.* [155]	2023	NSL-KDD, UNSW-NB15, CIC-IDS-2017, & BoT-IoT	Multiple Attacks	Multiple Results	High detection performance	Only one classifier was evaluated in the paper.	DNN
Rajesh *et al.* [138]	2024	IoT bot, NSL-KDD, CICIDS2017, & KDD99	Multiple Attacks	Multiple Results	High detection performance	Only two classifiers and multiclass classification were evaluated in the paper.	HSPFSID
Alsalman, [142]	2024	WUSTL EHMS 2020 & ICU-IOMT	Binary Attacks	Multiple Results	High detection performance	Only binary classification was evaluated in the paper.	FusionNet
Maz *et al.* [146]	2024	BoT-IoT	Multiple Attacks	97.67% Accuracy	High detection performance	The paper did not report their experiments' training and test time.	Ensemble Classifier Majority Voting
Saheed *et al.* [147]	2024	BoT-IoT & UNSW-NB15	Multiclass Attacks	100% Accuracy	High detection performance	The paper did not report their experiments' training and test time.	Voting GWO
Zhu and Liu [148]	2024	UNSW-NB15	Multiclass Attacks	97.05% Accuracy	High detection performance	The dataset used in the paper is not IoT related.	Ensemble Learning
Jayalatchumy *et al.* [149]	2024	NSL-KDD & UNSW-NB15	Multiclass Attacks	Multiple Results	High detection performance	The datasets used in the paper are not IoT related.	ICSA-FS Ensemble Model
Inuwa *et al.* [150]	2024	ToN-IoT & BoT-IoT	Multiclass Attacks	Multiple Results	High detection performance	No ensemble classifier was considered in the paper	NN
Talukder *et al.* [151]	2024	WSN-DS	Binary & Multiclass Attacks	Multiple Results	High detection performance	The paper did not report their experiments' training and test time.	RF
Maghrabi, [152]	2024	UNSW-NB15	Multiclass Attacks	90.17% Accuracy	High detection performance	The dataset used in the paper is not IoT related.	RF
Chander *et al.* [153]	2024	UNSW-NB15 & UCI SEMCOM	Multiclass Attacks	Multiple Results	High detection performance	The dataset used in the paper is not IoT related.	EPOA-EVAD

These classifiers are tested across multiple IoT datasets, such as the Na-BaIoT and Bot-IoT datasets. We then examined the challenges of using ML to detect IoT intrusions. These include deployment, IoT device processing and storage limitations, and the need to evaluate and explain results.

The survey article discusses the next phases and future research on ML-built NID for the IoT. We suggest more research to develop more accurate and effective attack detection systems to manage dynamic and diverse IoT networks. Overall, the article examines network intrusion detection from an IoT and non-IoT perspective.

XII. CONCLUSION

The present survey highlights the persisting challenge of intrusion detection in the Internet of Things (IoT) environment, which has been exacerbated by the evolution of the Internet, with a shift in emphasis from connectivity to data. The IoT has generated significant anticipation owing to its capacity to transform tangible objects from diverse application domains into Internet-enabled hosts. Notwithstanding, malevolent actors may exploit the vast potential of the IoT to compromise the confidentiality and integrity of personal data. Consequently, developing security solutions for the IoT is imperative. The IDS ensures IoT security like conventional networks. The present study is primarily concerned with research that employs machine learning classifiers to detect intrusions, safeguard data and ensure its security. The evolution of network datasets was expounded upon, commencing with the DARPA dataset that depicted the network during the latter part of the 1990s. Deficiencies were identified in KDD99 and its updated iteration, NSL-KDD. Some novel datasets have surfaced in

response to the demand for datasets that accurately reflect modern network configurations and services. Our research primarily centred on datasets related to the IoT and studies utilising machine learning techniques to assess the efficacy of various classifiers in detecting intrusions within the IoT ecosystem.

Ultimately, we shared our perspectives regarding the prospective trajectory of research on IoT NID. The necessity of establishing a consensus regarding the optimal approach for IoT NID was observed. Facilitating a consistent comparison of performance outcomes across various research endeavours would be advantageous. Moreover, classifiers developed based on datasets generated by simulated networks may not comprehensively reflect the actual network behaviours in the real world. It is imperative to gather additional datasets about IoT networks from authentic networks to accurately capture the behaviours of said networks and verify the findings of research endeavours. We have proposed a future investigation of IoT datasets on vehicular networks, attestation, and drones. It is recommended that additional datasets about IoT networks be gathered from authentic networks to accurately capture the behaviours of IoT networks and authenticate the findings of research endeavours.

ACKNOWLEDGMENT

This work was partly sponsored by Horizon 2020 Marie Skłodowska-Curie Actions under the project SwiftV2X (grant agreement ID 101008085). We would also like to acknowledge the support of the 5GIC/6GIC members.

REFERENCES

[1] M. Chen and A. A. Ghorbani, "A survey on user profiling model for anomaly detection in cyberspace," *J. Cyber Secur. Mobility*, vol. 8, no. 1, pp. 75–112, 2019.

[2] C. Gilmore and J. Haydaman, "Anomaly detection and machine learning methods for network intrusion detection: An industrially focused literature review," in *Proc. Int. Conf. Secur. Manage. (SAM)*, 2016, p. 292.

[3] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of Things: Vision, applications and research challenges," *Ad Hoc Netw.*, vol. 10, no. 7, pp. 1497–1516, Sep. 2012.

[4] J. Bradley, J. Barbier, and D. Handler, "Embracing the Internet of Everything to capture your share of $14.4 trillion," Cisco Internet Bus. Solutions Group (IBSG), Cisco Syst., San Jose, CA, USA, White Paper, 2013.

[5] D. Singh, G. Tripathi, and A. J. Jara, "A survey of Internet-of-Things: Future vision, architecture, challenges and services," in *Proc. IEEE World Forum Internet Things (WF-IoT)*, Mar. 2014, pp. 287–292.

[6] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," *Bus. Horizons*, vol. 58, no. 4, pp. 431–440, Jul. 2015.

[7] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead," *Comput. Netw.*, vol. 76, pp. 146–164, Jan. 2015.

[8] (2021). *Number of Internet of Things (IoT) Connected Devices Worldwide From 2019 to 2030 (in Billions)*. Accessed: Feb. 2, 2022. [Online]. Available: https://www.statista.com/

[9] S. Axelsson, "Intrusion detection systems: A survey and taxonomy," Tech. Rep., 2000.

[10] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *Exp. Syst. Appl.*, vol. 36, no. 10, pp. 11994–12000, 2009.

[11] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung, "Intrusion detection system: A comprehensive review," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, 2013.

[12] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Proc. Comput. Sci.*, vol. 60, pp. 708–713, Jan. 2015.

[13] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016.

[14] A. Özgür and H. Erdem, "A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015," *PeerJ Preprints*, vol. 4, Apr. 2010, Art. no. e1954v1.

[15] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *J. Netw. Comput. Appl.*, vol. 60, pp. 19–31, Jan. 2016.

[16] 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.

[17] B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network intrusion detection," *IEEE Netw.*, vol. 8, no. 3, pp. 26–41, May 1994.

[18] D. Larson, "Distributed denial of service attacks—Holding back the flood," *Netw. Secur.*, vol. 2016, no. 3, pp. 5–7, Mar. 2016.

[19] S. Venkatraman and M. Alazab, "Use of data visualisation for zero-day malware detection," *Secur. Commun. Netw.*, vol. 2018, pp. 1–13, Dec. 2018.

[20] H. N. Saha, A. Mandal, and A. Sinha, "Recent trends in the Internet of Things," in *Proc. IEEE 7th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2017, pp. 1–4.

[21] V. Subrahmanyam and K. Aruna, "Future automobile an introduction of IoT," *Int. J. Trend Res. Develop.*, vol. 2, pp. 88–90, Apr. 2017.

[22] H. Ahmadi, G. Arji, L. Shahmoradi, R. Safdari, M. Nilashi, and M. Alizadeh, "The application of Internet of Things in healthcare: A systematic literature review and classification," *Universal Access Inf. Soc.*, vol. 18, no. 4, pp. 837–869, Nov. 2019.

[23] N. Đurević, A. Labus, Z. Bogdanović, and M. Despotović-Zrakić, "Internet of Things in marketing and retail," *Int. J. Adv. Comput. Sci. Appl*, vol. 6, no. 3, pp. 20–24, 2017.

[24] B. L. Risteska Stojkoska and K. V. Trivodaliev, "A review of Internet of Things for smart home: Challenges and solutions," *J. Cleaner Prod.*, vol. 140, pp. 1454–1464, Jan. 2017.

[25] M. Bacco, L. Boero, P. Cassara, M. Colucci, A. Gotta, M. Marchese, and F. Patrone, "IoT applications and services in space information networks," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 31–37, Apr. 2019.

[26] *2020 Unit 42 IoT Threat Report*, Palo Alto Networks, Santa Clara, CA, USA, 2020.

[27] R. Heady, G. Luger, A. Maccabe, and M. Servilla, "The architecture of a network level intrusion detection system," Los Alamos National Lab, New Mexico Univ., Albuquerque, Albuquerque, NM, USA, Tech. Rep., 1990.

[28] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 686–728, 1st Quart., 2019.

[29] T. Sommestad, H. Holm, and D. Steinvall, "Variables influencing the effectiveness of signature-based network intrusion detection systems," *Inf. Secur. J., A Global Perspective*, vol. 31, no. 6, pp. 711–728, Nov. 2022.

[30] R. Werlinger, K. Hawkey, K. Muldner, P. Jaferian, and K. Beznosov, "The challenges of using an intrusion detection system: Is it worth the effort?" in *Proc. 4th Symp. Usable Privacy Secur.*, vol. 24, Jul. 2008, pp. 107–118.

[31] J. R. Goodall, W. G. Lutters, and A. Komlodi, "Developing expertise for network intrusion detection," *Inf. Technol. People*, vol. 22, no. 2, pp. 92–108, Jun. 2009.

[32] T. Sommestad and A. Hunstad, "Intrusion detection and the role of the system administrator," *Inf. Manage. Comput. Secur.*, vol. 21, no. 1, pp. 30–40, Mar. 2013.

[33] M. Zaman and C.-H. Lung, "Evaluation of machine learning techniques for network intrusion detection," in *Proc. NOMS IEEE/IFIP Netw. Oper. Manage. Symp.*, Apr. 2018, pp. 1–5.

[34] H. Holm, "Signature based intrusion detection for zero-day attacks: (Not) a closed chapter?" in *Proc. 47th Hawaii Int. Conf. Syst. Sci.*, Jan. 2014, pp. 4895–4904.

[35] T. F. Lunt, "Automated audit trail analysis and intrusion detection: A survey," in *Proc. 11th Nat. Comput. Secur. Conf.*, vol. 353, 1988, pp. 65–73.

[36] G. Fernandes, J. J. P. C. Rodrigues, L. F. Carvalho, J. F. Al-Muhtadi, and M. L. Proença, "A comprehensive survey on network anomaly detection," *Telecommun. Syst.*, vol. 70, no. 3, pp. 447–489, Mar. 2019.

[37] A. Tabassum, A. Erbad, and M. Guizani, "A survey on recent approaches in intrusion detection system in IoTs," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2019, pp. 1190–1197.

[38] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network intrusion detection for IoT security based on learning techniques," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2671–2701, 3rd Quart., 2019.

[39] J. C. S. Sicato, S. K. Singh, S. Rathore, and J. H. Park, "A comprehensive analyses of intrusion detection system for IoT environment," *J. Inf. Process. Syst.*, vol. 16, no. 4, pp. 975–990, 2020.

[40] R. P. V. and R. Sandhu, "POSTER: Security enhanced administrative role based access control models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1802–1804.

[41] P. V. Rajkumar, S. K. Ghosh, and P. Dasgupta, "An end to end correctness verification approach for application specific usage control," in *Proc. Int. Conf. Ind. Inf. Syst. (ICIIS)*, vol. 33, Dec. 2009, pp. 1–6.

[42] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 538–566, 1st Quart., 2023.

[43] A. Alotaibi and M. A. Rassam, "Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense," *Future Internet*, vol. 15, no. 2, p. 62, Jan. 2023.

[44] S. V. N. S. Kumar, M. Selvi, and A. Kannan, "A comprehensive survey on machine learning-based intrusion detection systems for secure communication in Internet of Things," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, pp. 330–348, Jan. 2023.

[45] J. Song, H. Takakura, Y. Okabe, M. Eto, D. Inoue, and K. Nakao, "Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation," in *Proc. 1st Workshop Building Anal. Datasets Gathering Exper. Returns Secur.*, Apr. 2011, pp. 29–36.

[46] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, and A. Baak, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, Mar. 2016, Art. no. 160018.

[47] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and M. A. Zissman, "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation," in *Proc. DARPA Inf. Survivability Conf. Expo. (DISCEX)*, vol. 2, Sep. 1998, pp. 12–26.

[48] W. Lee and S. J. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM Trans. Inf. Syst. Secur.*, vol. 3, no. 4, pp. 227–261, Nov. 2000.

[49] J. Song, H. Takakura, and Y. Okabe. (2006). *Description of Kyoto University Benchmark Data*. Accessed: Mar. 15, 2016. [Online]. Available: http://www.takakura.com/Kyoto_data/BenchmarkData-Description-v5.pdf

[50] 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.

[51] 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.

[52] (2012). *Capture Files From MID-Atlantic CCDC*. Accessed: Apr. 17, 2023. [Online]. Available: https://www.netresec.com/?page=MACCDC

[53] 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.

[54] 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. 1–6.

[55] W. Haider, J. Hu, J. Slay, B. P. Turnbull, and Y. Xie, "Generating realistic intrusion detection system dataset based on fuzzy qualitative modeling," *J. Netw. Comput. Appl.*, vol. 87, pp. 185–192, Jun. 2017.

[56] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and VPN traffic using time-related features," in *Proc. 2nd Int. Conf. Inf. Syst. Secur. Privacy*, 2016, pp. 407–414.

[57] E. K. Viegas, A. O. Santin, and L. S. Oliveira, "Toward a reliable anomaly-based intrusion detection in real-world environments," *Comput. Netw.*, vol. 127, pp. 200–216, Nov. 2017.

[58] 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.

[59] 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.

[60] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, pp. 51–62, Dec. 2017.

[61] R. Choudhary and H. K. Gianey, "Comprehensive review on supervised machine learning algorithms," in *Proc. Int. Conf. Mach. Learn. Data Sci. (MLDS)*, Dec. 2017, pp. 37–43.

[62] M. Mohri, "Foundations of machine learning," Tech. Rep., 2018.

[63] M. Bicego and M. Loog, "Weighted K-nearest neighbor revisited," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1642–1647.

[64] J. Zhao, J. Ouenniche, and J. De Smedt, "Survey, classification and critical analysis of the literature on corporate bankruptcy and financial distress prediction," *Mach. Learn. With Appl.*, vol. 15, Mar. 2024, Art. no. 100527.

[65] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[66] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, "Introduction to machine learning, neural networks, and deep learning," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, p. 14, 2020.

[67] R. Polikar, "Ensemble learning," in *Ensemble Machine Learning: Methods and Applications*. Boston, MA, USA: Springer, 2012, pp. 1–34.

[68] S. S. More and P. P. Gaikwad, "Trust-based voting method for efficient malware detection," *Proc. Comput. Sci.*, vol. 79, pp. 657–667, Jan. 2016.

[69] L. Wen and M. Hughes, "Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques," *Remote Sens.*, vol. 12, no. 10, p. 1683, May 2020.

[70] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, and Y. Elovici, "N-BaIoT—Network-based detection of IoT botnet attacks using deep autoencoders," *IEEE Pervasive Comput.*, vol. 17, no. 3, pp. 12–22, Jul. 2018.

[71] M.-O. Pahl and F.-X. Aubet, "All eyes on you: Distributed multi-dimensional IoT microservice anomaly detection," in *Proc. 14th Int. Conf. Netw. Service Manage. (CNSM)*, Nov. 2018, pp. 72–80.

[72] M. Teixeira, T. Salman, M. Zolanvari, R. Jain, N. Meskin, and M. Samaka, "SCADA system testbed for cybersecurity research using machine learning approach," *Future Internet*, vol. 10, no. 8, p. 76, Aug. 2018.

[73] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Gener. Comput. Syst.*, vol. 100, pp. 779–796, Nov. 2019.

[74] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, "Kitsune: An ensemble of autoencoders for online network intrusion detection," 2018, *arXiv:1802.09089*.

[75] M. Zolanvari, M. A. Teixeira, L. Gupta, K. M. Khan, and R. Jain, "Machine learning-based network vulnerability analysis of industrial Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6822–6834, Aug. 2019.

[76] H. Kang, D. H. Ahn, G. M. Lee, J. Yoo, K. H. Park, and H. K. Kim, "IoT network intrusion dataset," *IEEE Dataport*, Jan. 2019.

[77] A. Parmisano, S. Garcia, and M. Erquiaga, *A Labeled Dataset With Malicious and Benign IoT Network Traffic*. Praha, Czech Republic: Stratosphere Laboratory, 2020.

[78] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165130–165150, 2020.

[79] A. Guerra-Manzanares, J. Medina-Galindo, H. Bahsi, and S. Nõmm, "MedBIoT: Generation of an IoT botnet dataset in a medium-sized IoT network," in *Proc. 6th Int. Conf. Inf. Syst. Secur. Privacy*, 2020, pp. 207–218.

[80] H. Hindy, C. Tachtatzis, R. Atkinson, E. Bayne, and X. Bellekens, "MQTT-IoT-IDS2020: MQTT Internet of Things intrusion detection dataset," *IEEE Dataport*, Jun. 2020.

[81] M. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, "X-IIoTID: A connectivity-agnostic and device-agnostic intrusion data set for industrial Internet of Things," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3962–3977, Mar. 2022.

[82] S. Dadkhah, H. Mahdikhani, P. K. Danso, A. Zohourian, K. A. Truong, and A. A. Ghorbani, "Towards the development of a realistic multidimensional IoT profiling dataset," in *Proc. 19th Annu. Int. Conf. Privacy, Secur. Trust (PST)*, Aug. 2022, pp. 1–11.

[83] 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.

[84] 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)*, vol. 10, Aug. 2023, pp. 1–9.

[85] E. C. P. Neto, H. Taslimasa, S. Dadkhah, S. Iqbal, P. Xiong, T. Rahman, and A. A. Ghorbani, "CICIoV2024: Advancing realistic IDS approaches against DoS and spoofing attack in IoV CAN bus," *Internet Things*, vol. 26, Jul. 2024, Art. no. 101209.

[86] M. Sabhnani and G. Serpen, "Why machine learning algorithms fail in misuse detection on KDD intrusion detection data set," *Intell. Data Anal.*, vol. 8, no. 4, pp. 403–415, Oct. 2004.

[87] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and S. Y. Philip, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.

[88] D. Aksu, S. Üstebay, M. A. Aydin, and T. Atmaca, "Intrusion detection with comparative analysis of supervised learning techniques and Fisher score feature selection algorithm," in *Proc. Int. Symp. Comput. Inf. Sci.*, 2018, pp. 141–149.

[89] M. B. Nawir, A. Amir, N. Yaakob, and B. Ong, "Multi-classification of UNSW-NB15 dataset for network anomaly detection system," *J. Theor. Appl. Inf. Technol.*, vol. 96, pp. 5094–5104, Aug. 2018.

[90] M. Nawir, A. Amir, O. B. Lynn, N. Yaakob, and R. Badlishah Ahmad, "Performances of machine learning algorithms for binary classification of network anomaly detection system," *J. Phys., Conf. Ser.*, vol. 1018, May 2018, Art. no. 012015.

[91] A. Bansal and S. Kaur, "Extreme gradient boosting based tuning for classification in intrusion detection systems," in *Proc. Int. Conf. Adv. Comput. Data Sci.*, Oct. 2018, pp. 372–380.

[92] A. A. Abdulrahman and M. K. Ibrahem, "Evaluation of DDoS attacks detection in a new intrusion dataset based on classification algorithms," *Iraqi J. Inf. Commun. Technol.*, vol. 1, no. 3, pp. 49–55, Feb. 2019.

[93] C. J. Ugochukwu and E. Bennett, "An intrusion detection system using machine learning algorithm," *Int. J. Comput. Sci. Math. Theory*, vol. 4, no. 1, pp. 39–47, 2018.

[94] S. M. Othman, F. M. Ba-Alwi, N. T. Alsohybe, and A. Y. Al-Hashida, "Intrusion detection model using machine learning algorithm on big data environment," *J. Big Data*, vol. 5, no. 1, p. 34, Dec. 2018.

[95] K. Park, Y. Song, and Y.-G. Cheong, "Classification of attack types for intrusion detection systems using a machine learning algorithm," in *Proc. IEEE 4th Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar. 2018, pp. 282–286.

[96] S. K. Biswas, "Intrusion detection using machine learning: A comparison study," *Int. J. Pure Appl. Math.*, vol. 118, no. 19, pp. 101–114, 2018.

[97] S. Hosseini and H. Seilani, "Anomaly process detection using negative selection algorithm and classification techniques," *Evolving Syst.*, vol. 12, no. 3, pp. 769–778, Sep. 2021.

[98] R. Abdulhammed, M. Faezipour, H. Musafer, and A. Abuzneid, "Efficient network intrusion detection using PCA-based dimensionality reduction of features," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Jun. 2019, pp. 1–6.

[99] I. Obeidat, N. Hamadneh, M. Alkasassbeh, M. Almseidin, and M. AlZubi, "Intensive pre-processing of kDD Cup 99 for network intrusion classification using machine learning techniques," *iJIM*, vol. 13, no. 1, p. 71, 2019.

[100] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "A detailed analysis of the CICIDS2017 data set," in *Information Systems Security and Privacy*, P. Mori, S. Furnell, and O. Camp, Eds., Cham, Switzerland: Springer, 2019, pp. 172–188. [Online]. Available: https://www.google.co.uk/search?q="Network+intrusion+detection+using+supervised+machine+learning+technique+with+feature+selection

[101] H. Malhotra and P. Sharma, "Intrusion detection using machine learning and feature selection," *Int. J. Comput. Netw. Inf. Secur.*, vol. 11, no. 4, p. 43, 2019.

[102] S. Sapre, P. Ahmadi, and K. Islam, "A robust comparison of the KDDCup99 and NSL-KDD IoT network intrusion detection datasets through various machine learning algorithms," 2019, *arXiv:1912.13204*.

[103] G. Madhukar and G. N. Kumar, "An intruder detection system based on feature selection using random forest algorithm," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 2, pp. 5525–5529, Dec. 2019.

[104] P. S. Saini, S. Behal, and S. Bhatia, "Detection of DDoS attacks using machine learning algorithms," in *Proc. 7th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2020, pp. 16–21.

[105] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset," *IEEE Access*, vol. 8, pp. 32150–32162, 2020.

[106] A. A. Sallam, M. N. Kabir, Y. M. Alginahi, A. Jamal, and T. K. Esmeel, "IDS for improving DDoS attack recognition based on attack profiles and network traffic features," in *Proc. 16th IEEE Int. Colloq. Signal Process. Appl. (CSPA)*, Feb. 2020, pp. 255–260.

[107] Z. Liu, N. Thapa, A. Shaver, K. Roy, X. Yuan, and S. Khorsandroo, "Anomaly detection on IoT network intrusion using machine learning," in *Proc. Int. Conf. Artif. Intell., Big Data, Comput. Data Commun. Syst. (icABCD)*, Aug. 2020, pp. 1–5.

[108] S. Thaseen, I, B. Poorva, and P. S. Ushasree, "Network intrusion detection using machine learning techniques," in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng. (ic-ETITE)*, Feb. 2020, pp. 1–7.

[109] V. Kumar, V. Choudhary, V. Sahrawat, and V. Kumar, "Detecting intrusions and attacks in the network traffic using anomaly based techniques," in *Proc. 5th Int. Conf. Commun. Electron. Syst. (ICCES)*, Jun. 2020, pp. 554–560.

[110] D. Stiawan, M. Y. B. Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 dataset feature analysis with information gain for anomaly detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020.

[111] S. Farhat, M. Abdelkader, A. Meddeb-Makhlouf, and F. Zarai, "Comparative study of classification algorithms for cloud IDS using NSL-KDD dataset in WEKA," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2020, pp. 445–450.

[112] M. Hooda, J. Babu, P. S. Vamsi, and G. Gopakumar, "An improved intrusion detection system based on KDD dataset using feature ranking and data sampling," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Jul. 2020, pp. 1128–1132.

[113] G. Sah and S. Banerjee, "Feature reduction and classifications techniques for intrusion detection system," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Jul. 2020, pp. 1543–1547.

[114] D. Rani and N. C. Kaushal, "Supervised machine learning based network intrusion detection system for Internet of Things," in *Proc. 11th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2020, pp. 1–7.

[115] M. A. Umar, C. Zhanfang, and Y. Liu, "Network intrusion detection using wrapper-based decision tree for feature selection," in *Proc. Int. Conf. Internet Comput. Sci. Eng.*, Jan. 2020, pp. 5–13.

[116] R. Magán-Carrión, D. Urda, I. Díaz-Cano, and B. Dorronsoro, "Towards a reliable comparison and evaluation of network intrusion detection systems based on machine learning approaches," *Appl. Sci.*, vol. 10, no. 5, p. 1775, Mar. 2020.

[117] X. Shi, Y. Cai, and Y. Yang, "Extreme trees network intrusion detection framework based on ensemble learning," in *Proc. IEEE Int. Conf. Adv. Electr. Eng. Comput. Applications (AEECA)*, Aug. 2020, pp. 91–95.

[118] Q. R. S. Fitni and K. Ramli, "Implementation of ensemble learning and feature selection for performance improvements in anomaly-based intrusion detection systems," in *Proc. IEEE Int. Conf. Ind. 4.0, Artif. Intell., Commun. Technol. (IAICT)*, Jul. 2020, pp. 118–124.

[119] I. Abrar, Z. Ayub, F. Masoodi, and A. M. Bamhdi, "A machine learning approach for intrusion detection system on NSL-KDD dataset," in *Proc. Int. Conf. Smart Electron. Commun. (ICOSEC)*, Sep. 2020, pp. 919–924.

[120] H. Rajadurai and U. D. Gandhi, "A stacked ensemble learning model for intrusion detection in wireless network," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 15387–15395, Sep. 2022.

[121] A. A. Hady, A. Ghubaish, T. Salman, D. Unal, and R. Jain, "Intrusion detection system for healthcare systems using medical and network data: A comparison study," *IEEE Access*, vol. 8, pp. 106576–106584, 2020.

[122] S. Rajagopal, P. P. Kundapur, and K. S. Hareesha, "A stacking ensemble for network intrusion detection using heterogeneous datasets," *Secur. Commun. Netw.*, vol. 2020, pp. 1–9, Jan. 2020.

[123] X. Larriva-Novo, C. Sánchez-Zas, V. A. Villagrá, M. Vega-Barbas, and D. Rivera, "An approach for the application of a dynamic multi-class classifier for network intrusion detection systems," *Electronics*, vol. 9, no. 11, p. 1759, Oct. 2020.
[124] (2020). *History of the Internet of Things (IoT)*. Accessed: May 6, 2022. [Online]. Available: https://www.itonlinelearning.com/blog-history-iot/
[125] S. S. Chawathe, "Monitoring IoT networks for botnet activity," in *Proc. IEEE 17th Int. Symp. Netw. Comput. Appl. (NCA)*, Nov. 2018, pp. 1–8.
[126] H. Bahsi, S. Nõmm, and F. B. La Torre, "Dimensionality reduction for machine learning based IoT botnet detection," in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2018, pp. 1857–1862.
[127] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, and A. Alazab, "A novel ensemble of hybrid intrusion detection system for detecting Internet of Things attacks," *Electronics*, vol. 8, no. 11, p. 1210, Oct. 2019.
[128] A. Guerra-Manzanares, H. Bahsi, and S. Nõmm, "Hybrid feature selection models for machine learning based botnet detection in IoT networks," in *Proc. Int. Conf. Cyberworlds (CW)*, Oct. 2019, pp. 324–327.
[129] Y. N. Soe, Y. Feng, P. I. Santosa, R. Hartanto, and K. Sakurai, "Rule generation for signature based detection systems of cyber attacks in IoT environments," *Bull. Netw., Comput., Syst., Softw.*, vol. 8, no. 2, pp. 93–97, 2019.
[130] Y. N. Soe, P. I. Santosa, and R. Hartanto, "DDoS attack detection based on simple ANN with SMOTE for IoT environment," in *Proc. 4th Int. Conf. Informat. Comput. (ICIC)*, Oct. 2019, pp. 1–5.
[131] E. Anthi, L. Williams, M. Slowinska, G. Theodorakopoulos, and P. Burnap, "A supervised intrusion detection system for smart home IoT devices," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 9042–9053, Oct. 2019.
[132] M. Injadat, A. Moubayed, and A. Shami, "Detecting botnet attacks in IoT environments: An optimized machine learning approach," in *Proc. 32nd Int. Conf. Microelectron. (ICM)*, Dec. 2020, pp. 1–4.
[133] M. Shafiq, Z. Tian, Y. Sun, X. Du, and M. Guizani, "Selection of effective machine learning algorithm and bot-IoT attacks traffic identification for Internet of Things in smart city," *Future Gener. Comput. Syst.*, vol. 107, pp. 433–442, Jun. 2020.
[134] M. Alqahtani, H. Mathkour, and M. M. Ben Ismail, "IoT botnet attack detection based on optimized extreme gradient boosting and feature selection," *Sensors*, vol. 20, no. 21, p. 6336, Nov. 2020.
[135] O. Toutsop, P. Harvey, and K. Kornegay, "Monitoring and detection time optimization of man in the middle attacks using machine learning," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Oct. 2020, pp. 1–7.
[136] S. Latif, Z. Zou, Z. Idrees, and J. Ahmad, "A novel attack detection scheme for the industrial Internet of Things using a lightweight random neural network," *IEEE Access*, vol. 8, pp. 89337–89350, 2020.
[137] J. L. Leevy, J. Hancock, T. M. Khoshgoftaar, and J. Peterson, "Detecting information theft attacks in the bot-IoT dataset," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 807–812.
[138] R. Rajesh, S. Hemalatha, S. M. Nagarajan, G. G. Devarajan, M. Omar, and A. K. Bashir, "Threat detection and mitigation for tactile Internet driven consumer IoT-healthcare system," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 4249–4257, Feb. 2024.
[139] I. Ullah and Q. H. Mahmoud, "Network traffic flow based machine learning technique for IoT device identification," in *Proc. IEEE Int. Syst. Conf. (SysCon)*, Apr. 2021, pp. 1–8.
[140] F. Abbasi, M. Naderan, and S. E. Alavi, "Anomaly detection in Internet of Things using feature selection and classification based on logistic regression and artificial neural network on N-BaIoT dataset," in *Proc. 5th Int. Conf. Internet Things Appl. (IoT)*, May 2021, pp. 1–7.
[141] R. Gandhi and Y. Li, "Comparing machine learning and deep learning for IoT botnet detection," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Aug. 2021, pp. 234–239.
[142] D. Alsalman, "A comparative study of anomaly detection techniques for IoT security using adaptive machine learning for IoT threats," *IEEE Access*, vol. 12, pp. 14719–14730, 2024.
[143] H. Chunduri, T. Gireesh Kumar, and P. S. Charan, "A multi class classification for detection of IoT botnet malware," in *Proc. Int. Conf. Comput. Sci., Commun. Secur.* Cham, Switzerland: Springer, 2021, pp. 17–29.
[144] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, S. Garg, and M. M. Hassan, "A distributed intrusion detection system to detect DDoS attacks in blockchain-enabled IoT network," *J. Parallel Distrib. Comput.*, vol. 164, pp. 55–68, Jun. 2022.
[145] M. Mohy-Eddine, A. Guezzaz, S. Benkirane, M. Azrour, and Y. Farhaoui, "An ensemble learning based intrusion detection model for industrial IoT security," *Big Data Mining Anal.*, vol. 6, no. 3, pp. 273–287, Sep. 2023.
[146] Y. A. Maz, M. Anbar, S. Manickam, S. D. A. Rihan, B. A. Alabsi, and O. M. Dorgham, "Majority voting ensemble classifier for detecting keylogging attack on Internet of Things," *IEEE Access*, vol. 12, pp. 19860–19871, 2024.
[147] Y. K. Saheed and S. Misra, "A voting gray wolf optimizer-based ensemble learning models for intrusion detection in the Internet of Things," *Int. J. Inf. Secur.*, vol. 23, no. 3, pp. 1557–1581, Jun. 2024.
[148] J. Zhu and X. Liu, "An integrated intrusion detection framework based on subspace clustering and ensemble learning," *Comput. Electr. Eng.*, vol. 115, Apr. 2024, Art. no. 109113.
[149] D. Jayalatchumy, R. Ramalingam, A. Balakrishnan, M. Safran, and S. Alfarhood, "Improved crow search-based feature selection and ensemble learning for IoT intrusion detection," *IEEE Access*, vol. 12, pp. 33218–33235, 2024.
[150] M. M. Inuwa and R. Das, "A comparative analysis of various machine learning methods for anomaly detection in cyber attacks on IoT networks," *Internet Things*, vol. 26, Jul. 2024, Art. no. 101162.
[151] M. A. Talukder, S. Sharmin, M. A. Uddin, M. M. Islam, and S. Aryal, "MLSTL-WSN: Machine learning-based intrusion detection using SMOTETomek in WSNs," *Int. J. Inf. Secur.*, vol. 23, no. 3, pp. 2139–2158, Jun. 2024.
[152] L. A. Maghrabi, "Automated network intrusion detection for Internet of Things: Security enhancements," *IEEE Access*, vol. 12, pp. 30839–30851, 2024.
[153] N. Chander and M. U. Kumar, "Enhanced pelican optimization algorithm with ensemble-based anomaly detection in industrial Internet of Things environment," *Cluster Comput.*, vol. 27, no. 5, pp. 6491–6509, Aug. 2024.
[154] M. Mohy-Eddine, A. Guezzaz, S. Benkirane, and M. Azrour, "An efficient network intrusion detection model for IoT security using K-NN classifier and feature selection," *Multimedia Tools Appl.*, vol. 82, no. 15, pp. 23615–23633, Jun. 2023.
[155] A. Thakkar and R. Lohiya, "Attack classification of imbalanced intrusion data for IoT network using ensemble learning-based deep neural network," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11888–11895, Jul. 2023.
[156] I. Syarif, R. F. Afandi, and F. A. Saputra, "Feature selection algorithm for intrusion detection using cuckoo search algorithm," in *Proc. Int. Electron. Symp. (IES)*, Sep. 2020, pp. 430–435.

SULYMAN AGE ABDULKAREEM (Student Member, IEEE) received the B.Sc. degree from the Computer Science Department, University of Ilorin, Nigeria, in 2014, and the M.Sc. degree in management information systems from the Faculty of Computing and Engineering, Coventry University, U.K., in 2017. He is currently pursuing the Ph.D. degree with the 5G/6G Innovation Centre, School of Computer Science and Electronic Engineering, University of Surrey. His main research interests include machine learning for network security, information systems, IT strategy, and project management.

CHUAN HENG FOH (Senior Member, IEEE) received the M.Sc. degree from Monash University, Melbourne, VIC, Australia, in 1999, and the Ph.D. degree from the University of Melbourne, Melbourne, in 2002. After the Ph.D. degree, he spent six months as a Lecturer with Monash University. In December 2002, he joined Nanyang Technological University, Singapore, as an Assistant Professor, until 2012. He is currently a Senior Lecturer with the University of Surrey, Guildford, U.K. He has authored or co-authored more than 180 refereed papers in international journals and conferences. His research interests include protocol design, machine learning application, and performance analysis of various computer networks, including wireless local area networks, mobile ad-hoc and sensor networks, vehicular networks, the Internet of Things, 5G/6G networks, and open RAN. He served as the Vice Chair (Europe/Africa) for IEEE TCGCC, in 2015 and 2017. He is currently the Vice-Chair of the IEEE VTS Ad Hoc Committee on Mission Critical Communications. He is on the editorial boards of several international journals.

MOHAMMAD SHOJAFAR (Senior Member, IEEE) received the Ph.D. degree in ICT from the Sapienza University of Rome, Rome, Italy, in 2016. He is currently a Senior Lecturer (Associate Professor) in network security and an Intel Innovator, a Professional ACM Member and ACM Distinguished Speaker, a fellow of the Higher Education Academy, and a Marie Curie Alumni, working with the 5G & 6G Innovation Centre (5G/6GIC), Institute for Communication Systems, University of Surrey, U.K. Before joining 5G/6GIC, he was a Senior Researcher and a Marie Curie Fellow of the SPRITZ Security and Privacy Research Group, University of Padua, Italy. He received the ''Excellent'' Award for the Ph.D. degree. He secured around $1.7M as PI in various EU/U.K. projects, including D-XPERT (funded by I-U.K./U.K., 2024), 5G Mode (funded by DSIT/U.K., 2023), 5G ONE4HDD (funded by DSIT/U.K., 2023), TRACE-V2X (funded by EU/MSCA-SE, 2023), AUTOTRUST (funded by ESA/EU, 2021), PRISENODE (funded by EU/MSCA-IF: 2019), and SDN-Sec (funded by Italian Government: 2018). He was also the COI of various U.K./EU Projects like HiPER-RAN (funded by DSIT/U.K., 2023), APTd5G Project (funded by EPSRC/UKI-FNI: 2022), ESKMARALD (funded by U.K./NCSC, 2022), GAUChO, S2C, and SAMMClouds (funded by Italian Government, 2016 and 2018). He is an Associate Editor of IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, and *Computer Networks*.

FRANÇOIS CARREZ (Senior Member, IEEE) received the Ph.D. degree in theoretical computer science from the University of Nancy, France, in 1991. After 18 years of research with Alcatel Research (now Bell Laboratories) in Telecommunication, Multi-Agent Systems and Security, he joined the University of Surrey, in 2007. Then, he has been, in particular, leading the ICT-FP7 Internet of Things Initiative (IOT-I) Coordination Action, which is the origin of the IoT International Forum. He has also been the WP Leader on Architecture for IoT-A, the flagship European project on architecture for the Internet of Things, which eventually released the comprehensive Architecture Reference Model for the IoT (IoT ARM). More recently, he has been involved in H2020 projects COSMOS, FIESTA, and CPaaS.io, working in particular on IoT-A-inspired system architecture and from 2021 to 2023, in DEDICAT 6G, where he was leading the system architecture activities. His research interests include the IoT, 5G/6G architecture, machine learning/AI and system architecture, GPU programming, and medical science (cardiology, physiology, and neurology).

KLAUS MOESSNER (Senior Member, IEEE) is currently a Professor in communications engineering with the University of Technology Chemnitz and a Professor in cognitive networks with the Institute for Communication Systems and the 5G Innovation Centre, University of Surrey. He was involved in many cognitive communications, service provision, and IoT projects. He was responsible for the work on cognitive decision-making mechanisms in the CR Project ORACLE. He led the work on radio awareness in the ICT FP7 Project QoSMOS and led the H2020 Speed5G Project. His research interests include cognitive networks, the IoT deployments, sensor data-based knowledge generation, reconfiguration, and resource management. He leads the EU-Taiwan Project Clear5G, investigating the extensions of 5G systems needed to serve the particular requirements of future factories. He was the Founding Chair of the IEEE DYSPAN Working Group (WG6) on Sensing Interfaces for Future and Cognitive Communication Systems.