Review

Optimizing Intrusion Detection for IoT: A Systematic Review of Machine Learning & Deep Learning Approaches with Feature Selection & Data Balancing

S Kumar Reddy Mallidi * and Rajeswara Rao Ramisetty

Posted Date: 11 March 2025

doi: 10.20944/preprints202503.0706.v1

Keywords: Internet of Things (IoT); Systematic Review; Feature Selection; Machine Learning; Deep Learning; Data Balancing; SMOTE

 doi:10.20944/preprints202503.0706.v1

Review

Optimizing Intrusion Detection for IoT: A Systematic Review of Machine Learning & Deep Learning Approaches with Feature Selection & Data Balancing

S Kumar Reddy Mallidi [1,2,*] and Rajeswara Rao [3]

[1] Computer Science and Engineering, Jawaharlal Nehru Technological University, Kakinada, 533003, AP, India
[2] Computer Science and Engineering, Sri Vasavi Engineering College, Tadepalligudem, 534101, AP, India
[3] Computer Science and Engineering, Jawaharlal Nehru Technological University Gurajada, Vizianagaram, 535003, AP, India
* Correspondence: satya.cnis@gmail.com

Abstract: As the Internet of Things (IoT) continues expanding its footprint across various sectors, robust security systems to mitigate associated risks are more critical than ever. Intrusion Detection Systems (IDS) are fundamental in safeguarding IoT infrastructures against malicious activities. This systematic review aims to guide future research by addressing six pivotal research questions that underscore the development of advanced IDS tailored for IoT environments. Specifically, the review concentrates on applying Machine Learning (ML) and Deep Learning (DL) technologies to enhance IDS capabilities. It explores various feature selection methodologies aimed at developing lightweight IDS solutions that are both effective and efficient for IoT scenarios. Additionally, the review assesses different datasets and balancing techniques, which are crucial for training IDS models to perform accurately and reliably. Through a comprehensive analysis of existing literature, this review highlights significant trends, identifies current research gaps, and suggests future studies to optimize IDS frameworks for the ever-evolving IoT landscape.

Keywords: IoT security; intrusion detection systems; machine learning; deep learning; feature selection; dataset balancing

1. Introduction

The Internet of Things (IoT) refers to a network of physical objects or "things" embedded with sensors, software, and other technologies to connect and exchange data with other devices and systems over the Internet. These "things" can range from ordinary household objects to sophisticated industrial tools. IoT aims to enable these objects to collect and share data, monitor and control physical environments, and perform specific actions through actuators, ultimately improving efficiency, productivity, and decision-making across various domains.

There are 15.9 billion connected IoT devices worldwide as of 2023, accounting for approximately 75% of all internet-connected devices. As illustrated in Figure 1, this number is expected to continue rising significantly in the coming years [1,2].

The scope of the IoT spans across various sectors, with significant contributions from device manufacturing (30%), consumer IoT (24%), healthcare (18%), transportation (12%), and utilities (9%) [3]. According to a market analysis report by Markets and Markets [4], IoT is anticipated to generate revenue of over 650 billion USD by 2026.

 doi:10.20944/preprints202503.0706.v1

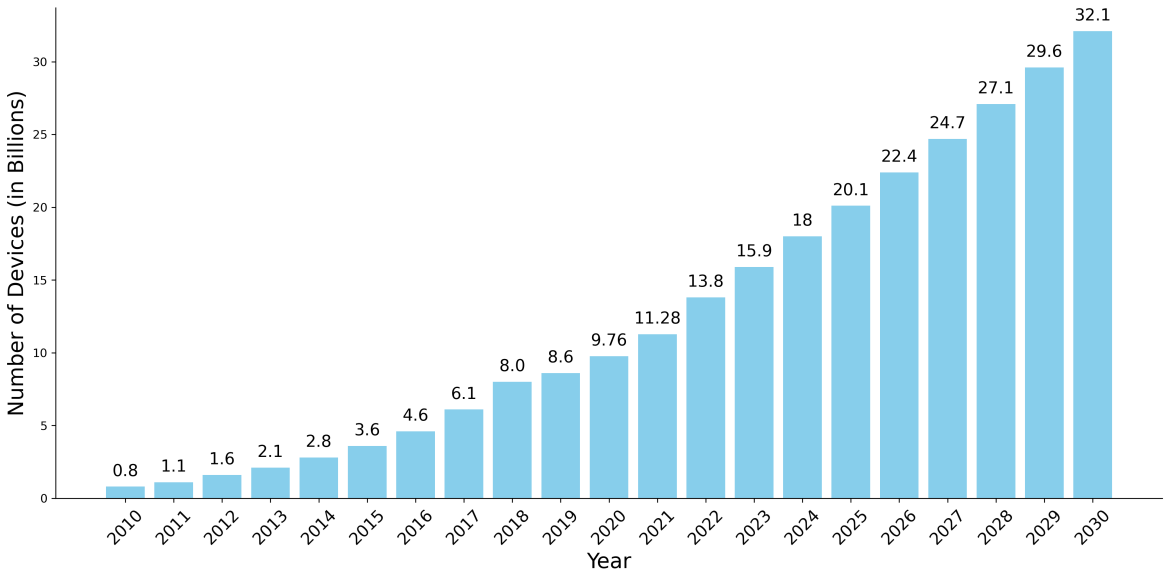

Figure 1. Number of Connected IoT Devices Worldwide (2010-2030)

The rapid expansion of the IoT presents numerous challenges that impact its potential. A significant challenge is the need for global cooperation and standardization. Diverse and fragmented IoT ecosystems, evident in the different approaches taken by various regions, obstruct the development of a unified global framework. This lack of standardization undermines interoperability, seamless integration, and the security of IoT devices and systems across the world [5].

Additionally, the vast number of devices expected to connect within IoT networks poses substantial scalability and energy efficiency challenges. Managing these devices and their immense data demands robust architectures and protocols. Batteries power many IoT devices, and the impracticality of replacing batteries in large-scale deployments raises concerns about energy sustainability. The absence of unified architecture and standard protocols complicates interoperability and exacerbates security and integration issues [6,7]. Moreover, the impact of IoT on privacy and ethics, including issues like surveillance and data-driven life, calls for the development of comprehensive ethical frameworks and protective technologies to safeguard user rights and data integrity [5,7].

1.1. Threats and Security Measures in IoT

Security and privacy represent some of the most significant and persistent challenges in the IoT ecosystem. As IoT devices expand, the landscape of potential cyber threats broadens, exposing these interconnected systems to various sophisticated cyber-attacks. The number of malware attacks targeting IoT networks rose dramatically, from 32 million incidents in 2018 to 112 million in 2022. During this period, the finance sector experienced a Y-o-Y growth of 252%, retail saw an increase of 159%, and education faced a growth of 146% in cyber incidents, highlighting a pressing need for robust security measures [8–10].

The Zscaler 2023 Enterprise IoT and OT Threat Report [11] indicated a staggering 400% increase in attacks in 2023, with manufacturing and education sectors being the prime targets. Geographically, the United States and Mexico were the most affected. The report underscores the predominance of botnets and Distributed Denial of Service (DDoS) attacks, which are becoming increasingly complex and challenging to mitigate.

Major industry reports have focused on emerging security dimensions in response to the escalating threat landscape. For instance, the Globalstar satellite IoT initiative aims to enhance efficiency, reduce costs, and introduce complex security implications due to the expanded attack surface that such enhanced connectivity entails [12]. Furthermore, the integration of Artificial Intelligence (AI) in IoT, as

 doi:10.20944/preprints202503.0706.v1

highlighted by Telenor, offers advanced operational intelligence and automation capabilities. However, these advancements also introduce additional vulnerabilities, especially in data privacy and network security [13].

In 2024, IoT Analytics [14] emphasized ongoing concerns about the inadequacy of current security measures to address evolving threats effectively. This was echoed by the Palo Alto Networks 2023 Benchmark Report on IoT Security, which called for more robust, integrated security solutions that can scale with the increasing number of connected devices [15]. A recent Bitdefender study highlighted that specific malware threats have adapted to exploit IoT vulnerabilities, signaling an urgent need for continuous updates to security protocols and practices [16].

To counter these threats, the IoT security strategy has evolved to focus on both cryptographic measures and detection-based approaches. Cryptographic techniques aim to secure data transmission across IoT devices with lightweight, efficient encryption methods suitable for the constrained nature of many IoT devices [17,18]. On the detection front, IDS has increasingly incorporated ML and DL to enhance their effectiveness, providing advanced capabilities to identify and mitigate potential threats dynamically [19,20].

Despite these efforts, the continuous advancement in IoT technology and the corresponding evolution of cyber threats pose ongoing challenges. It is imperative for research and industry practices to keep pace with these developments, ensuring that security solutions not only address current vulnerabilities but are also adaptable to future threats. This proactive approach to IoT security is essential to safeguard the vast networks of interconnected devices that are increasingly integral to personal and professional environments.

1.2. Limitations and Challenges Associated with IoT Security and Attack Taxonomy

Understanding the limitations and challenges associated with IoT systems is crucial for providing adequate security. Traditional network security measures are not directly suitable for the IoT environment due to its distinct characteristics and the complex nature of its network interactions [21–23]. The principal challenges include:

- *Hardware Limitations:* IoT devices often have constrained computational power, memory, and energy resources, limiting complex security measures' implementation. These inherent limitations make it challenging to deploy robust cryptographic and intrusion detection systems that are resource-intensive.
- *Deployment and Scalability:* The large-scale deployment of IoT devices complicates the management of security measures. A significant challenge is ensuring consistent security protocols across many devices without centralized control.
- *Connectivity:* IoT devices typically connect through heterogeneous networks that can be insecure or unreliable. This variability in connectivity exposes IoT systems to increased risks of network-based attacks.
- *Heterogeneity:* The diversity of IoT devices in terms of operating systems, hardware capabilities, and communication protocols complicates the implementation of uniform security strategies, thus hampering interoperability and seamless integration of security solutions.
- *Embedded and Outdated Software:* Many IoT devices run on embedded software that may not be regularly updated, leading to vulnerabilities. The lack of regular updates and patches increases the risk of security breaches.
- *Big Data:* The vast amount of data generated by IoT devices necessitates effective protection measures to ensure privacy and security. Managing the security of such large-scale data without compromising system performance is daunting.
- *Dynamic Topology:* The dynamic nature of IoT networks, where devices can frequently join and leave, makes it challenging to maintain a stable security infrastructure and manage the integrity of the network.

 doi:10.20944/preprints202503.0706.v1

With these limitations in mind, securing IoT ecosystems becomes increasingly challenging, making them vulnerable targets for various attacks [23]. Understanding the attack taxonomy is crucial for developing effective security strategies. Figure 2 illustrates the different types of attacks typically performed on IoT networks, highlighting the necessity for tailored security approaches that address the diverse threats encountered in such environments.

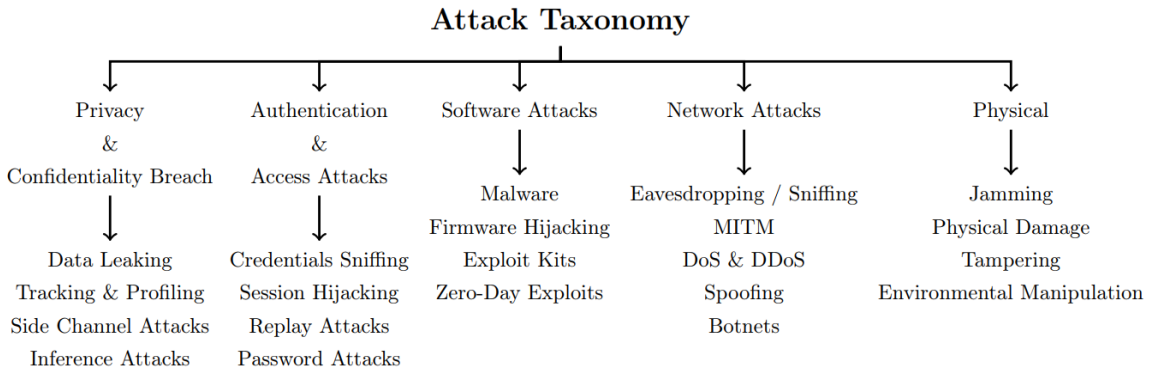

Figure 2. Different types of attacks on IoT networks.

The IoT attack taxonomy outlined in Figure 2 categorizes the attacks into five main categories: Physical Attacks, Network Attacks, Software Attacks, Authentication and Access Attacks, and Privacy and Confidentiality Breaches. Physical attacks, such as jamming and tampering, often require physical access to devices and can be mitigated through robust physical security measures and tamper-evident designs. Network attacks, including eavesdropping, MITM attacks, and botnets, exploit vulnerabilities in the communication channels and can be addressed through secure communication practices and IDS. Software attacks, such as malware and zero-day exploits, target the software stack of IoT devices and necessitate regular updates, patch management, and IDS. Authentication and access attacks focus on compromising user credentials and session integrity, requiring robust authentication mechanisms and continuous monitoring. Privacy breaches, including data leakage and inference attacks, highlight the need for data encryption, access controls, and privacy-preserving techniques.

The majority of attacks performed on IoT networks are software and network-based attacks [11]. Given that IDSs are built to thwart most of these attacks, it seems sensible to explore the development of lightweight IDS for IoT. This study focuses on driving future research towards building such lightweight IDS for IoT systems using ML and DL algorithms.

1.3. Trends in IDS Research: Focus on IoT and Feature Selection

To underline the significance of these developments, Figure 3 showcases the trends in IDS research related to IoT and feature selection extracted from IEEE publications from 2015 to 2024. This analysis illustrates how the research community increasingly focuses on refining IDS capabilities specifically for IoT environments through advanced methodologies like feature selection.

These trends highlight the rising importance of tailored security solutions for IoT and emphasize the critical role of feature selection in enhancing IDS operational efficiency. Such insights are instrumental in driving future research and technological advancements in IoT security.

 doi:10.20944/preprints202503.0706.v1

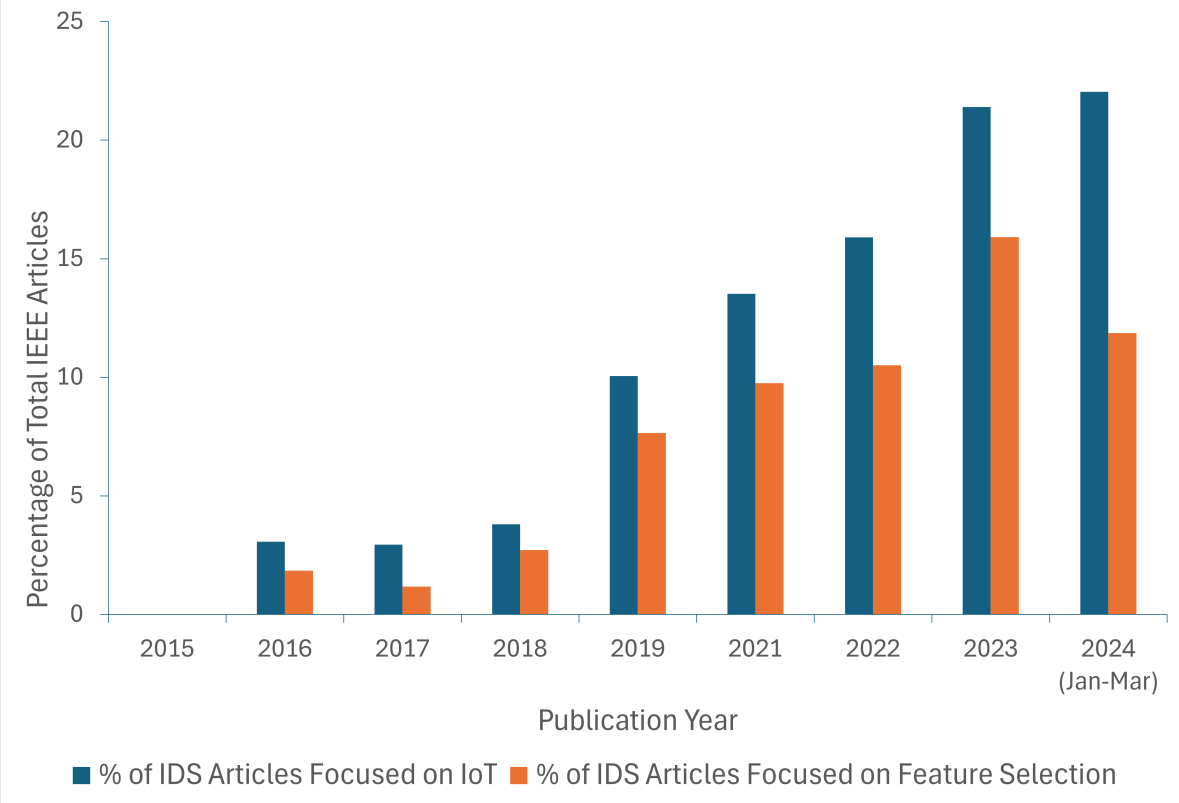

Figure 3. IEEE Trends in IDS Research (Jan 2015-Mar 2024)

1.4. Importance of Dataset Optimization in ML/DL-Based IDS Development

The rapid expansion of IoT has introduced unique challenges in data management, especially in the domain of IDS that utilize advanced ML and DL technologies. These systems depend heavily on comprehensive and accurately curated datasets to perform effectively. A thorough understanding of ML and DL methodologies and their application in cybersecurity is crucial for developing robust IDS. Feature extraction and preparation ensure that only the most relevant and impactful data is processed, reducing computational overhead and enhancing real-time analytical performance. This is particularly crucial in IoT contexts, where devices often have limited processing power and memory [24].

Feature selection plays a pivotal role in the optimization of IDS datasets. Identifying and retaining only the most significant features that influence detection accuracy can significantly reduce the computational burden on IoT devices. This step enhances ML and DL models' performance and conserves energy—a critical consideration in sustained IoT operations. Efficient feature selection methods have been proven to improve model training times and enhance the generalizability of the IDS [25,26].

Furthermore, the balance of datasets is essential for the efficacy of ML/DL-based IDS. Unbalanced datasets can lead to biased models that fail to generalize well, particularly in detecting less frequent, albeit more harmful, attacks. Techniques such as synthetic data generation, under-sampling, and over-sampling are vital in addressing these imbalances. Such methods help ensure that IDS is well-equipped to recognize and respond to various intrusion attempts, thereby maintaining system integrity across different operational scenarios [27].

In conclusion, the development of lightweight and efficient IDS for IoT, leveraging ML and DL, hinges significantly on the quality of the datasets utilized. Optimizing these datasets through strategic feature selection and ensuring their balance is indispensable. Such efforts are fundamental to enhancing the accuracy and responsiveness of IDS solutions, thereby securing IoT networks against an increasingly sophisticated landscape of cyber threats. This focus on dataset optimization lays the

 doi:10.20944/preprints202503.0706.v1

groundwork for the subsequent sections, which delve into specific ML/DL methodologies and the technical implementation of IDS in diverse IoT settings.

1.5. Contributions and Comparative Analysis

This systematic literature review (SLR) offers significant contributions to the field of IoT security, emphasizing the integration of ML and DL in IDS. First, it provides a comprehensive synthesis of the existing research on IoT security threats and various IDS solutions, giving readers a clear overview of current challenges and the state of the art. It underscores the crucial role of dataset optimization, mainly through effective feature selection and data balancing, to enhance the performance of ML/DL algorithms in IDS.

Secondly, the paper thoroughly reviews various ML and DL techniques applicable to IDS, especially in IoT ecosystems, considering the typical constraints such as limited computational power and energy resources. This review does not evaluate these techniques but instead categorizes and discusses their potential based on existing research, highlighting effective strategies for implementing lightweight and efficient IDS solutions.

In addition, this review extensively covers the available datasets most commonly used for training and testing IDS models. A detailed discussion of these datasets highlights their characteristics. This examination is crucial for understanding current IDS approaches' strengths and limitations and identifying gaps where future research could contribute more robust datasets.

Table 1 compares our systematic review against notable survey papers published from 2021 onwards on IDS. This underscores the comprehensive nature of our approach, highlighting our unique emphasis on both ML and DL techniques, as well as critical aspects like feature selection and data imbalance.

Table 1. Comparative analysis of reviews focusing on ML/DL applications in IDS across various studies.

Review Article	Year	Systematic Study	IoT Focused	ML	DL	Feature Selection	Data Imbalance
Mijwil et al. [28]	2023	✗	✓	✓	✗	✗	✗
Lyu et al. [29]	2023	✗	✗	✓	✓	✓	✗
Sarker et al. [30]	2023	✗	✓	✓*	✓*	✓	✗
Sarker [31]	2023	✗	✗	✓	✓	✗	✗
Dasgupta et al. [32]	2022	✗	✗	✓	✓	✗	✗
Halbouni et al. [27]	2022	✗	✗	✓	✓	✗	✗
Gyamfi et al. [33]	2022	✗	✓	✗	✗	✗	✗
Farooq et al. [34]	2022	✗	✓	✓	✓	✗	✗
Faiz et al. [26]	2022	✗	✗	✓	✗	✓*	✗
Bharati et al. [24]	2022	✗	✓	✓	✓	✗	✗
Dixit et al. [35]	2021	✗	✗	✗	✓	✗	✗
Adnan et al. [36]	2021	✗	✓	✓	✗	✗	✗
Lansky et al. [37]	2021	✓	✗	✗	✓	✗	✗
Ahmad et al. [38]	2021	✓	✗	✓	✓	✗	✗
Thakkar et al. [39]	2021	✗	✓	✓	✓	✗	✗
Geetha et al. [40]	2021	✗	✗	✓	✓	✗	✗
This Systematic Review		✓	✓	✓	✓	✓	✓

**Discussed but not in detail.*

Our review is notable for its rigorous approach. It utilizes a comprehensive methodology that assesses ML and DL applications in IDS and emphasizes the importance of optimized feature selection and data balance. This makes our study particularly invaluable for researchers and practitioners aiming to implement efficient IDS in IoT settings where performance and accuracy are crucial.

1.6. Structure of the Study

The study is meticulously organized to guide the reader through the complexities of applying ML and DL in IDS for IoT environments:

 doi:10.20944/preprints202503.0706.v1

- **Section 1: Introduction** - Outlines the IoT landscape, security challenges, and the role of ML/DL in addressing these issues.
- **Section 2: Research Methodology** - Describes the methods used to gather and analyze relevant data systematically.
- **Section 3: IDS Architectures and Advancements in IoT** - Provides an in-depth exploration of various IDS architectures.
- **Section 4: ML/DL Use in IDS** - Examines how different ML and DL models enhance detection capabilities.
- **Section 5: Optimizing IDS with Feature Selection** - Discusses the critical role of feature selection in enhancing IDS efficiency.
- **Section 6: Datasets and Data Balancing in IDS** - Examins a range of datasets specific to IoT and broader network environments and discusses various techniques to ensure data balance, which is crucial for the effectiveness of IDS models in handling real-world uneven data distributions.
- **Section 7: Observations, Challenges, and Future Directions** - Reflects on the findings and outlines potential areas for future research.
- **Section 8: Conclusion** - Summarizes the essential findings and contributions of the study, emphasizing their implications for IoT security.

2. Research Methodology

This SLR investigates the application of ML and DL in IDS, encompassing both general systems and those tailored for the IoT. Spanning research published from 2018 to 2023, the study methodically collects, reviews, and synthesizes findings from a broad range of sources to establish a comprehensive understanding of the field and its evolution towards lightweight IDS solutions for IoT.

An SLR is characterized by a structured and replicable approach that minimizes bias and provides comprehensive coverage of relevant literature. SLRs involve clearly defined questions, predetermined methodology, and specified study eligibility criteria. This process includes comprehensive literature searches, systematic data extraction, and the rigorous evaluation of the quality of each identified study. This structured approach aims to summarize the evidence on a topic and identify and analyze trends and gaps to guide future studies [41]. The methodological rigor ensures the reliability and validity of the review's conclusions, which is essential for advancing the field of IDS, particularly in the context of IoT.

2.1. Objectives and Research Questions

This subsection outlines the objectives and research questions guiding the systematic literature review. These questions are designed to address critical areas in the development of IDS using ML and DL methodologies, particularly within the IoT domain.

Table 2 presents the formulated research questions and the corresponding objectives. Each research question and objective is strategically designed to build upon the insights gathered from the literature, ensuring a comprehensive understanding of the current landscape and future prospects of IDS in IoT environments. This systematic approach helps pinpoint gaps in the existing research and suggests pathways for future inquiries.

 doi:10.20944/preprints202503.0706.v1

Table 2. Research Questions and Objectives

S.No	Research Question	Research Objective
1	What are the prevalent types of cyberattacks targeting IoT devices, and what makes securing IoT ecosystems inherently challenging?	To study the spectrum of cyberattacks on IoT devices and understand the inherent challenges of securing IoT ecosystems.
2	What countermeasures are currently employed to safeguard IoT networks?	To explore the current strategies and technologies for IoT security, specifically focusing on the effectiveness and role of IDS in mitigating these threats.
3	Which ML and DL methodologies are currently applied in IDS, especially for IoT networks, and what are their associated challenges?	To identify and assess the ML and DL techniques utilized in IDS, highlighting their effectiveness and the challenges they face, particularly in IoT networks
4	What datasets are predominantly utilized for evaluating ML or DL-based IDS, and how are issues of data imbalance addressed within these datasets?	To investigate the security datasets frequently used for testing ML or DL-based IDS in both general and IoT-specific networks and to understand the methodologies implemented to tackle issues of data imbalance.
5	What feature selection strategies are employed in preparing datasets for IDS, and how do they contribute to developing lightweight IDS solutions?	To examine the feature selection strategies applied in preprocessing IDS datasets for IoT, identifying their contributions to the development of efficient and lightweight IDS models.
6	What are the emerging trends and future research directions for AI-powered lightweight IDS solutions in the IoT domain?	To identify emerging trends and future research directions in developing AI-powered lightweight IDS solutions for IoT ecosystems.

2.2. Research Study Selection

The research study selection process was meticulously structured to ensure the inclusion of relevant and high-quality articles on IDS using ML and DL within the IoT environment. This section details each step of the systematic approach used to identify and evaluate the literature, as visually summarized in the accompanying flowchart (Figure 4).

Stage 1: Formulation of Problem Statement

The problem statement was clearly defined to guide the systematic review, focusing on the application of ML and DL in enhancing IDS in IoT settings. It centers on the vulnerabilities and scalability challenges that IoT environments face due to high-dimensional and imbalanced data, as well as the limitations of existing IDS solutions. The review aims to explore and critically analyze how advancements in ML and DL, especially through innovative feature selection and data balancing techniques, can lead to more efficient and effective IDS frameworks specifically tailored for the IoT context.

Stage 2: Identify the Keywords and Deciding the Duration of the Study

Keywords were carefully selected based on their relevance to the core topics of IDS, IoT, and ML/DL technologies. The primary keywords included those listed in Table 3. Various combinations of these and equivalent keywords were used to ensure comprehensive topic coverage. The duration of the study was set from 2019 to 2024 to focus on the most recent developments in the field.

 doi:10.20944/preprints202503.0706.v1

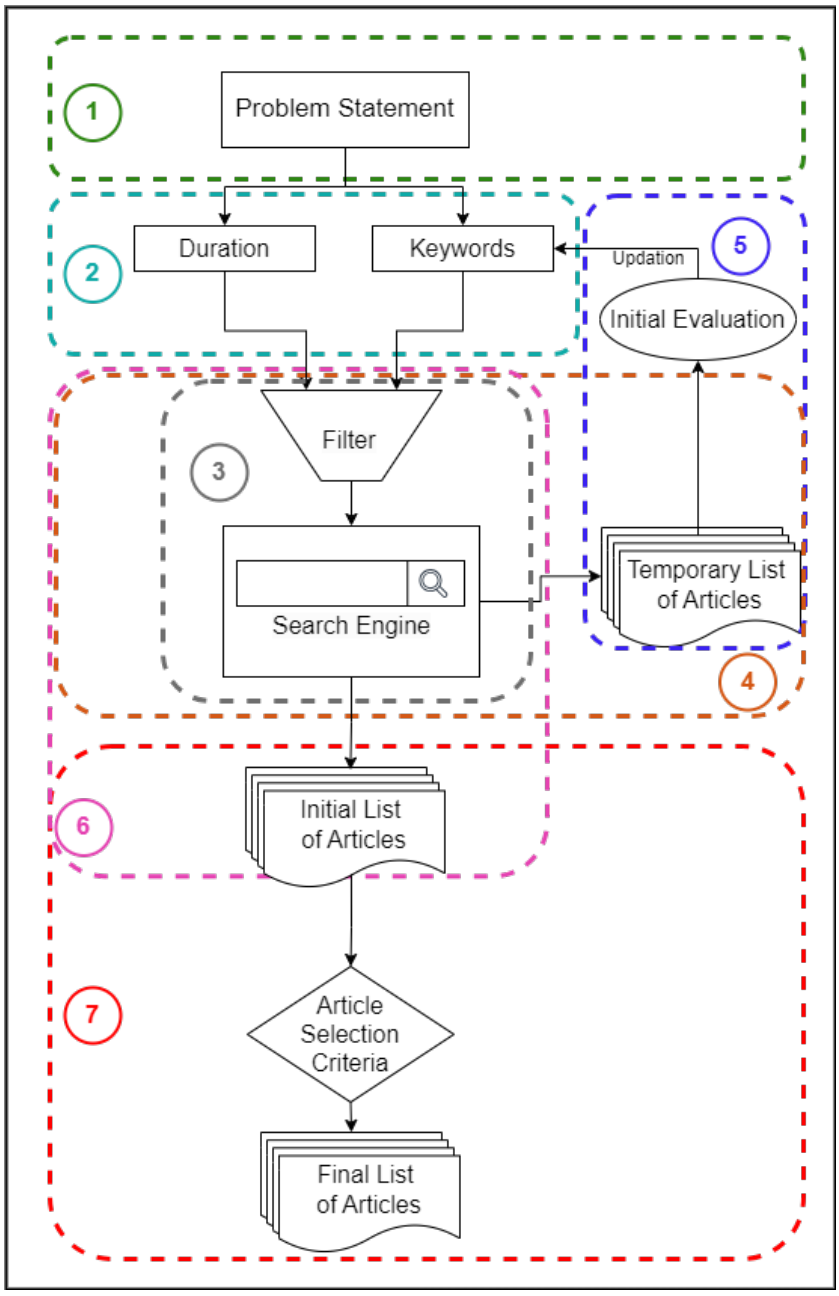

Figure 4. Flowchart of the Research Study Selection Process

Table 3. Keywords used in the search process

S.No	Keywords
1	Intrusion Detection System
2	Internet of Things
3	Security
4	Cyber Security
5	Network Security
6	Machine Learning
7	Deep Learning
8	Lightweight Intrusion Detection System
9	Feature Selection
10	Data (Im)Balance

 doi:10.20944/preprints202503.0706.v1

Stage 3: Identification of Search Engine and Necessary Filters

The search utilized several renowned academic databases with suitable filters to ensure a broad and diverse source of literature, as listed in Table 4.

Table 4. Search Engines and Filters used in the search process

Search Engine	Filters
IEEE Xplore	Publication Years: 2019 to 2024
Springer	Article Type: Research Articles or Conference Proceedings
ScienceDirect	Language: English
Google Scholar	
ResearchGate	

Stage 4: Perform Preliminary Search and Generate Temporary List of Articles

An initial search was conducted using the specified keywords and filters, resulting in a temporary list of potentially relevant articles.

Stage 5: Extract Information from the Temporary List and Update the Keywords

Information was extracted from the temporary list, and keywords were refined to enhance the specificity and relevance of subsequent searches.

Stage 6: Perform the Initial Search and Generate Initial List of Articles

The refined keywords generated a more focused search and an initial list of articles that closely matched the review's objectives.

Stage 7: Apply Article Selection Criteria and Generate the Final List

The final stage of the article selection process involved meticulously applying predefined selection criteria to the initial list of articles. Each article underwent a detailed review to ascertain whether it comprehensively described its methodology and presented practical comparisons or empirical data to substantiate its conclusions. Articles that did not meet these stringent standards were excluded from further analysis. This rigorous vetting process ensured the inclusion of only the most relevant and methodologically robust articles in the review.

Ultimately, this process led to the selection of 98 IDS-related articles, of which 71 specifically addressed IDS implementations in IoT environments. The remaining 27 articles explored ML or DL techniques in IDS implementations and discussed the application of either feature selection or data-balancing techniques.

3. IDS Architectures and Advancements in IoT

IDS is an essential tool in cybersecurity, tasked with monitoring network or system activities to detect unauthorized access, misuse, or policy violations. Their primary function is to alert system administrators or security personnel to potential threats, safeguarding data integrity and operational continuity [42]. As network complexities increase, the role of IDS becomes increasingly crucial, particularly in detecting sophisticated cyber threats that continually evolve to exploit network vulnerabilities [43].

The evolution of IDS from traditional rule-based systems, which rely on known signatures of malicious activities, to more sophisticated systems has been significantly influenced by advancements in computational technologies [44]. This shift has been necessitated by the dynamic nature of modern network environments, where static detection methodologies are insufficient. Today's IDS can incorporate ML and AI to dynamically learn from network behavior and adapt to new, previously unseen threats as they emerge [45].

Modern IDS are designed to be highly adaptive, utilizing advanced computational technologies to handle the increased scalability and diversity of network environments. These systems are particularly vital in settings like the IoT, where the number of connected devices presents unique security challenges.

 doi:10.20944/preprints202503.0706.v1

By employing advanced ML and DL techniques, IDS can detect complex patterns and anomalies indicative of new or evolving threats, ensuring robust defense against a broad spectrum of cyber attacks [46,47].

3.1. Types of IDS and Their Applicability to IoT

IDS can be categorized based on their monitoring strategies and deployment as:

Network-based IDS (NIDS)

NIDS monitors network traffic for signs of intrusion and is strategically placed to oversee all traffic within the network. In IoT environments, NIDS must handle diverse devices and high data volumes, making advanced ML algorithms essential for effective traffic analysis and anomaly detection. Recent studies highlight the effectiveness of deep learning models in improving the accuracy and efficiency of NIDS in IoT settings [48].

Host-based IDS (HIDS)

HIDS is installed directly on devices, monitoring all system interactions and network traffic to detect potential threats. This placement particularly benefits devices that manage sensitive data or critical operations. Optimizing HIDS for IoT involves balancing detection capabilities with the device's computational limits to minimize performance impacts while maintaining robust security [49].

Anomaly-based IDS (AIDS)

AIDS is designed to detect unusual patterns that may signify a threat by comparing activities to a baseline of "normal" network or system behavior. This type of IDS is highly effective in identifying novel attacks that have not been previously encountered or defined in threat databases. The strength of anomaly-based IDS lies in their ability to adapt to network security's dynamic and evolving landscapes, making them especially suited for IoT environments where device behaviors and interactions can be highly variable and unpredictable. These systems employ a range of machine learning and deep learning techniques to continuously refine their understanding of what is considered normal, improving their accuracy over time [50].

Integrating any of these IDS types into IoT architectures requires careful planning to address IoT environments' specific security needs and resource constraints. Each type offers unique benefits: NIDS provides broad network coverage, HIDS offers detailed scrutiny directly on the devices, and AIDS identifies unusual activities through behavior analysis. Together, they form a comprehensive defense strategy against various cyber threats, ensuring robust security across the IoT landscape.

3.2. Advances Through ML and DL

The integration of ML and DL has significantly transformed IDS's capabilities, particularly in addressing the scalability and complexity of IoT environments.

ML in IDS

ML techniques, such as decision tree (DT), Support Vector Machine (SVM), and clustering, have been crucial in improving the anomaly detection capabilities of IDS. DT, for example, effectively categorizes network behaviors, making them suitable for the dynamic nature of IoT networks. SVM, known for its robustness in high-dimensional spaces, is adept at distinguishing between normal and malicious activities with high accuracy. These ML techniques adapt to evolving threat landscapes, continuously improving their predictive accuracy without human intervention [51].

DL in IDS

DL, utilizing architectures like Artificial Neural Networks (ANN), Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN), enhances IDS by detecting complex patterns and anomalies that are difficult for traditional methods to capture. In IoT, where devices generate large volumes of data, neural networks can analyze this data to detect sophisticated attack patterns. The strength of DL techniques like CNN is their ability to learn feature representations from vast amounts of data, automating feature extraction and significantly boosting detection rates. This capability is particularly

 doi:10.20944/preprints202503.0706.v1

beneficial in IoT environments, where the diversity of devices and interactions complicates the security landscape [52].

These advancements in ML and DL improve IDS systems' detection capabilities and ensure they can scale effectively with the expanding IoT infrastructure. The deep integration of these technologies into IDS frameworks enhances their ability to preemptively identify and mitigate threats, thereby fortifying IoT networks against a wide array of cyber threats.

4. ML/DL Use in IDS

Building on the advances discussed in Section 3.2, this section delves deeper into the application of ML and DL in IDS. We will explore the full spectrum of the ML/DL process from data collection to model deployment, detailed in Figure 5, which outlines each critical phase of developing and deploying an ML/DL-based IDS.

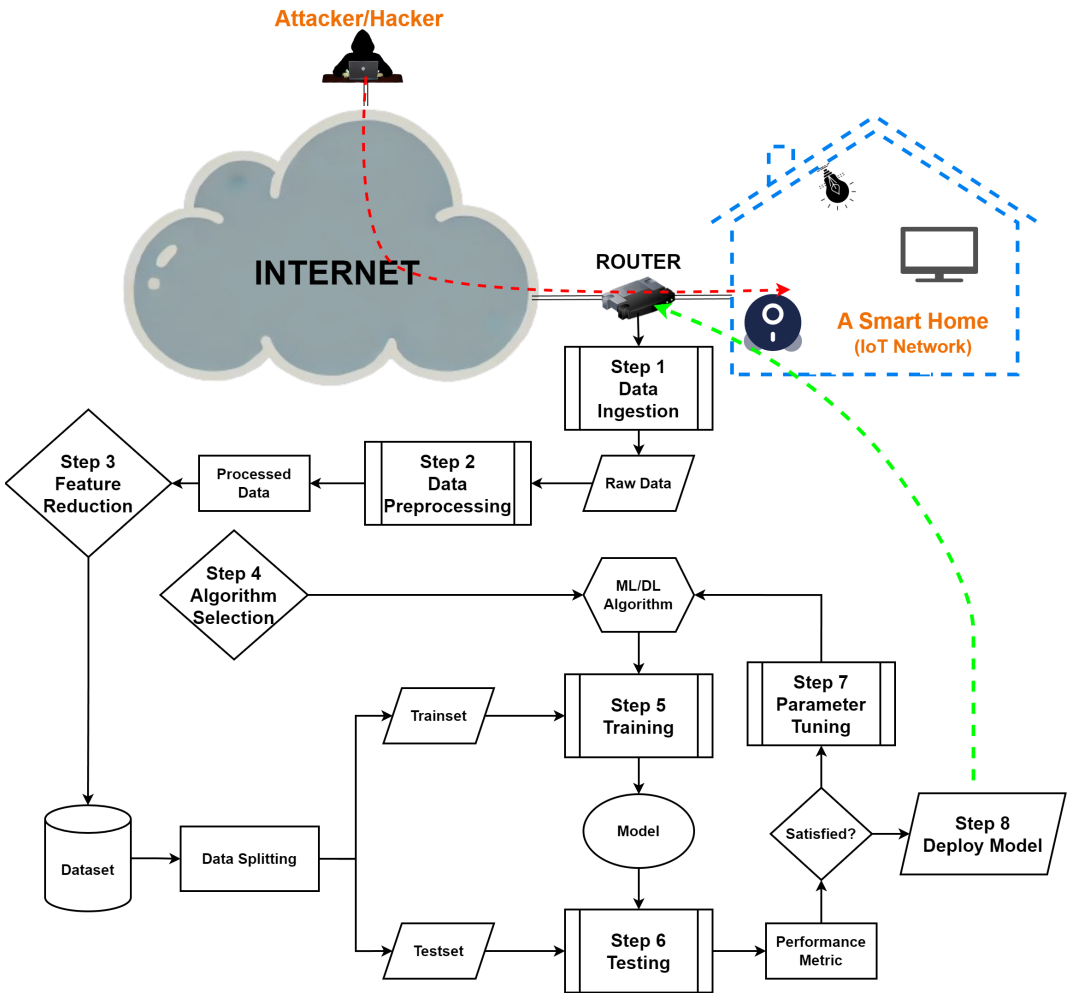

Figure 5. ML/DL Training Process for IDS

Data Ingestion

Data ingestion in IoT IDS involves collecting a broad spectrum of network traffic data, which is crucial for training robust detection systems. This foundational step ensures that the IDS is equipped with comprehensive data reflective of both normal operations and potential security threats.

Data Pre-Processing

Data pre-processing in IoT IDS involves transforming raw data into a structured format that is ready for analysis. This includes using tools like the CICFlowMeter to convert raw network data into

 doi:10.20944/preprints202503.0706.v1

analyzable formats [53]. In recent studies, deep learning techniques such as CNN [54,55] and Long Short-term Memory (LSTM) [56,57] are employed for efficient feature extraction. Data balancing is another crucial aspect of this stage, ensuring the model is not biased towards the majority class, which will be discussed more comprehensively in Section 6.

Feature Reduction

Feature reduction streamlines the data further, enhancing model efficiency and focus using techniques like Feature Selection [58] and Principle Component Analysis (PCA) [59]. More details on feature selection methods and their impact on IDS performance will be explored in Section 5. Post-reduction, the dataset is split into training and testing sets.

Algorithm Selection

Choosing a suitable algorithm is pivotal. This decision is based on the nature of the data and the specific requirements of the IDS. The selection process considers factors like data complexity, the volume of data, and the expected types of intrusion behaviors to be detected. Subsections 4.1 and 4.2 have more detailed discussions focusing on the specific ML and DL algorithms used in IDS.

Training

The training phase involves feeding the prepared and reduced training dataset into the selected ML or DL model. Based on the dataset, the model learns to identify and categorize behavior as usual or potentially malicious.

Testing

Post-training, the model is rigorously tested using a separate testing set to evaluate its performance. This phase is critical for assessing the model's performance and ability to generalize to new data without overfitting the training set.

Parameter Tuning

Parameter tuning is often necessary if the initial testing phase does not yield satisfactory results. This may involve manual adjustments or the application of advanced optimization techniques like the Slime Mold Algorithm (SMA) [60], Modified Red Fox Optimizer (MRFO) [61], Tunicate Swarm Algorithm (TSA) [62], Chaotic Butterfly Optimization Algorithm (CBOA)[63], Genetic Algorithm (GA) [64] and others. The process iterates—training, testing, and tuning—until the model meets the predefined performance criteria.

Deploy Model

Once optimized, the model is deployed within an IDS framework. It operates in real-time, monitors network traffic, filters, and alerts security personnel to potential threats.

The Table 5 summarizes the articles reviewed that discuss various ML and DL techniques used in IDS in IoT environments, providing a snapshot of how these technologies are applied in the field.

Table 5. ML/DL Techniques in IoT-IDS

Reference	ML	DL	Algorithms
[65]	✗	✓	ANN
[59]	✗	✓	DBN
[66]	✓	✗	RF
[67]	✓	✓	NB, BN, DT, LR, SVM, RF, MLP
[68]	✓	✗	SVM, RF, LR, DT
[69]	✓	✓	LR, SVM, DT, RF, ANN
[70]	✓	✗	SVM
[71]	✓	✗	RBM Based Clustering
[72]	✗	✓	AE
[73]	✗	✓	DNN

Continued on next page

 doi:10.20944/preprints202503.0706.v1

Table 5 – continued from previous page

Reference	ML	DL	Algorithms
[74]	✗	✓	MLP, CNN, LSTM
[75]	✓	✓	RF, AdaBoost, GBM, XGB, ETC, DT, MLP
[52]	✗	✓	CNN, LSTM
[76]	✗	✓	DBN
[77]	✗	✓	RNN
[78]	✗	✓	RaNN
[79]	✗	✓	CNN
[80]	✗	✓	CNN, LSTM
[81]	✓	✗	NB, BN, DT, RF, RT
[82]	✗	✓	DCNN
[83]	✓	✗	BIRCH
[84]	✓	✗	kNN, SGD, RF, LR, NB
[85]	✓	✗	RF, kNN, XGB
[86]	✗	✓	ANN
[87]	✗	✓	CGAN
[88]	✓	✓	Bagging, DT, kNN, MLP
[89]	✗	✓	CFNN
[90]	✗	✓	LSTM
[91]	✗	✓	CNN, LSTM, GRU
[56]	✗	✓	CNN
[92]	✓	✓	SVM, kNN, LDA, QDA, DT, MLP, LSTM, AE
[93]	✗	✓	BiLSTM, GRU
[94]	✗	✓	CNN
[95]	✗	✓	DNN
[96]	✗	✓	SDPN
[97]	✓	✗	SVM, LR, RF
[98]	✓	✗	RF, XGB, LGBM, CatBoost
[99]	✓	✗	XGB, CatBoost, kNN, SVM, QDA, NB
[100]	✗	✓	DCNN
[101]	✗	✓	DNN
[102]	✓	✗	RELM
[103]	✗	✓	CNN
[104]	✗	✓	ANN
[105]	✓	✗	LR, LDA, NB, DT, RF, SVM, GBM
[106]	✓	✗	RF
[107]	✗	✓	CNN, LSTM
[108]	✓	✓	SVM, NB, CGAN
[109]	✓	✗	kNN, XGB, DT, RF
[60]	✗	✓	MESNN
[110]	✓	✗	LR, DT, RF, NB, kNN, SVM
[63]	✗	✓	CNN, QRNN
[111]	✗	✓	CVAE
[112]	✓	✓	LR, DT, RF, NB, GRU, RNN, LSTM, BiLSTM
[113]	✓	✓	LR, FFNN, CNN, AE, Vanilla RNN, LSTM, GRU
[114]	✓	✗	Ensemble Learning
[115]	✓	✓	LR, SVM, MLP

Continued on next page

 doi:10.20944/preprints202503.0706.v1

Table 5 – continued from previous page

Reference	ML	DL	Algorithms
[53]	✓	✓	Adaboost, NB, kNN, QDA, DT, RF, MLP
[58]	✓	✓	RF, DT, kNN, MLP
[116]	✓	✓	Adaboost, DT, RF, ANN, LSTM, AE
[117]	✗	✓	CNN
[118]	✓	✗	An Ensemble of RF, XG, kNN, DT
[119]	✗	✓	REDNN, REFDNN
[120]	✗	✓	LSTM
[121]	✗	✓	LSTM
[122]	✓	✗	SVM
[123]	✗	✓	CNN, LSTM
[64]	✗	✓	Deep Transfer Learning with GA
[124]	✓	✗	Adaboost, kNN, NB
[125]	✗	✓	CNN, LSTM
[126]	✗	✓	CNN, LSTM, GRU
[127]	✓	✗	NB, RF, kNN, SVM

4.1. ML Application in IDS

In this subsection, we explore using ML algorithms in IDS, focusing on traditional and ensemble learning approaches. ML techniques are pivotal in IDS because they can efficiently process and classify vast data, identifying patterns that signify potential security threats.

4.1.1. Traditional Machine Learning Algorithms

Traditional ML algorithms are the foundation of many predictive modeling systems in IDS due to their effectiveness and simplicity in handling various data types. The following are some of the key algorithms used:

- **Logistic Regression (LR)**: Often used for binary classification tasks, LR models the probabilities for classification problems, such as distinguishing between normal and malicious activities [97].
- **Support Vector Machine (SVM)**: SVM is robust in high-dimensional spaces, making it suitable for IDS where it constructs a hyperplane in a multidimensional space to separate different classes [70].
- **Decision Trees (DT)**: DTs are popular for their simplicity and interpretability. They use a tree-like model of decisions and their possible consequences [68].
- **Naïve Bayes (NB)**: This algorithm applies Bayes' Theorem with the assumption of independence between features. It's particularly effective in IDS for its speed and performance with large datasets [81].
- **k-Nearest Neighbors (kNN)**: A non-parametric method for classifying data based on the closest training examples in the feature space [84].

4.1.2. Ensemble Learning Algorithms

Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone [114]. Some of the notable ensemble algorithms include:

- **Random Forest (RF)**: An ensemble of Decision Trees, typically trained with the "bagging" method, RF is very effective in IDS due to its ability to reduce overfitting and improve model accuracy [66].

 doi:10.20944/preprints202503.0706.v1

- **AdaBoost**: Works by combining multiple weak classifiers to form a robust classifier. AdaBoost is used in IDS to enhance classification accuracy by focusing more on complex cases [124].
- **Gradient Boosting Machine (GBM)**: GBM builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions, making it versatile for various IDS tasks [75].
- **Extreme Gradient Boosting (XGB)**: Known for its efficiency, flexibility, and portability, XGB delivers high performance and speed when processing large volumes of data [75].
- **CatBoost**: This algorithm excels in handling categorical features and is less prone to overfitting than other methods [98].

4.1.3. Innovative ML Implementations in IDS:

This section outlines various innovative implementations of ML across different studies, illustrating their contribution to enhancing security capabilities in IDS systems.

- **RF in Smart Cities:** The Anomaly Detection IoT (AD-IoT) system [66] utilizes the RF algorithm to detect compromised IoT devices within smart city networks. This system is designed for distributed fog networks, aiming to enhance the responsiveness and scalability of attack detection.
- **Multi-Algorithm IDS for Smart Homes:** Anthi et al. [67] developed a novel three-layer IDS for IoT networks within smart homes. This system employs multiple algorithms such as NB, Bayesian Networks (BN), DT, LR, SVM, and RF to classify device behaviors and detect malicious activities.
- **DDoS Attack Detection:** Chaudhary & Gupta [68] addressed significant security vulnerabilities associated with DDoS attacks through compromised IoT devices. The authors propose a new ML approach for efficient DDoS traffic detection within local network routers. Upon careful testing and evaluation, RF outperformed LR, SVM, and DT.
- **Comprehensive Algorithm Evaluation:** A study by Hasan et al. [69] evaluated the effectiveness of various ML algorithms like LR, SVM, DT, and RF in detecting attacks within IoT systems.
- **Edge Computing in Smart Homes:** The implementation discussed in [128] leverages an ML classifier, specifically a Radial Basis Function (RBF)- SVM, within a smart home system simulation developed on Alibaba ECS, focusing on enhancing network security detection using edge computing technology.
- **DoS Protection with ML:** Verma & Ranga [75] explored the application of various ML classifiers to protect IoT systems against DoS attacks, assessing their performance on prominent IoT datasets and implementing them on IoT-specific hardware.
- **ML Classifier Evaluation Framework:** The framework introduced by Shafiq et al. [81] incorporates a hybrid algorithm based on a bijective soft set approach to evaluate and select the most effective ML algorithm from several available options for identifying cyber-attacks in IoT networks.
- **Network Intrusion Detection:** The study by Kocher and Kumar [84] applied multiple ML classifiers to the recent UNSW-NB15 dataset to evaluate their performance in network intrusion detection.

4.2. DL Application in IDS

DL has significantly enhanced the capabilities of IDS through its ability to perform complex feature extraction and pattern recognition tasks. DL models excel in identifying subtle patterns in high-dimensional data, often indicative of cybersecurity threats.

 doi:10.20944/preprints202503.0706.v1

4.2.1. Traditional DL Algorithms

- **Multi-Layer Perceptron (MLP)**: MLPs are fundamental neural networks with one or more hidden layers between input and output layers. They are effective for pattern recognition tasks due to their ability to learn non-linear decision boundaries [115].
- **Artificial Neural Networks (ANN) and Deep Neural Networks (DNN)**: ANNs are the backbone of many DL approaches [65], with DNNs representing an extension of ANNs that contain multiple hidden layers [95]. These structures are adept at processing complex datasets commonly found in IDS. DNNs, in particular, enhance the capability to capture deeper levels of data abstraction.
- **Convolutional Neural Networks (CNN)**: CNNs are particularly beneficial for feature extraction because they can process data in a grid-like topology, such as images or time series. They use convolutional layers to detect essential features automatically without any human supervision. CNNs are utilized in both one-dimensional (1D) and two-dimensional (2D) forms for analyzing network traffic and log data [79].
- **Recurrent Neural Networks (RNN)**: RNNs are designed to handle sequential data, making them suitable for time-series analysis in IDS [77]. LSTM units [121] and Gated Recurrent Unit (GRU) [126] are enhancements over traditional RNNs, providing solutions to the vanishing gradient problem and improving the memory capacity of the model.
- **Autoencoders (AE)**: AEs are used for unsupervised learning of efficient coding. They learn to compress (encode) the input into a more miniature representation and then reconstruct (decode) the output from this representation. Convolutional Variational Auto Encoders (CVAE) [72] is a type of AE that provides a probabilistic manner for describing observations in latent space.
- **Generative Adversarial Networks (GAN)**: GANs involve two neural networks, a generator and a discriminator, which compete against each other. This structure is highly effective for generating new data samples. Conditional Generative Adversarial Networks (CGANs) extend GANs by adding a condition to the generation process, enhancing their applicability in IDS for generating realistic attack scenarios to test systems [87].
- **Deep Belief Networks (DBN)**: DBNs are generative models that consist of multiple layers of stochastic, latent variables. They are effective in feature extraction and classification tasks [59].

4.2.2. Select Deep Learning Implementations in IDS:

This subsection highlights diverse research efforts that have effectively integrated DL techniques to advance the capabilities of IDS, focusing on feature extraction and comprehensive modeling.

- **DDoS Attack Detection Using ANN:** Soe et al. [65] developed a detection system tailored for IoT environments to identify DDoS attacks caused by malware like Mirai efficiently. This implementation leverages an ANN.
- **Hybrid IDS Using DBN:** In the work presented by Yang et al. [59], a hybrid IDS was proposed using DBN combined with a modified density peak clustering algorithm (MDPCA). This system segments the training data into manageable clusters, thereby reducing data imbalance and enhancing the detection of minority class attacks, while DBNs are used for high-level feature extraction.
- **Three-Layer IDS for Smart Homes:** The research by Latif et al. [64] introduced a novel three-layer IDS designed explicitly for Industrial Internet of Things (IIoT) networks. Using deep transfer learning (DTL) and a tri-layer architecture combining CNN, GA, and bootstrap aggregation.
- **Innovative IDS with DNN and Transfer Learning:** The study outlined by Qureshi et al. [72] utilized a DNN combined with ASTL. This approach integrates features extracted from a pre-trained network with original data features, enhancing the IDS's capability to detect network security breaches.

 doi:10.20944/preprints202503.0706.v1

- **Hybrid CNN-LSTM Model for IoT Security:** Roopak et al. [74] explored a hybrid CNN-LSTM model focused on combating DDoS attacks in IoT networks. This model efficiently handles spatial and temporal data, significantly improving the system's capability to detect and respond to cybersecurity threats.
- **Advanced RNN for Fog Computing:** The implementation by Almiani et al. [77] detailed an advanced IDS tailored for the security needs of Fog computing environments essential for IoT. This system employed a multi-layered RNN to detect various cyber threats effectively.

4.3. Performance Measures in IDS

Various metrics are utilized to evaluate the performance of IDS, which employs ML and DL techniques effectively. These metrics help quantify the effectiveness of the IDS in identifying legitimate threats while minimizing false alarms. Below, we explore these measures in detail.

4.3.1. Confusion Matrix

The confusion matrix not only gives a count of correct and incorrect classifications, but it also provides a clear visual representation of the model's performance, especially in terms of distinguishing between classes. Here's a deeper look at each element and its significance:

- **True Positives (TP)**: Correctly predicted positive observations.
- **True Negatives (TN)**: Correctly predicted negative observations.
- **False Positives (FP)**: Incorrectly predicted as positive, also known as Type I error.
- **False Negatives (FN)**: Incorrectly predicted as negative, also known as Type II error.

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

4.3.2. Key Performance Metrics

Accuracy: Represents the ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision (Positive Predictive Value): Indicates the proportion of identifications that were actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall/Sensitivity (True Positive Rate): Reflects the proportion of actual positives correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity (True Negative Rate): Shows the proportion of actual negatives that were correctly identified.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

F1-Score: Combines precision and recall into a single metric using their harmonic mean, useful for unbalanced classes.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Area Under the ROC Curve (AUC): The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's performance. As shown in Figure 6 it plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC provides a

 doi:10.20944/preprints202503.0706.v1

single scalar value to summarize the classifier's ability to discriminate between positive and negative classes. A higher AUC value indicates better classifier performance. The plot in Figure 6 illustrate the ROC curves for different classifiers as follows:

- **Perfect Classifier**: This classifier reaches the top left corner of the plot, indicating a TPR of 1 with a FPR of 0. It has an AUC of 1.00.
- **High-Performance Classifier**: This classifier performs well with a high TPR and a low FPR, achieving an AUC of 0.88.
- **Low-Performance Classifier**: This classifier performs modestly, with an AUC of 0.65. It shows a higher FPR for a given TPR than the high-performance classifier.
- **Random Classifier**: This classifier performs no better than random guessing, resulting in an AUC of 0.50, represented by a diagonal line from (0,0) to (1,1).

The ROC curve and AUC are essential tools for evaluating the performance of binary classifiers in various applications, including medical diagnosis, fraud detection, and machine learning model assessment.

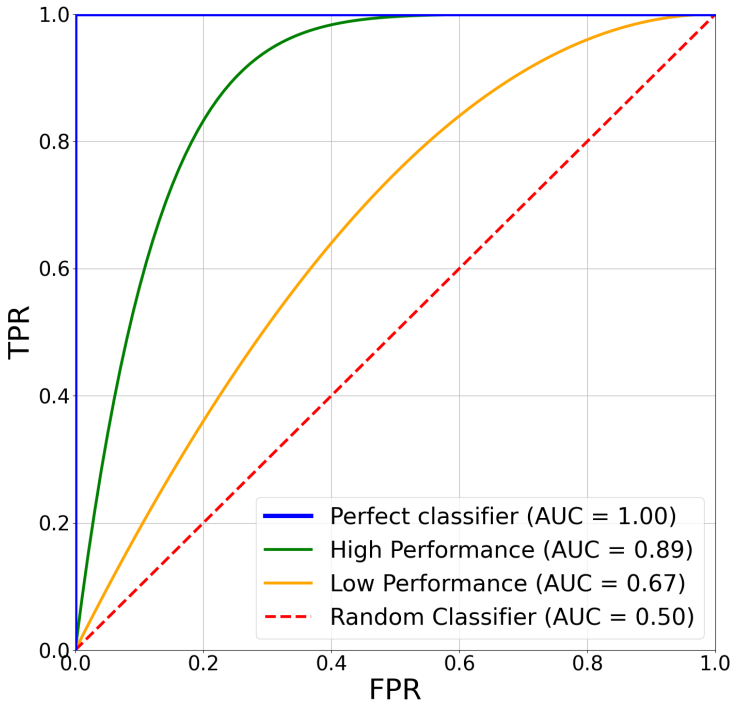

Figure 6. Sample ROC curve showing the performance of different classifiers.

Detection Rate (DR): Similar to recall, indicates how well the system identifies positive instances.

$$\text{DR} = \frac{TP}{TP + FN}$$

False Alarm Rate (FAR): Measures the probability of falsely classifying a negative observation as positive.

$$\text{False Alarm Rate} = \frac{FP}{FP + TN}$$

 doi:10.20944/preprints202503.0706.v1

Matthews Correlation Coefficient (MCC): Provides a balanced measure that considers both the size of positive and negative elements in the dataset.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Kappa: Assesses the agreement between observed and expected accuracy, providing insight into the accuracy obtained by chance.

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement, and P_e is the expected agreement by chance.

These metrics provide a comprehensive toolkit for analyzing the performance of IDS models, offering insights into their strengths and weaknesses in real-world scenarios.

To provide a comprehensive overview, a detailed summary located in Appendix 1 extends beyond the discussions in Section 4. It presents an in-depth study of various implementations of ML and DL algorithms for IDS, collating research efforts from IoT ecosystems and traditional network environments. This summary illustrates the wide-ranging applications and effectiveness of these advanced computational techniques in enhancing the capabilities of IDSs. It serves as a valuable resource for understanding diverse methodologies and their impacts across different security contexts.

5. Optimizing IDS with Feature Selection

As discussed in Section 1.4 on the importance of dataset optimization for ML/DL-based IDS development, feature selection emerges as a critical strategy for creating lightweight IDS for IoT. By eliminating unnecessary features in attack classification, feature selection enhances the IDS's attack detection capabilities and significantly reduces the computational load on IoT devices. This step is crucial in improving the performance of ML and DL models and conserving energy, which is vital for the sustainable operation of IoT systems.

This section will delve deeper into various feature selection techniques and their application in IDS, mainly focusing on how they contribute to developing lightweight and efficient IDS solutions for IoT environments. We will explore different approaches to feature selection, including filter, wrapper, and embedded methods, each offering unique advantages and suited for various scenarios within the IDS framework. q

5.1. Feature Selection & Types

Feature selection is a critical process in machine learning that involves selecting a subset of relevant features for model construction. It helps enhance the performance of machine learning models by eliminating irrelevant or redundant data, reducing dimensionality, and speeding up the learning process [129]. Feature selection techniques were broadly classified into three categories.

- **Filter Methods:** These methods apply a statistical measure to assign a scoring to each feature. The measure ranks the features and is either selected to keep or removed from the dataset. Filter methods are generally faster and less computationally expensive than other feature selection methods because they do not involve training models. A common filter method is the correlation-based feature selection, which measures the correlation between each feature and the target variable [25].
- **Wrapper Methods:** Wrapper methods consider selecting a set of features as a search problem, where different combinations are prepared, evaluated, and compared to other combinations. A predictive model evaluates a combination of features and assigns a score based on model accuracy. Wrapper methods can be very computationally intensive and usually involve training a new model for each feature (or combination of features) added or removed [25].

 doi:10.20944/preprints202503.0706.v1

- **Embedded Methods:** Embedded methods perform feature selection during the model training process and are specific to given learning algorithms. These methods integrate feature selection as part of the training process and are more efficient than wrapper methods since they include feature selection as part of the model construction process. Techniques like Lasso and Ridge regression are embedded methods that regularize coefficients to zero to reduce the number of features [130].

5.2. General Approaches for Feature Selection in IDS

5.2.1. Filter-Based Feature Selection in IDS

Filter-based methods assess the importance of features using statistical tests. These methods are favored for their efficiency and are especially suitable for high-dimensional data in IDS. Here's a detailed look at several statistical measures used in filter-based feature selection, alongside their practical implementations in the field of IDS:

Correlation: Measures the linear relationship between two variables, focusing on how closely changes in one variable are associated with changes in another. In IDS, correlation is used to select features that strongly relate to the detection of threats. For instance, the research by Woo et al. [131] and Dat-Thinh et al.[92] utilize correlation-based filters to streamline feature sets for enhanced intrusion detection accuracy.

$$\text{Correlation} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Information Gain (IG): Assesses how much information a feature provides about the class, calculating the reduction in entropy. IG has been effectively applied in IDS to filter out less informative features, enhancing model performance as shown in studies by Kasongo & Sun [132] and by Wirawan et al.[133].

$$\text{IG}(T, f) = H(T) - H(T|f)$$

where $H(T)$ is the entropy of the target variable, and $H(T|f)$ is the conditional entropy of the target given feature f.

Mutual Information (MI): A measure that captures the amount of information obtained about one random variable through another. It is instrumental in IDS for identifying features that share MI with the class attribute. The implementation of MI in IDS is evident in works like [134] and [58], where it helps select features that contribute significantly to classification accuracy.

$$\text{MI}(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log\left(\frac{p(x, y)}{p(x)p(y)}\right)$$

where $p(x, y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y, respectively.

Chi-Square: This test measures the lack of independence between a feature and the class. It is widely used in IDS to determine which features are statistically significant to the class outcomes. Chi-Square tests are highlighted in the works of Kocher & Kumar [84] and Gad et al. [109], helping to identify features strongly associated with the presence or absence of intrusions.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency, and E_i is the expected frequency under the null hypothesis.

 doi:10.20944/preprints202503.0706.v1

Gain Ratio: An extension of IG that normalizes its values to reduce bias towards multi-valued features. This method is particularly effective in environments where features vary in the number of levels. Studies by Kumar & Gupta [85] and Albulayhi et al. [88] have employed the Gain Ratio to ensure a balanced consideration of feature importance across different feature types.

$$\text{Gain Ratio} = \frac{\text{IG}(T, f)}{H(f)}$$

where $H(f)$ is the entropy of feature f.

5.2.2. Wrapper and Embedded Feature Selection in IDS

Wrapper and embedded methods represent sophisticated approaches to feature selection in IDS. Wrapper methods evaluate feature subsets using the performance of predictive models, while embedded methods incorporate feature selection directly into the learning algorithm.

Wrapper Methods in IDS

Wrapper methods assess feature subsets based on the predictive accuracy of a model. Although computationally intensive, they often yield higher-performing feature sets than filter methods. These methods utilize a specific ML algorithm to evaluate the effectiveness of feature combinations iteratively:

- **SVM-Based Wrapper**: A study by Taher et al. [135] uses an SVM to evaluate feature subsets, aiming to maximize classification accuracy directly.
- **Extra Tree (ET) Wrapper**: The ET classifier serves as a wrapper in the study by Kasongo & Sun [136], assessing feature sets and leveraging its ensemble nature to evaluate feature importance effectively.
- **Non-dominated Sorting Genetic Algorithm (NSGA) with LR**: Khammassi & Krichen [137] combined the NSGA with LR in a wrapper approach to optimize feature combinations, balancing IDS complexity and performance.
- **Rf Wrapper**: Research by Kumar & Gupta [85] employs RF to evaluate features, utilizing its inherent feature importance measures to guide the selection process.
- **Multiple Classifier Wrappers**: The work by Rahman et al. [86] uses various classifiers, including SVM, DT, and NB, within a wrapper framework to identify the most compelling features for detecting network intrusions.

Embedded Methods in IDS

Embedded methods integrate the feature selection process within the training phase of the model, making them particularly efficient as they include feature selection as part of the algorithm's objective function:

- **DT Embedded Method**: DT inherently evaluates feature importance during model training, effectively reducing the feature space without separate validation, as shown in the work of Sarker et al. [138].
- **DL Embedding**: Yu & Bian [139] employed CNN and DNN to automatically select features through their training process, adjusting weights in network layers that correspond directly to feature relevance.
- **RF Regressor Embedded Method**: As detailed in the work by Malathi and Padmaja [53], embedding feature selection within an RF Regressor allows for simultaneous learning and feature evaluation, enhancing the model's focus on the most predictive features.

 doi:10.20944/preprints202503.0706.v1

5.3. Nature-Inspired Optimization Algorithms for Feature Selection in IDS

Nature-inspired optimization algorithms mimic biological processes and natural phenomena to solve complex optimization problems. These algorithms are particularly effective for feature selection in IDS, optimizing the feature set to enhance detection accuracy while reducing computational complexity.

5.3.1. Filter Implementation in IDS

In the filter approach, nature-inspired algorithms assess feature relevance independently of any learning algorithm, focusing on the data's intrinsic properties.

- **Reptile Search Algorithm (RSA)**: The authors Dahou et al. [56] utilized RSA for its efficiency in exploring and exploiting the search space, optimizing feature selection by mimicking reptilian motion, which significantly improves IDS performance.
- **Spider Monkey Optimization (SMO)**: The study by Otoum et al. [96] employs SMO, inspired by the foraging behavior of spider monkeys, to optimize the selection of features, thereby enhancing the detection capabilities of IDS systems.
- **Particle Swarm Optimization (PSO)**: PSO's mechanism of social information sharing is utilized to guide the search towards optimal feature sets, significantly improving computational efficiency and detection rates in IDS [124].
- **Improved Dynamic Sticky Binary PSO (IDSBPSO)**: This variant [106] introduces dynamic adjustments to particles' behavior, sticking to promising areas of the search space to find the best feature subsets.
- **African Vulture Optimization (AVO)**: Implemented by Alsirhani et al. [140], AVO draws inspiration from the scavenging behavior of vultures to efficiently scan and select optimal feature sets for IDS.

5.3.2. Wrapper Implementation in IDS

Wrapper methods utilize nature-inspired algorithms to select features based on the performance improvements they offer to a specific predictive model used within the IDS.

- **Discrete Variant of the Cuttlefish Algorithm (DF-CFA)**: Al-Daweri et al. [141] apply this algorithm, which mimics the adaptive coloration of cuttlefish, to optimize the feature space and enhance model accuracy specifically tailored for intrusion detection.
- **Firefly Optimization (FFO)**: Saraeian & Golchi [142] and Karthikeyan et al. [122] have implemented FFO algorithm for feature selection. It simulates the behavior of fireflies, which is particularly effective in finding global optima in the search space, thereby selecting the most relevant features for IDS.
- **Genetic Algorithm (GA)**: GA's evolutionary strategies are leveraged by Zhang et al. [143] to evaluate and select features that maximize the performance of the IDS.
- **Flamingo Search Algorithm (FSA)**: This algorithm optimizes feature selection by mimicking the foraging behavior of flamingos, focusing on efficiently identifying relevant features to improve detection rates [60].
- **Modified Equilibrium Optimization Algorithm (MEOA)**: This algorithm enhances traditional equilibrium optimization by introducing modifications that adapt better to the IDS's requirements, selecting features that significantly boost system performance [62].
- **Modified FFO**: The work by Almuqren et al. [63] enhances the standard FFO algorithm to suit the complex feature spaces in IDS better, improving both the efficiency and accuracy of the intrusion detection system.

 doi:10.20944/preprints202503.0706.v1

- **Chaotic Binary Pelican Optimization Algorithm (CBPOA)**: As used by Alrowais et al. [111], this algorithm introduces a variation of POA using chaotic sequences, enhancing the exploration capabilities and ensuring a more diverse search for optimal features in IDS.

6. Datasets and Data Balancing in IDS

In IDS, the quality and balance of datasets are crucial for training effective ML/DL models. Data balancing addresses the common class imbalance issue within IDS datasets, where specific network threats or behaviors are significantly underrepresented. This imbalance can lead to biased models that perform well on overrepresented classes but poorly on underrepresented ones, ultimately compromising the IDS's ability to detect and respond to various threats accurately.

This section explores the datasets used in IDS and the importance of balancing these datasets. Effective data balancing ensures that the trained IDS models can generalize well across all types of network behavior and attacks, improving the sensitivity and specificity of threat detection. Subsequent discussions will delve into various techniques implemented to achieve data balance, which is vital for enhancing IDS' overall performance and reliability in real-world scenarios.

6.1. IDS Datasets

Table 6 lists 9 IoT-specific and 12 general IDS datasets used in research, crucial for training and testing IDS solutions across various network scenarios.

Table 6. List of IDS datasets.

Dataset	Year	IoT Specific?
CICIoT2023 [144]	2023	✓
EDGE-IIOTSET [145]	2022	✓
WUSTL-IIOT-2021 [146]	2021	✓
IOTID20 [147]	2020	✓
MQTT-IoT-IDS2020 [148]	2020	✓
TON-IoT [149]	2020	✓
IoT-23 [150]	2019	✓
BoT-IoT [151]	2018	✓
CSE-CICDIS2018 [152]	2018	✗
DS2OS [153]	2018	✗
Kitsune Network Attack Dataset [154]	2018	✗
N-BAIOT [155]	2018	✓
CIDDS001 [156]	2017	✗
CICDoS (2017) [157]	2017	✗
CIC-IDS2017 [158]	2017	✗
UNSW-NB15 [159]	2015	✗
CTU-13 [160]	2013	✗
ISCXIDS2012 [161]	2012	✗
ISCX NSL-KDD [162]	2009	✗
Kyoto 2006+ [163]	2006	✗
KDDCup99 [164]	1999	✗

6.1.1. IoT IDS Datasets

CICIoT2023 Dataset

The CICIoT2023 dataset ([144]) is a comprehensive collection designed to reflect the complexities of IoT environments and facilitate advanced IDS testing. This dataset encompasses a variety of attack types, making it particularly valuable for evaluating IDS capabilities across different IoT scenarios. Notable for its diversity, CICIoT2023 includes numerous categories of attacks, from common DDoS attacks to sophisticated IoT-specific threats like Mirai and Bashlite. As visualized in Figure 7, the distribution of attack types highlights a predominant focus on DDoS attacks across various protocols

 doi:10.20944/preprints202503.0706.v1

and techniques, ranging from SYN floods to UDP-based attacks. This array provides researchers with a realistic spectrum of challenges, simulating a variety of intrusion attempts that an IoT network might face in real-world deployments. The dataset's detailed classification of attack types aids in fine-tuning IDS algorithms to detect subtle nuances between different malicious activities and benign traffic, ensuring a robust defense mechanism for IoT networks.

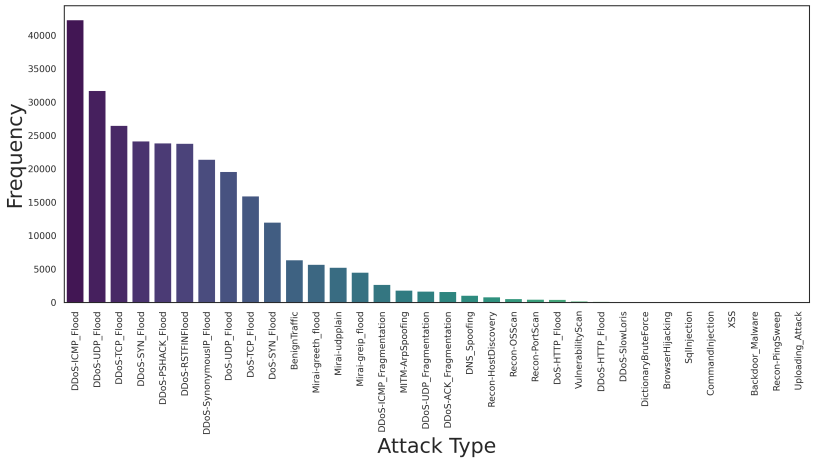

Figure 7. CICIoT2023 Dataset Distribution.

Edge-IIoT Dataset

The Edge-IIoT Dataset dataset, compiled in 2022, is tailored for intrusion detection within edge computing environments integral to IIoT architectures. The data composition, shown in Figure 8, illustrates a balanced mixture of attack types and normal activities, crucial for developing robust IDS models. This dataset includes a diverse range of attack types, such as DDoS over different protocols (UDP, ICMP), ransomware, SQL injections, and less frequent but critical threats like XSS and MITM attacks. Notably, the dataset's variety supports the development of models that recognize and differentiate between nuanced attack vectors and normal network behavior, which is essential for maintaining security in dynamic edge computing scenarios.

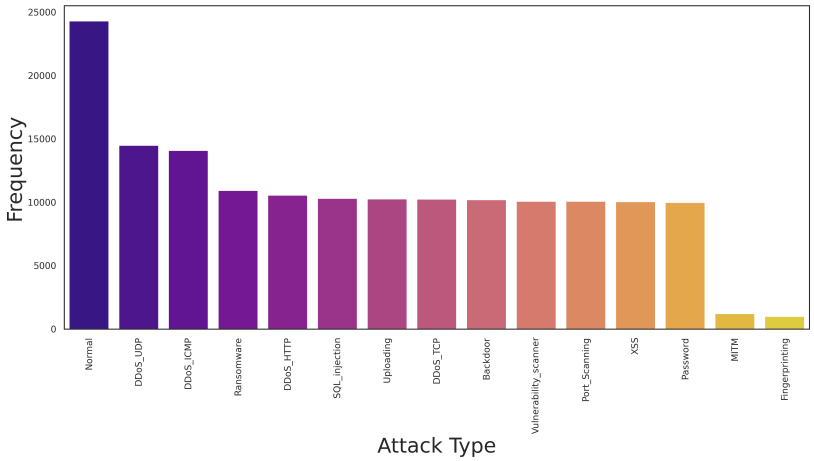

Figure 8. Edge-IIoT Dataset Distribution.

WUSTL-IIoT-2021 Dataset

The WUSTL-IIoT-2021 dataset focuses on the security of IIoT environments and is designed to address complex security challenges in these settings. It features a rich collection of attack scenarios relevant to industrial networks, such as Mirai and Bashlite botnet attacks, demonstrating how com-

 doi:10.20944/preprints202503.0706.v1

promised IoT devices can be used to conduct massive DDoS attacks. The dataset provides a diverse array of attack types, including various forms of DDoS, reconnaissance, and infiltration activities, offering researchers a comprehensive resource to develop and validate IDS solutions for the IIoT. The attack distribution within the dataset, illustrated in Figure 9, highlights the prevalence of DDoS attacks, underscoring their significance in IIoT security challenges.

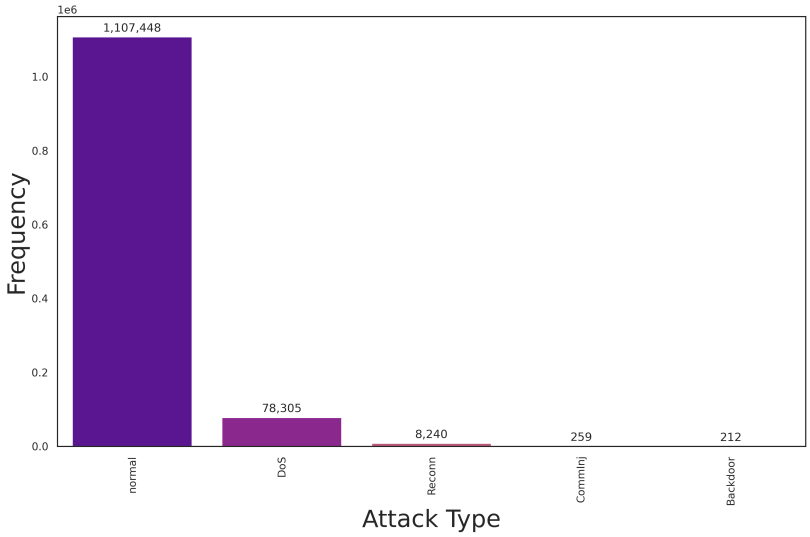

Figure 9. WUSTL-IIoT Dataset Distribution.

IoTID20 Dataset

The IoTID20 dataset, developed in 2020, explicitly addresses security in IoT networks and provides a detailed view of various attack vectors prevalent in IoT environments. This dataset is highly regarded for its comprehensive collection of IoT-specific attack types, including DDoS, command injection, and several others tailored to IoT scenarios. The distribution of attack types in the IoTID20 dataset, as shown in Figure 10, illustrates various attack categories, highlighting the dataset's utility in training and testing IDS designed to safeguard IoT ecosystems. This extensive range of attacks helps fine-tune IDS capabilities to effectively detect and mitigate IoT-specific threats.

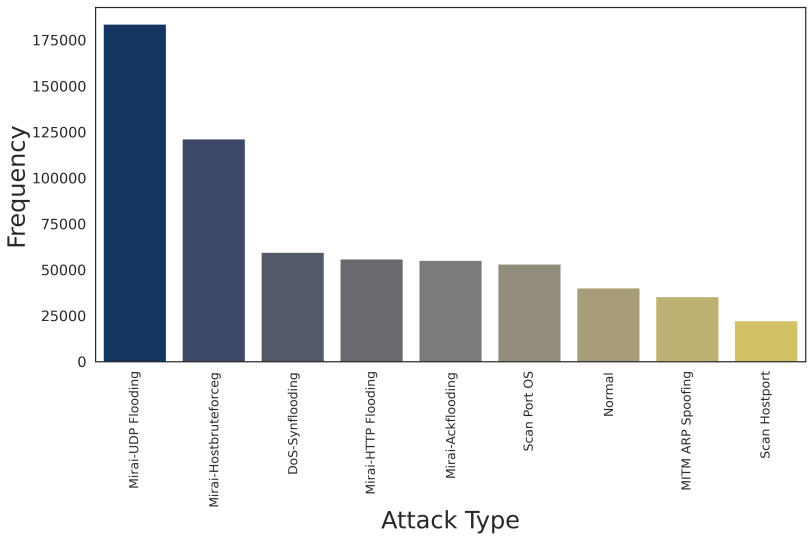

Figure 10. IoTID20 Dataset Distribution.

 doi:10.20944/preprints202503.0706.v1

MQTT-IoT-IDS2020 Dataset

The MQTT-IoT-IDS2020 dataset, designed to mirror the behavior of IoT networks communicating via the MQTT protocol, is characterized by a prominent imbalance in data classes, showcasing a high prevalence of normal activities followed by aggressive scans, UDP scans, and SSH brute-force attacks, among others. The dataset is a fundamental resource for developing and testing IDS solutions that focus on typical IoT-specific threats, such as MQTT brute force attacks, reflecting highly relevant scenarios in real-world applications. This dataset's distribution of attack types, which ranges from the most frequent normal activities to several forms of network probing and intrusion attempts, highlights the variety of attack vectors that an effective IoT IDS must handle. This broad spectrum of attack types is visualized in Figure 11, visually representing the dataset's composition and anomaly detection and classification challenges.

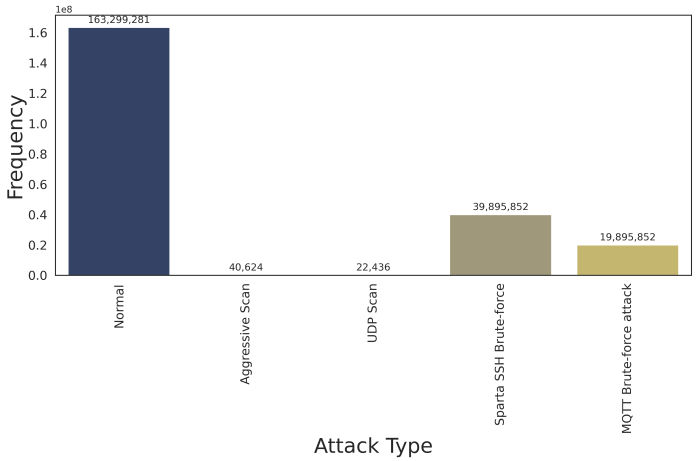

Figure 11. MQTT-IoT-IDS2020 Dataset Distribution.

TON-IoT Dataset

The TON-IoT dataset, created in 2020, is a comprehensive data source for testing IDS, specifically within IoT and IIoT environments. This dataset captures a variety of network traffic scenarios, including both normal activities and a spectrum of attack types such as Backdoor, DDoS, Injection, Normal activities, Password, Ransomware, Scanning, and XSS attacks. The bar plot in Figure 12 displays the distribution of these activities, showing a significant prevalence of normal traffic followed by DDoS and Injection attacks, highlighting the dataset's versatility in representing realistic network behaviors for rigorous IDS testing.

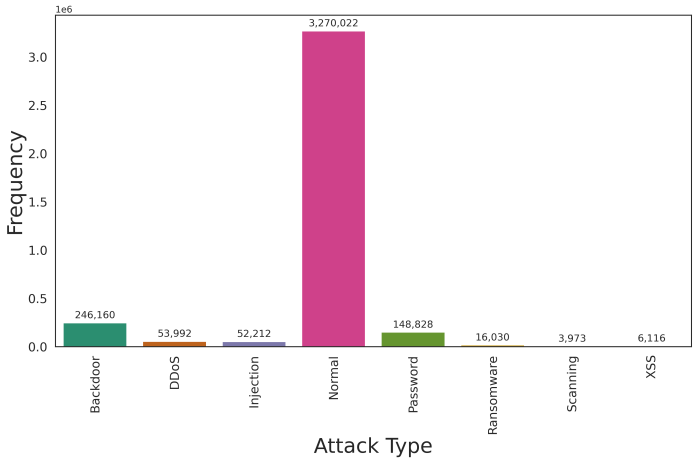

Figure 12. TON-IoT Dataset Distribution.

 doi:10.20944/preprints202503.0706.v1

IoT-23 Dataset Overview

The IoT-23 dataset features a rich collection of IoT-specific attack scenarios and benign samples, making it a critical tool for developing intrusion detection systems tailored to the IoT environment. The dataset primarily consists of labeled network flows, capturing various attack types from various IoT devices. Notable for its volume and variety, IoT-23 encompasses common threats like Mirai and Bashlite alongside less frequent but equally concerning attacks such as partOfAHorizontalPortScan and Okiru. The distribution of these attack types, as illustrated in Figure 13, reveals the predominant focus on probing and malware dissemination, which are pivotal for developing robust detection algorithms. The varied nature of the dataset helps in testing the resilience and effectiveness of IDS solutions across different attack vectors and intensities, providing a comprehensive platform for empirical research and development in IoT security.

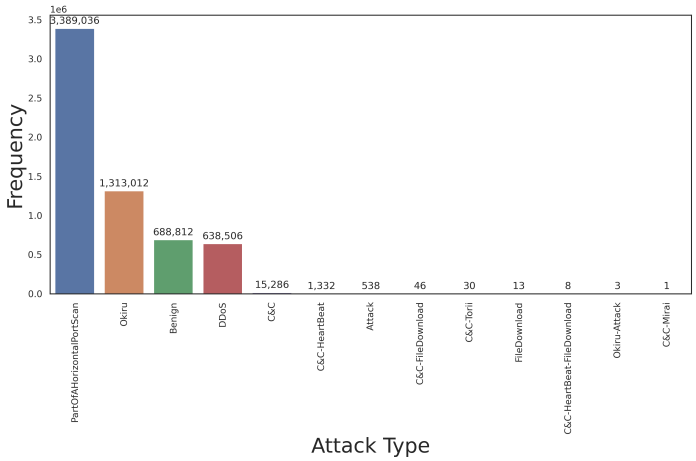

Figure 13. IoT23 Dataset Distribution.

BOT-IoT Dataset Overview

The BoT-IoT dataset, created by Koroniotis et al., is a comprehensive dataset for IoT and network security research, extensively used to evaluate IDS tailored for IoT networks. It notably includes various simulated IoT environments to model attacks, such as DDoS, DoS, reconnaissance, and more exotic threats like theft and backdoor attacks. The distribution of attack types in the BoT-IoT dataset, as shown in Figure 14, illustrates a significant prevalence of DDoS attacks, followed by DoS, which highlights the dataset's focus on these types of network disruptions that are critically impactful in IoT contexts. The dataset also includes examples of reconnaissance and theft, providing a nuanced view into less frequent but equally critical threats and enabling comprehensive testing of IDS capabilities.

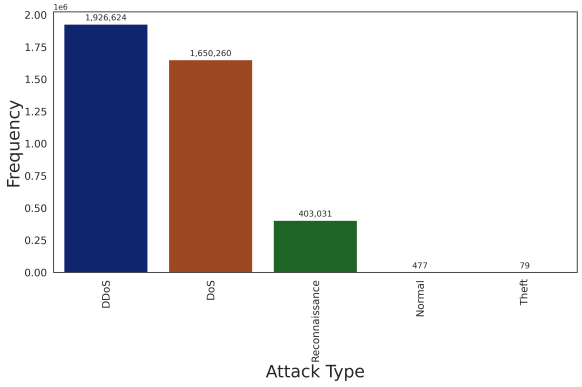

Figure 14. BOT-IoT Dataset Distribution.

 doi:10.20944/preprints202503.0706.v1

N-BAIOT Dataset Overview

The N-BAIOT dataset is specifically designed to address the intricacies of network traffic within IoT environments. It features various attacks, including Mirai UDP, Mirai SYN, Mirai ACK, and several Gafgyt attack types, such as UDP, TCP, and Combo, reflecting a broad spectrum of attack vectors pertinent to IoT security. This diversity aids in developing robust detection models that can handle various intrusion scenarios prevalent in IoT settings. The dataset's comprehensive nature is illustrated in Figure 15, which details the distribution of attack types, offering a visual breakdown of normal versus malicious activities and showcasing the dataset's suitability for testing anomaly detection algorithms.

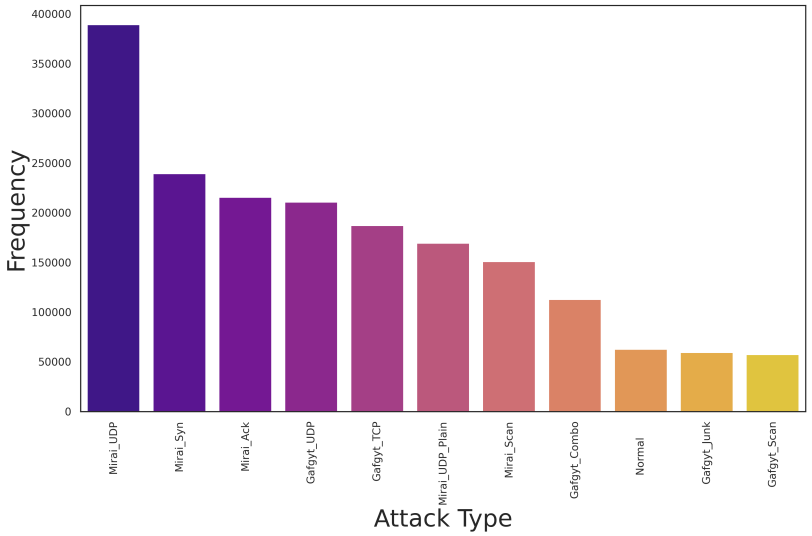

Figure 15. N-BAIOT Dataset Distribution.

6.1.2. Regular IDS Datasets

The range of regular IDS datasets developed over the years is critical in advancing intrusion detection research. Beginning with the well-recognized KDDCup99 dataset [164] from 1999, which has been fundamental in shaping initial intrusion detection systems, the evolution of datasets has seen significant refinement and diversification. The KDDCup99 dataset, despite its age, continues to be a benchmark for new detection algorithms due to its comprehensive set of features and attack types.

Progressing to more contemporary datasets, the CSE-CICDIS2018 [152], Kitsune Network Attack Dataset [154], and CIC-IDS2017 [158] provide more affluent, more modern environments that reflect current network configurations and advanced attack scenarios. For instance, the CSE-CICDIS2018 and CIC-IDS2017 datasets include a variety of attack simulations that help researchers develop and evaluate the efficacy of their IDS models against a spectrum of sophisticated threats. These datasets, enhanced by detailed traffic profiles and diverse attack representations, support the development of systems capable of distinguishing between benign and malicious activities with higher accuracy.

Lastly, the transition into more specialized datasets such as CTU-13 [160] and ISCXIDS2012 [161], which include botnet traffic and more granular attack details, underscores the progression towards datasets that not only challenge the robustness of intrusion detection models but also push the boundaries of what these systems can recognize and mitigate in evolving network environments.

6.2. Balancing IDS Datasets

Balancing techniques in IDS datasets is crucial for addressing the skewed distribution of attack classes as shown by almost all datasets in section 6.1, which can heavily influence the performance of IDS. Various approaches have been developed to ensure balanced datasets, enhancing the predictive

 doi:10.20944/preprints202503.0706.v1

accuracy and reliability of IDS models. These techniques are generally categorized into resampling techniques, algorithmic approaches, and data augmentation methods.

6.2.1. Resampling Techniques

Resampling techniques adjust the class distribution in a dataset by either undersampling the majority class or oversampling the minority class. A popular method for oversampling is the Synthetic Minority Over-sampling Technique (SMOTE), which has been applied to various datasets like BoT-IoT [65], ToN-IoT [109], and several others, including UNSW BoT IoT [116] and NSL-KDD [113]. These studies have demonstrated that SMOTE can significantly improve models' performance metrics by creating synthetic examples rather than over-replicating existing ones.

For undersampling, Random Under Sampling for the Majority Class has been used in diverse datasets like IoTID20 and CIC-IDS-2017 to reduce the number of majority class samples, thereby balancing the dataset more effectively [92]. Hybrid techniques combining both oversampling and undersampling, such as those described in studies using NSL-KDD and CICIDS2017 datasets [165], aim to optimize the balance between class distributions.

6.2.2. Algorithmic Approaches

Algorithmic balancing involves modifying existing data clustering algorithms to manage imbalanced datasets better. The MDPCA utilized in conjunction with datasets like NSL-KDD and UNSW-NB15 [59] enhances the discrimination power of the algorithm by focusing on the density-based clustering of minority class examples. This method helps identify core samples of minority classes crucial for training robust detection models.

6.2.3. Data Augmentation

Data augmentation involves artificially increasing the size and diversity of training datasets by creating modified versions of existing data points. GANs have been used for data augmentation, as they can generate new, synthetic data samples that are realistic yet diverse. Studies utilizing GANs for datasets like UNSW-NB15 and CICIDS2017 [94] have shown that these networks can produce new samples that help train models more effectively distinguish between different network behaviors.

7. Observations, Challenges, and Future Directions

The development of IDS for IoT and broader network environments continues to be a dynamic research area driven by the rapid evolution of technology and emerging cyber threats. This section consolidates critical findings from recent studies, identifies significant challenges current systems face, and outlines promising directions for future research to enhance IDS effectiveness and adaptability.

7.1. Observations

- **Advances in ML and DL**: Integrating ML and DL in IDS has significantly improved detection capabilities, particularly in recognizing complex patterns and anomalies that traditional methods might miss. These approaches enhance the ability to detect sophisticated attacks, including zero-day vulnerabilities, by learning from large datasets.
- **Shift from ML to DL Techniques**: There is a noticeable trend toward employing DL techniques, such as CNN, for both feature extraction and attack detection. With their ability to process raw data and automatically extract features, DL models have shown superior performance in IDS tasks compared to traditional ML methods.
- **Effectiveness of Feature Selection and Data Balancing**: Techniques like feature selection and data balancing have been critical in optimizing IDS performance. Feature selection helps reduce model complexity and computational overhead, which is crucial for deployment in resource-constrained

 doi:10.20944/preprints202503.0706.v1

environments like IoT. Data balancing techniques such as SMOTE and GANs address the class imbalance issue, ensuring the models are trained on diverse attack scenarios.

- **Increased Adoption of Hybrid and Ensemble Approaches**: There has been a noticeable shift towards hybrid and ensemble IDS that combine multiple detection techniques or integrate various data handling strategies. This approach enhances the system's robustness by leveraging the strengths of different methods, improving detection rates and accuracy.
- **Emphasis on Real-Time Detection**: The necessity for real-time threat detection has driven the development of IDS to analyze data streams in real-time, reducing the latency between threat identification and response. This capability is critical for mitigating cyber-attacks impact on IoT systems and networks.

7.2. Challenges

- **Dynamic and Sophisticated Threats**: Cyber threats are becoming more sophisticated, with attackers constantly developing new strategies to bypass security measures. IDS must evolve to detect zero-day attacks and advanced persistent threats, often exhibiting subtle and low-frequency characteristics.
- **Big Data and Scalability Issues**: The exponential growth in IoT devices results in massive data streams that IDS must process, often in real time. Scalability remains a formidable challenge, requiring efficient data management and processing capabilities to handle such volumes without degradation in performance.
- **Resource Consumption and Lightweight IDS**: Developing IDS that are both lightweight and efficient in resource consumption is critical, particularly for IoT environments where devices have limited processing power and battery life. Ensuring that IDS solutions do not overburden the host devices remains a significant challenge.
- **Integration and Standardization**: Integrating IDS seamlessly across diverse platforms and ensuring compatibility with new IoT protocols is increasingly challenging. There is also a lack of standardization in security protocols across devices and networks, complicating the deployment of universal solutions.
- **Maintaining High Accuracy with Imbalanced Data**: Imbalanced datasets, where some attack types are underrepresented, can lead to biased models that perform well on frequent classes but poorly on rare yet critical attack types. This remains a significant hurdle in developing robust IDS.

7.3. Future Directions

- **Development of Adaptive Systems**: Future IDS should focus on adaptability, utilizing online learning or transfer learning to continuously adjust to new data or attack patterns without full retraining cycles. This would help maintain high detection rates over time.
- **Cross-Domain Security Solutions**: There is a critical need for solutions operating across different network layers and environments—from edge devices to the cloud—providing end-to-end security. This includes developing IDS that can handle diverse IoT devices and communication protocols.
- **Leveraging Emerging Technologies**: Integrating innovative technologies like blockchain for decentralized data sharing and AI for predictive threat intelligence could significantly enhance the detection capabilities and reliability of IDS. Blockchain can ensure data integrity and traceability, while AI can predict and preemptively counter emerging threats.
- **Enhanced Anomaly Detection Techniques**: Investing in research that improves anomaly detection algorithms to reduce false positives and adapt to network traffic's evolving behavior can help preemptively identify potential threats. This includes developing more sophisticated behavioral analysis techniques and anomaly detection models.

 doi:10.20944/preprints202503.0706.v1

- **Low-Cost, Energy-Efficient Solutions**: Future research should also focus on developing low-cost, energy-efficient IDS solutions that can be widely adopted, especially in developing regions and applications with significant cost constraints. This includes optimizing algorithms for minimal resource consumption and developing hardware-accelerated IDS.
- **Real-Time Adaptability and Response**: Future IDS systems should incorporate mechanisms for real-time adaptability and automated response to detected threats. This includes integrating ML models that can dynamically update themselves based on new attack patterns and developing automated response strategies to mitigate threats as they are detected.

Expanding on these areas, further research should also explore the development of IDS-specific benchmarks and datasets that reflect real-world environments to evaluate better and compare the efficacy of different IDS approaches. Additionally, more focus on developing low-cost, energy-efficient solutions would facilitate broader adoption, especially in developing regions and applications with significant cost constraints.

8. Conclusion

This paper explores a systematic review of IDS within IoT and general network environments, highlighting the pivotal role that ML and DL play in enhancing cybersecurity measures. As technology evolves and the digital landscape expands, the complexity of cyber threats grows, necessitating more sophisticated and robust IDS solutions. This review has synthesized the recent advancements in IDS, mainly focusing on integrating advanced computational techniques such as feature selection, data balancing, and innovative algorithms to address network threats' dynamic and increasingly sophisticated nature.

Our discussions underscored several critical trends: the shift from traditional ML to more complex DL models, the adoption of hybrid approaches combining various techniques, and the emphasis on developing lightweight and efficient systems suitable for resource-constrained IoT devices. Moreover, the challenges identified through this review, including scalability and the need for real-time threat detection, reflect the ongoing need for research and development in this field. These challenges are not only technological but also extend into the realms of standardization and regulatory compliance, which are crucial for the widespread adoption of these systems.

Looking forward, the future of IDS appears promising yet demanding, with numerous opportunities for breakthroughs in algorithmic efficiency, cross-domain applications, and the use of emerging technologies such as blockchain and AI for enhanced security measures. The insights gathered here should serve as a foundation for ongoing and future research efforts, aiming to develop IDS solutions that are effective, efficient, and adaptable to the ever-changing landscape of cybersecurity threats. Integrating these advanced systems into real-world applications will significantly secure our digital infrastructures against increasingly sophisticated cyber threats.

This paper provides an overview of optimized intrusion detection strategies for IoT environments underpinned by advanced ML and DL techniques. For further details on the specific systems discussed and a complete list of acronyms, readers are directed to Appendix 1 and Appendix 2, respectively.

References

1. Vailshery, L.S. Number of IoT connected devices worldwide 2022 to 2023, with forecasts from 2024 to 2033. https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/, 2022. Accessed: 2022-06-20.
2. Duarte, F. Number of IoT devices. https://explodingtopics.com/blog/number-of-iot-devices, 2024. Accessed: 2024-02-19.
3. Singh, H. IoT deconstructed. https://www.capgemini.com/insights/expert-perspectives/iot-deconstructed/, 2020. Accessed: 2024-02-11.

 doi:10.20944/preprints202503.0706.v1

4. Markets and Markets. IoT Market by Component (Hardware, Software Solutions and Services), Organization Size, Focus Area (Smart Manufacturing, Smart Energy and Utilities,and Smart Retail) and Region–Global Forecasts to 2026. https://www.marketsandmarkets.com/Market-Reports/internet-of-things-market-573.html, 2022. Accessed: 2023-10-18.
5. van Kranenburg, R.; Bassi, A. IoT Challenges. *Communications in Mobile Computing* **2012**, *1*, 9. https://doi.org/10.1186/2192-1121-1-9.
6. Balaji, S.; Nathani, K.; Santhakumar, R. IoT Technology, Applications and Challenges: A Contemporary Survey. *Wireless Personal Communications* **2019**, *108*, 363–388. https://doi.org/10.1007/s11277-019-06407-w.
7. Sobin, C.C. A Survey on Architecture, Protocols and Challenges in IoT. *Wireless Personal Communications* **2020**, *112*, 1383–1429. https://doi.org/10.1007/s11277-020-07108-5.
8. Petrosyan, A. Increase of customers targeted by IoT malware 2022, by sector, 2023. Accessed: 2023-11-22.
9. Petrosyan, A. Global annual number of IoT cyber attacks 2018-2022, 2023. Accessed: 2024-01-05.
10. Staff, C. Guarding Against IoT Attacks: Strategies and Best Practices, 2024. Accessed: 2024-03-14.
11. Infisim. 2024 IoT malware trends: Navigating the evolving landscape of cyber threats, 2024. Accessed: 2024-3-18.
12. Daws, R. Luis Mirabal, Globalstar: Satellite IoT for increased efficiency and cost reduction, 2024. Accessed: 2024-06-15.
13. TelenorIoT. AIoT to Emerge as the Defining Enabler of Digital Transformation, 2024. Accessed: 2024-02-12.
14. Analytics, I. State of IoT Spring 2024, 2024. Accessed: 2024-01-03.
15. Networks, P.A. 2023 Benchmark Report on IoT Security, 2023. Accessed: 2023-12-11.
16. Daws, R. IoT Security Remains a Top Concern for Enterprises in 2024, 2024. Accessed: 2024-02-25.
17. Dhanda, S.S.; Singh, B.; Jindal, P. Lightweight Cryptography: A Solution to Secure IoT. *Wireless Personal Communications* **2020**, *112*, 1947–1980. https://doi.org/10.1007/s11277-020-07134-3.
18. Mousavi, S.K.; Ghaffari, A.; Besharat, S.; Afshari, H. Security of internet of things based on cryptographic algorithms: a survey. *Wireless Networks* **2021**, *27*, 1515–1555. https://doi.org/10.1007/s11276-020-02535-5.
19. Santhosh Kumar, S.V.N.; Selvi, M.; Kannan, A. A Comprehensive Survey on Machine Learning-Based Intrusion Detection Systems for Secure Communication in Internet of Things. *Computational Intelligence and Neuroscience* **2023**, *2023*, 8981988, [https://onlinelibrary.wiley.com/doi/pdf/10.1155/2023/8981988]. https://doi.org/https://doi.org/10.1155/2023/8981988.
20. Aldhaheri, A.; Alwahedi, F.; Ferrag, M.A.; Battah, A. Deep learning for cyber threat detection in IoT networks: A review. *Internet of Things and Cyber-Physical Systems* **2024**, *4*, 110–128. https://doi.org/https://doi.org/10.1016/j.iotcps.2023.09.003.
21. Malhotra, P.; Singh, Y.; Anand, P.; Bangotra, D.K.; Singh, P.K.; Hong, W.C. Internet of Things: Evolution, Concerns and Security Challenges. *Sensors* **2021**, *21*. https://doi.org/10.3390/s21051809.
22. Tawalbeh, L.; Muheidat, F.; Tawalbeh, M.; Quwaider, M. IoT Privacy and Security: Challenges and Solutions. *Applied Sciences* **2020**, *10*. https://doi.org/10.3390/app10124102.
23. Khanam, S.; Ahmedy, I.B.; Idna Idris, M.Y.; Jaward, M.H.; Bin Md Sabri, A.Q. A Survey of Security Challenges, Attacks Taxonomy and Advanced Countermeasures in the Internet of Things. *IEEE Access* **2020**, *8*, 219709–219743. https://doi.org/10.1109/ACCESS.2020.3037359.
24. Bharati, S.; Podder, P. Machine and Deep Learning for IoT Security and Privacy: Applications, Challenges, and Future Directions. *Security and Communication Networks* **2022**, *2022*, 41. https://doi.org/10.1155/2022/8951961.
25. Thakkar, A.; Lohiya, R. A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. *Artificial Intelligence Review* **2022**, *55*, 453–563. https://doi.org/10.1007/s10462-021-10037-9.
26. Faiz, M.N.; Somantri, O.; Supriyono, A.R.; Wirawan, A. Impact of Feature Selection Methods on Machine Learning-based for Detecting DDoS Attacks : Literature Review. *Journal of Informatics and Telecommunication Engineering* **2022**, *5*, 305–314. https://doi.org/10.31289/jite.v5i2.6112.
27. Halbouni, A.; Member, G.S. Machine Learning and Deep Learning Approaches for CyberSecurity : A Review. *IEEE Access* **2022**, *10*, 19572–19585. https://doi.org/10.1109/ACCESS.2022.3151248.
28. Mijwil, M.; Salem, I.E.; Ismaeel, M.M. The Significance of Machine Learning and Deep Learning Techniques in Cybersecurity: A Comprehensive Review. *Iraqi Journal For Computer Science and Mathematics* **2023**, *4*, 87–101. https://doi.org/10.52866/ijcsm.2023.01.01.008.

 doi:10.20944/preprints202503.0706.v1

29. Lyu, Y.; Feng, Y.; Sakurai, K. A Survey on Feature Selection Techniques Based on Filtering Methods for Cyber Attack Detection. *Information* **2023**, *14*. https://doi.org/10.3390/info14030191.
30. Sarker, I.H.; Khan, A.I.; Abushark, Y.B.; Alsolami, F. Internet of Things (IoT) Security Intelligence: A Comprehensive Overview, Machine Learning Solutions and Research Directions. *Mobile Networks and Applications* **2023**, *28*, 296–312. https://doi.org/10.1007/s11036-022-01937-3.
31. Sarker, I.H. Machine Learning for Intelligent Data Analysis and Automation in Cybersecurity: Current and Future Prospects. *Annals of Data Science* **2023**, *10*, 1473–1498. https://doi.org/10.1007/s40745-022-00444-2.
32. Dasgupta, D.; Akhtar, Z.; Sen, S. Machine learning in cybersecurity: a comprehensive survey. *The Journal of Defense Modeling and Simulation* **2022**, *19*, 57–106. https://doi.org/10.1177/1548512920951275.
33. Gyamfi, E.; Jurcut, A. Intrusion Detection in Internet of Things Systems: A Review on Design Approaches Leveraging Multi-Access Edge Computing, Machine Learning, and Datasets. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22103744.
34. Farooq, U.; Tariq, N.; Asim, M.; Baker, T.; Al-Shamma'a, A. Machine learning and the Internet of Things security: Solutions and open challenges. *Journal of Parallel and Distributed Computing* **2022**, *162*, 89–104. https://doi.org/https://doi.org/10.1016/j.jpdc.2022.01.015.
35. Dixit, P.; Silakari, S. Deep Learning Algorithms for Cybersecurity Applications: A Technological and Status Review. *Computer Science Review* **2021**, *39*, 100317. https://doi.org/https://doi.org/10.1016/j.cosrev.2020.100317.
36. Adnan, A.; Muhammed, A.; Abd Ghani, A.A.; Abdullah, A.; Hakim, F. An Intrusion Detection System for the Internet of Things Based on Machine Learning: Review and Challenges. *Symmetry* **2021**, *13*. https://doi.org/10.3390/sym13061011.
37. Lansky, J.; Ali, S.; Mohammadi, M.; Majeed, M.K.; Karim, S.H.T.; Rashidi, S.; Hosseinzadeh, M.; Rahmani, A.M. Deep Learning-Based Intrusion Detection Systems: A Systematic Review. *IEEE Access* **2021**, *9*, 101574–101599. https://doi.org/10.1109/ACCESS.2021.3097247.
38. Ahmad, Z.; Shahid Khan, A.; Wai Shiang, C.; Abdullah, J.; Ahmad, F. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies* **2021**, *32*, e4150, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/ett.4150]. https://doi.org/10.1002/ett.4150.
39. Thakkar, A.; Lohiya, R. A Review on Machine Learning and Deep Learning Perspectives of IDS for IoT: Recent Updates, Security Issues, and Challenges. *Archives of Computational Methods in Engineering* **2021**, *28*, 3211–3243. https://doi.org/10.1007/s11831-020-09496-0.
40. Geetha, R.; Thilagam, T. A Review on the Effectiveness of Machine Learning and Deep Learning Algorithms for Cyber Security. *Archives of Computational Methods in Engineering* **2021**, *28*, 2861–2879. https://doi.org/10.1007/s11831-020-09478-2.
41. Petticrew, M.; Roberts, H. *Systematic Reviews in the Social Sciences: A Practical Guide*; Blackwell Publishing, 2006. https://doi.org/10.1002/9780470754887.
42. Garcia-Teodoro, P.; Diaz-Verdejo, J.; Macia-Fernandez, G.; Vazquez, E. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security* **2014**, *28*, 18–28. https://doi.org/10.1016/j.cose.2008.08.003.
43. Sommer, R.; Paxson, V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In Proceedings of the Proceedings of the IEEE Symposium on Security and Privacy. IEEE, 2010. https://doi.org/10.1109/SP.2010.25.
44. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009, pp. 1–6. https://doi.org/10.1109/CISDA.2009.5356528.
45. Zuech, R.; Khoshgoftaar, T.M.; Wald, R. Intrusion detection and Big Heterogeneous Data: A Survey. *Journal of Big Data* **2015**, *2*. https://doi.org/10.1186/s40537-015-0013-4.
46. Aminanto, E.; Kim, K. Deep Learning in Intrusion Detection System: An Overview **2016**.
47. Ahmed, M.; Mahmood, A.N.; Hu, J. A Survey of Network Anomaly Detection Techniques. *Journal of Network and Computer Applications* **2017**, *60*, 19–31. https://doi.org/10.1016/j.jnca.2015.11.016.
48. Elnakib, O.; Shaaban, E.; Mahmoud, M.; Emara, K. EIDM: deep learning model for IoT intrusion detection systems. *The Journal of Supercomputing* **2023**, *79*, 13241–13261. https://doi.org/10.1007/s11227-023-05197-0.

 doi:10.20944/preprints202503.0706.v1

49. Elrawy, M.F.; Awad, A.I.; Hamed, H.F.A. Intrusion detection systems for IoT-based smart environments: a survey. *Journal of Cloud Computing* **2018**, *7*, 21. https://doi.org/10.1186/s13677-018-0123-6.
50. Bhavsar, M.; Roy, K.; Kelly, J.; Olusola, O. Anomaly-based intrusion detection system for IoT application. *Discover Internet of Things* **2023**, *3*, 5. https://doi.org/10.1007/s43926-023-00034-5.
51. García-Teodoro, P.; Díaz-Verdejo, J.; Maciá-Fernández, G.; Vázquez, E. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security* **2009**, *28*, 18–28. https://doi.org/https://doi.org/10.1016/j.cose.2008.08.003.
52. Zeng, Y.; Qiu, M.; Zhu, D.; Xue, Z.; Xiong, J.; Liu, M. DeepVCM: A Deep Learning Based Intrusion Detection Method in VANET. In Proceedings of the 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS), 2019, pp. 288–293. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2019.00060.
53. Malathi, C.; Padmaja, I.N. Identification of cyber attacks using machine learning in smart IoT networks. *Materials Today: Proceedings* **2023**, *80*, 2518–2523. SI:5 NANO 2021, https://doi.org/https://doi.org/10.1016/j.matpr.2021.06.400.
54. Hassan, M.M.; Gumaei, A.; Alsanad, A.; Alrubaian, M.; Fortino, G. A hybrid deep learning model for efficient intrusion detection in big data environment. *Information Sciences* **2020**, *513*, 386–396. https://doi.org/https://doi.org/10.1016/j.ins.2019.10.069.
55. Xiao, G.; Li, J.; Chen, Y.; Li, K. MalFCS: An effective malware classification framework with automated feature extraction based on deep convolutional neural networks. *Journal of Parallel and Distributed Computing* **2020**, *141*, 49–58. https://doi.org/https://doi.org/10.1016/j.jpdc.2020.03.012.
56. Dahou, A.; Abd Elaziz, M.; Chelloug, S.A.; Awadallah, M.A.; Al-Betar, M.A.; Al-qaness, M.A.A.; Forestiero, A. Intrusion Detection System for IoT Based on Deep Learning and Modified Reptile Search Algorithm. *Computational Intelligence and Neuroscience* **2022**, *2022*, 6473507, [https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/6473507]. https://doi.org/https://doi.org/10.1155/2022/6473507.
57. Liang, X.; Znati, T. A Long Short-Term Memory Enabled Framework for DDoS Detection. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), 2019, pp. 1–6. https://doi.org/10.1109/GLOBECOM38437.2019.9013450.
58. Pantelakis, V.; Bountakas, P.; Farao, A.; Xenakis, C. Adversarial Machine Learning Attacks on Multiclass Classification of IoT Network Traffic. In Proceedings of the Proceedings of the 18th International Conference on Availability, Reliability and Security, New York, NY, USA, 2023; ARES '23. https://doi.org/10.1145/3600160.3605086.
59. Yang, Y.; Zheng, K.; Wu, C.; Niu, X.; Yang, Y. Building an Effective Intrusion Detection System Using the Modified Density Peak Clustering Algorithm and Deep Belief Networks. *Applied Sciences* **2019**, *9*. https://doi.org/10.3390/app9020238.
60. Alajmi, M.; Mengash, H.A.; Alqahtani, H.; Aljameel, S.S.; Hamza, M.A.; Salama, A.S. Automated Threat Detection Using Flamingo Search Algorithm With Optimal Deep Learning on Cyber-Physical System Environment. *IEEE Access* **2023**, *11*, 127669–127678. https://doi.org/10.1109/ACCESS.2023.3332213.
61. Alamro, H.; Mahmood, K.; Aljameel, S.S.; Yafoz, A.; Alsini, R.; Mohamed, A. Modified Red Fox Optimizer With Deep Learning Enabled False Data Injection Attack Detection. *IEEE Access* **2023**, *11*, 79256–79264. https://doi.org/10.1109/ACCESS.2023.3298056.
62. Aljebreen, M.; Alrayes, F.S.; Maray, M.; Aljameel, S.S.; Salama, A.S.; Motwakel, A. Modified Equilibrium Optimization Algorithm With Deep Learning-Based DDoS Attack Classification in 5G Networks. *IEEE Access* **2023**, *11*, 108561–108570. https://doi.org/10.1109/ACCESS.2023.3318176.
63. Almuqren, L.; Alqahtani, H.; Aljameel, S.S.; Salama, A.S.; Yaseen, I.; Alneil, A.A. Hybrid Metaheuristics With Machine Learning Based Botnet Detection in Cloud Assisted Internet of Things Environment. *IEEE Access* **2023**, *11*, 115668–115676. https://doi.org/10.1109/ACCESS.2023.3322369.
64. Latif, S.; Boulila, W.; Koubaa, A.; Zou, Z.; Ahmad, J. DTL-IDS: An optimized Intrusion Detection Framework using Deep Transfer Learning and Genetic Algorithm. *Journal of Network and Computer Applications* **2024**, *221*, 103784. https://doi.org/https://doi.org/10.1016/j.jnca.2023.103784.

 doi:10.20944/preprints202503.0706.v1

65. Soe, Y.N.; Santosa, P.I.; Hartanto, R. DDoS Attack Detection Based on Simple ANN with SMOTE for IoT Environment. In Proceedings of the 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, pp. 1–5. https://doi.org/10.1109/ICIC47613.2019.8985853.
66. Alrashdi, I.; Alqazzaz, A.; Aloufi, E.; Alharthi, R.; Zohdy, M.; Ming, H. AD-IoT: Anomaly Detection of IoT Cyberattacks in Smart City Using Machine Learning. In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), 2019, pp. 0305–0310. https://doi.org/10.1109/CCWC.2019.8666450.
67. Anthi, E.; Williams, L.; Słowińska, M.; Theodorakopoulos, G.; Burnap, P. A Supervised Intrusion Detection System for Smart Home IoT Devices. *IEEE Internet of Things Journal* **2019**, *6*, 9042–9053. https://doi.org/10.1109/JIOT.2019.2926365.
68. Chaudhary, P.; Gupta, B.B. DDoS Detection Framework in Resource Constrained Internet of Things Domain. In Proceedings of the 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), 2019, pp. 675–678. https://doi.org/10.1109/GCCE46687.2019.9015465.
69. Hasan, M.; Islam, M.M.; Zarif, M.I.I.; Hashem, M. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. *Internet of Things* **2019**, *7*, 100059. https://doi.org/https://doi.org/10.1016/j.iot.2019.100059.
70. Ioannou, C.; Vassiliou, V. Classifying Security Attacks in IoT Networks Using Supervised Learning. In Proceedings of the 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), 2019, pp. 652–658. https://doi.org/10.1109/DCOSS.2019.00118.
71. Otoum, S.; Kantarci, B.; Mouftah, H.T. On the Feasibility of Deep Learning in Sensor Network Intrusion Detection. *IEEE Networking Letters* **2019**, *1*, 68–71. https://doi.org/10.1109/LNET.2019.2901792.
72. Qureshi, A.S.; Khan, A.; Shamim, N.; Durad, M.H. Intrusion detection using deep sparse auto-encoder and self-taught learning. *Neural Computing and Applications* **2020**, *32*, 3135–3147. https://doi.org/10.1007/s00521-019-04152-6.
73. Rezvy, S.; Luo, Y.; Petridis, M.; Lasebae, A.; Zebin, T. An efficient deep learning model for intrusion classification and prediction in 5G and IoT networks. In Proceedings of the 2019 53rd Annual Conference on Information Sciences and Systems (CISS), 2019, pp. 1–6. https://doi.org/10.1109/CISS.2019.8693059.
74. Roopak, M.; Tian, G.Y.; Chambers, J.A. Deep Learning Models for Cyber Security in IoT Networks. *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)* **2019**, pp. 0452–0457.
75. Verma, A.; Ranga, V. Machine Learning Based Intrusion Detection Systems for IoT Applications. *Wireless Personal Communications* **2020**, *111*, 2287–2310. https://doi.org/10.1007/s11277-019-06986-8.
76. Zhang, Y.; Li, P.; Wang, X. Intrusion Detection for IoT Based on Improved Genetic Algorithm and Deep Belief Network. *IEEE Access* **2019**, *7*, 31711–31722. https://doi.org/10.1109/ACCESS.2019.2903723.
77. Almiani, M.; AbuGhazleh, A.; Al-Rahayfeh, A.; Atiewi, S.; Razaque, A. Deep recurrent neural network for IoT intrusion detection system. *Simulation Modelling Practice and Theory* **2020**, *101*, 102031. Modeling and Simulation of Fog Computing, https://doi.org/https://doi.org/10.1016/j.simpat.2019.102031.
78. Latif, S.; Zou, Z.; Idrees, Z.; Ahmad, J. A Novel Attack Detection Scheme for the Industrial Internet of Things Using a Lightweight Random Neural Network. *IEEE Access* **2020**, *8*, 89337–89350. https://doi.org/10.1109/ACCESS.2020.2994079.
79. Li, Y.; Xu, Y.; Liu, Z.; Hou, H.; Zheng, Y.; Xin, Y.; Zhao, Y.; Cui, L. Robust detection for network intrusion of industrial IoT based on multi-CNN fusion. *Measurement* **2020**, *154*, 107450. https://doi.org/https://doi.org/10.1016/j.measurement.2019.107450.
80. Roopak, M.; Tian, G.Y.; Chambers, J. An Intrusion Detection System Against DDoS Attacks in IoT Networks. In Proceedings of the 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), 2020, pp. 0562–0567. https://doi.org/10.1109/CCWC47524.2020.9031206.
81. Shafiq, M.; Tian, Z.; Sun, Y.; Du, X.; Guizani, M. Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for internet of things in smart city. *Future Generation Computer Systems* **2020**, *107*, 433–442. https://doi.org/https://doi.org/10.1016/j.future.2020.02.017.
82. Song, H.M.; Woo, J.; Kim, H.K. In-vehicle network intrusion detection using deep convolutional neural network. *Vehicular Communications* **2020**, *21*, 100198. https://doi.org/https://doi.org/10.1016/j.vehcom.2019.100198.
83. Huč, A.; Trček, D. Anomaly Detection in IoT Networks: From Architectures to Machine Learning Transparency. *IEEE Access* **2021**, *9*, 60607–60616. https://doi.org/10.1109/ACCESS.2021.3073785.

84. Kocher, G.; Kumar, G. Analysis of Machine Learning Algorithms with Feature Selection for Intrusion Detection Using UNSW-NB15 Dataset **2021**. *13*, 21–31. https://doi.org/10.2139/ssrn.3784406.
85. Kumar, P.; Gupta, G.P.; Tripathi, R. Toward Design of an Intelligent Cyber Attack Detection System using Hybrid Feature Reduced Approach for IoT Networks. *Arabian Journal for Science and Engineering* **2021**, *46*, 3749–3778. https://doi.org/10.1007/s13369-020-05181-3.
86. Rahman, M.A.; Asyhari, A.T.; Wen, O.W.; Ajra, H.; Ahmed, Y.; Anwar, F. Effective combining of feature selection techniques for machine learning-enabled IoT intrusion detection. *Multimedia Tools and Applications* **2021**, *80*, 31381–31399. https://doi.org/10.1007/s11042-021-10567-y.
87. Ullah, I.; Mahmoud, Q.H. A Framework for Anomaly Detection in IoT Networks Using Conditional Generative Adversarial Networks. *IEEE Access* **2021**, *9*, 165907–165931. https://doi.org/10.1109/ACCESS.2021.3132127.
88. Albulayhi, K.; Abu Al-Haija, Q.; Alsuhibany, S.A.; Jillepalli, A.A.; Ashrafuzzaman, M.; Sheldon, F.T. IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method. *Applied Sciences* **2022**, *12*. https://doi.org/10.3390/app12105015.
89. Alghamdi, M.I. A Hybrid Model for Intrusion Detection in IoT Applications. *Wireless Communications and Mobile Computing* **2022**, *2022*, 4553502, [https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/4553502]. https://doi.org/https://doi.org/10.1155/2022/4553502.
90. Adefemi Alimi, K.O.; Ouahada, K.; Abu-Mahfouz, A.M.; Rimer, S.; Alimi, O.A. Refined LSTM Based Intrusion Detection for Denial-of-Service Attack in Internet of Things. *Journal of Sensor and Actuator Networks* **2022**, *11*. https://doi.org/10.3390/jsan11030032.
91. Banaamah, A.M.; Ahmad, I. Intrusion Detection in IoT Using Deep Learning. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22218417.
92. Dat-Thinh, N.; Xuan-Ninh, H.; Kim-Hung, L. MidSiot: A Multistage Intrusion Detection System for Internet of Things. *Wireless Communications and Mobile Computing* **2022**, *2022*, 9173291, [https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/9173291]. https://doi.org/https://doi.org/10.1155/2022/9173291.
93. EMEC, M.; OZCANHAN, M.H. A Hybrid Deep Learning Approach for Intrusion Detection in IoT Networks. *Advances in Electrical and Computer Engineering* **2022**, *22*, 3–12. https://doi.org/https://doi.org/10.4316/aece.2022.01001.
94. Le, K.H.; Nguyen, M.H.; Tran, T.D.; Tran, N.D. IMIDS: An Intelligent Intrusion Detection System against Cyber Threats in IoT. *Electronics* **2022**, *11*. https://doi.org/10.3390/electronics11040524.
95. Nguyen, X.H.; Nguyen, X.D.; Huynh, H.H.; Le, K.H. Realguard: A Lightweight Network Intrusion Detection System for IoT Gateways. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22020432.
96. Otoum, Y.; Liu, D.; Nayak, A. DL-IDS: a deep learning–based intrusion detection framework for securing IoT. *Transactions on Emerging Telecommunications Technologies* **2022**, *33*, e3803, [https://onlinelibrary.wiley.com/doi/pdf/10.1002/ett.3803]. e3803 ett.3803, https://doi.org/https://doi.org/10.1002/ett.3803.
97. Raghuvanshi, A.; Singh, U.K.; Sajja, G.S.; Pallathadka, H.; Asenso, E.; Kamal, M.; Singh, A.; Phasinam, K. Intrusion Detection Using Machine Learning for Risk Mitigation in IoT-Enabled Smart Irrigation in Smart Farming. *Journal of Food Quality* **2022**, *2022*, 3955514, [https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/3955514]. https://doi.org/https://doi.org/10.1155/2022/3955514.
98. Rani, D.; Gill, N.S.; Gulia, P.; Chatterjee, J.M. An Ensemble-Based Multiclass Classifier for Intrusion Detection Using Internet of Things. *Computational Intelligence and Neuroscience* **2022**, *2022*, 1668676, [https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/1668676]. https://doi.org/https://doi.org/10.1155/2022/1668676.
99. Kayode Saheed, Y.; Idris Abiodun, A.; Misra, S.; Kristiansen Holone, M.; Colomo-Palacios, R. A machine learning-based intrusion detection for detecting internet of things network attacks. *Alexandria Engineering Journal* **2022**, *61*, 9395–9409. https://doi.org/https://doi.org/10.1016/j.aej.2022.02.063.
100. Ullah, S.; Ahmad, J.; Khan, M.A.; Alkhammash, E.H.; Hadjouni, M.; Ghadi, Y.Y.; Saeed, F.; Pitropakis, N. A New Intrusion Detection System for the Internet of Things via Deep Convolutional Neural Network and Feature Engineering. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22103607.

 doi:10.20944/preprints202503.0706.v1

101. Vishwakarma, M.; Kesswani, N. DIDS: A Deep Neural Network based real-time Intrusion detection system for IoT. *Decision Analytics Journal* **2022**, *5*, 100142. https://doi.org/https://doi.org/10.1016/j.dajour.2022.100142.
102. Alrowais, F.; Althahabi, S.; Alotaibi, S.S.; Mohamed, A.; Hamza, M.A.; Marzouk, R. Automated Machine Learning Enabled Cybersecurity Threat Detection in Internet of Things Environment. *Computer Systems Science and Engineering* **2023**, *45*, 687–700. https://doi.org/10.32604/csse.2023.030188.
103. Basati, A.; Faghih, M.M. DFE: efficient IoT network intrusion detection using deep feature extraction. *Neural Computing and Applications* **2022**, *34*, 15175–15195. https://doi.org/10.1007/s00521-021-06826-6.
104. Gupta, L.; Salman, T.; Ghubaish, A.; Unal, D.; Al-Ali, A.K.; Jain, R. Cybersecurity of multi-cloud healthcare systems: A hierarchical deep learning approach. *Applied Soft Computing* **2022**, *118*, 108439. https://doi.org/https://doi.org/10.1016/j.asoc.2022.108439.
105. Kaushik, S.; Bhardwaj, A.; Alomari, A.; Bharany, S.; Alsirhani, A.; Mujib Alshahrani, M. Efficient, Lightweight Cyber Intrusion Detection System for IoT Ecosystems Using MI2G Algorithm. *Computers* **2022**, *11*. https://doi.org/10.3390/computers11100142.
106. Sarwar, A.; Alnajim, A.M.; Marwat, S.N.K.; Ahmed, S.; Alyahya, S.; Khan, W.U. Enhanced Anomaly Detection System for IoT Based on Improved Dynamic SBPSO. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22134926.
107. Aldhyani, T.H.H.; Alkahtani, H. Attacks to Automatous Vehicles: A Deep Learning Algorithm for Cybersecurity. *Sensors* **2022**, *22*. https://doi.org/10.3390/s22010360.
108. Abdulameer, H.; Musa, I.; Al-Sultani, N. Three level intrusion detection system based on conditional generative adversarial network **2023**. *13*, 2240–2258. https://doi.org/10.11591/ijece.v13i2.pp2240-2258.
109. Gad, A.R.; Haggag, M.; Nashat, A.A.; Barakat, T.M. A Distributed Intrusion Detection System using Machine Learning for IoT based on ToN-IoT Dataset. *International Journal of Advanced Computer Science and Applications* **2022**, *13*. https://doi.org/10.14569/IJACSA.2022.0130667.
110. Moody Alhanaya, K.H.A.A.S. Performance Analysis of Intrusion Detection System in the IoT Environment Using Feature Selection Technique. *Intelligent Automation & Soft Computing* **2023**, *36*, 3709–3724. https://doi.org/10.32604/iasc.2023.036856.
111. Alrowais, F.; Eltahir, M.M.; Aljameel, S.S.; Marzouk, R.; Mohammed, G.P.; Salama, A.S. Modeling of Botnet Detection Using Chaotic Binary Pelican Optimization Algorithm With Deep Learning on Internet of Things Environment. *IEEE Access* **2023**, *11*, 130618–130626. https://doi.org/10.1109/ACCESS.2023.3332690.
112. Ashraf, S.N.; Manickam, S.; Zia, S.S.; Abro, A.A.; Obaidat, M.; Uddin, M.; Abdelhaq, M.; Alsaqour, R. IoT empowered smart cybersecurity framework for intrusion detection in internet of drones. *Scientific Reports* **2023**, *13*, 18422. https://doi.org/10.1038/s41598-023-45065-8.
113. Jithish, J.; Alangot, B.; Mahalingam, N.; Yeo, K.S. Distributed Anomaly Detection in Smart Grids: A Federated Learning-Based Approach. *IEEE Access* **2023**, *11*, 7157–7179. https://doi.org/10.1109/ACCESS.2023.3237554.
114. Lai, T.; Farid, F.; Bello, A.; Sabrina, F. Ensemble learning based anomaly detection for IoT cybersecurity via Bayesian hyperparameters sensitivity analysis. *Cybersecurity* **2024**, *7*, 44. https://doi.org/10.1186/s42400-024-00238-4.
115. Lopez, M.M.; Shao, S.; Hariri, S.; Salehi, S. Machine Learning for Intrusion Detection: Stream Classification Guided by Clustering for Sustainable Security in IoT. In Proceedings of the Proceedings of the Great Lakes Symposium on VLSI 2023, New York, NY, USA, 2023; GLSVLSI '23, p. 691–696. https://doi.org/10.1145/3583781.3590271.
116. Sarwar, N.; Bajwa, I.S.; Hussain, M.Z.; Ibrahim, M.; Saleem, K. IoT Network Anomaly Detection in Smart Homes Using Machine Learning. *IEEE Access* **2023**, *11*, 119462–119480. https://doi.org/10.1109/ACCESS.2023.3325929.
117. Shtayat, M.M.; Hasan, M.K.; Sulaiman, R.; Islam, S.; Khan, A.U.R. An Explainable Ensemble Deep Learning Approach for Intrusion Detection in Industrial Internet of Things. *IEEE Access* **2023**, *11*, 115047–115061. https://doi.org/10.1109/ACCESS.2023.3323573.
118. Usoh, M.; Asuquo, P.; Ozuomba, S.; Stephen, B.; Inyang, U. A hybrid machine learning model for detecting cybersecurity threats in IoT applications. *International Journal of Information Technology* **2023**, *15*, 3359–3370. https://doi.org/10.1007/s41870-023-01367-8.

 doi:10.20944/preprints202503.0706.v1

119. Zakariyya, I.; Kalutarage, H.; Al-Kadri, M.O. Towards a robust, effective and resource efficient machine learning technique for IoT security monitoring. *Computers & Security* **2023**, *133*, 103388. https://doi.org/https://doi.org/10.1016/j.cose.2023.103388.
120. Alazab, M.; Awajan, A.; Alazzam, H.; Wedyan, M.; Alshawi, B.; Alturki, R. A Novel IDS with a Dynamic Access Control Algorithm to Detect and Defend Intrusion at IoT Nodes. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24072188.
121. Hazman, C.; Guezzaz, A.; Benkirane, S.; Azrour, M. Enhanced IDS with Deep Learning for IoT-Based Smart Cities Security. *Tsinghua Science and Technology* **2024**, *29*, 929–947. https://doi.org/10.26599/TST.2023.9010033.
122. Karthikeyan, M.; Manimegalai, D.; RajaGopal, K. Firefly algorithm based WSN-IoT security enhancement with machine learning for intrusion detection. *Scientific Reports* **2024**, *14*, 231. https://doi.org/10.1038/s41598-023-50554-x.
123. Yaras, S.; Dener, M. IoT-Based Intrusion Detection System Using New Hybrid Deep Learning Algorithm. *Electronics* **2024**, *13*. https://doi.org/10.3390/electronics13061053.
124. Sun, Z.; An, G.; Yang, Y.; Liu, Y. Optimized machine learning enabled intrusion detection 2 system for internet of medical things. *Franklin Open* **2024**, *6*, 100056. https://doi.org/https://doi.org/10.1016/j.fraope.2023.100056.
125. Morshedi, R.; Matinkhah, S.M.; Sadeghi, M.T. Intrusion Detection for IoT Network Security with Deep learning. *Journal of AI and Data Mining* **2024**, *12*, 37–55. https://doi.org/10.22044/jadm.2023.13539.2471.
126. Kilichev, D.; Turimov, D.; Kim, W. Next–Generation Intrusion Detection for IoT EVCS: Integrating CNN, LSTM, and GRU Models. *Mathematics* **2024**, *12*. https://doi.org/10.3390/math12040571.
127. Ayoob Almotairi, Samer Atawneh, O.A.K.; Khafajah, N.M. Enhancing intrusion detection in IoT networks using machine learning-based feature selection and ensemble models. *Systems Science & Control Engineering* **2024**, *12*, 2321381, [https://doi.org/10.1080/21642583.2024.2321381]. https://doi.org/10.1080/21642583.2024.2321381.
128. Hou, S.; Huang, X. Use of Machine Learning in Detecting Network Security of Edge Computing System. In Proceedings of the 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), 2019, pp. 252–256. https://doi.org/10.1109/ICBDA.2019.8713237.
129. Ren, K.; Zeng, Y.; Zhong, Y.; Sheng, B.; Zhang, Y. MAFSIDS: a reinforcement learning-based intrusion detection model for multi-agent feature selection networks. *Journal of Big Data* **2023**, *10*, 137. https://doi.org/10.1186/s40537-023-00814-4.
130. Alalhareth, M.; Hong, S.C. An Improved Mutual Information Feature Selection Technique for Intrusion Detection Systems in the Internet of Medical Things. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23104971.
131. Woo, J.; Song, J.Y.; Choi, Y.J. Performance Enhancement of Deep Neural Network Using Feature Selection and Preprocessing for Intrusion Detection. *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* **2019**, pp. 415–417.
132. Kasongo, S.M.; Sun, Y. A Deep Long Short-Term Memory based classifier for Wireless Intrusion Detection System. *ICT Express* **2020**, *6*, 98–103. https://doi.org/https://doi.org/10.1016/j.icte.2019.08.004.
133. Wirawan Muhammad, A.; Feresa Mohd Foozy, C.; .; Azhari, A. Machine Learning-Based Distributed Denial of Service Attack Detection on Intrusion Detection System Regarding to Feature Selection. *International Journal Of Artificial Intelligence Research* **2020**, *4*, 1–8. https://doi.org/10.29099/ijair.v4i1.156.
134. Maslan, A.; Mohamad, K.M.B.; Feresa, B.M.F. Feature selection for DDoS detection using classification machine learning techniques. *IAES International Journal of Artificial Intelligence (IJ-AI)* **2020**, *9*, 137–145. https://doi.org/10.11591/ijai.v9.i1.pp137-145.
135. Taher, K.A.; Mohammed Yasin Jisan, B.; Rahman, M.M. Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection. In Proceedings of the 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), 2019, pp. 643–646. https://doi.org/10.1109/ICREST.2019.8644161.
136. Kasongo, S.M.; Sun, Y. A deep learning method with wrapper based feature extraction for wireless intrusion detection system. *Computers & Security* **2020**, *92*, 101752. https://doi.org/https://doi.org/10.1016/j.cose.2020.101752.

 doi:10.20944/preprints202503.0706.v1

137. Khammassi, C.; Krichen, S. A NSGA2-LR wrapper approach for feature selection in network intrusion detection. *Computer Networks* **2020**, *172*, 107183. https://doi.org/https://doi.org/10.1016/j.comnet.2020.107183.
138. Sarker, I.H.; Abushark, Y.B.; Alsolami, F.; Khan, A.I. IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model. *Symmetry* **2020**, *12*. https://doi.org/10.3390/sym12050754.
139. Yu, Y.; Bian, N. An Intrusion Detection Method Using Few-Shot Learning. *IEEE Access* **2020**, *8*, 49730–49740. https://doi.org/10.1109/ACCESS.2020.2980136.
140. Alsirhani, A.; Mujib Alshahrani, M.; Hassan, A.M.; Taloba, A.I.; Abd El-Aziz, R.M.; Samak, A.H. Implementation of African vulture optimization algorithm based on deep learning for cybersecurity intrusion detection. *Alexandria Engineering Journal* **2023**, *79*, 105–115. https://doi.org/https://doi.org/10.1016/j.aej.2023.07.077.
141. Al-Daweri, M.S.; Zainol Ariffin, K.A.; Abdullah, S.; Md. Senan, M.F.E. An Analysis of the KDD99 and UNSW-NB15 Datasets for the Intrusion Detection System. *Symmetry* **2020**, *12*. https://doi.org/10.3390/sym12101666.
142. Saraeian, S.; Golchi, M.M. Application of Deep Learning Technique in an Intrusion Detection System. *International Journal of Computational Intelligence and Applications* **2020**, *19*, 2050016, [https://doi.org/10.1142/S1469026820500169]. https://doi.org/10.1142/S1469026820500169.
143. Zhang, J.; Ling, Y.; Fu, X.; Yang, X.; Xiong, G.; Zhang, R. Model of the intrusion detection system based on the integration of spatial-temporal features. *Computers & Security* **2020**, *89*, 101681. https://doi.org/https://doi.org/10.1016/j.cose.2019.101681.
144. Neto, E.C.P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; Ghorbani, A.A. CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors* **2023**, *23*. https://doi.org/10.3390/s23135941.
145. Ferrag, M.A., Friha, O., Hamouda, D., Maglaras, L., Janicke, H.: Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications: Centralized and Federated Learning. https://doi.org/10.21227/mbc1-1h68.
146. Zolanvari, M.; Teixeira, M.A.; Gupta, L.; Khan, K.M.; Jain, R. WUSTL-IIOT-2021 Dataset for IIoT Cybersecurity Research, 2021.
147. Ullah, I.; Mahmoud, Q.H. A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks. In Proceedings of the Advances in Artificial Intelligence; Goutte, C.; Zhu, X., Eds., Cham, 2020; pp. 508–520.
148. Hindy, H., Tachtatzis, C., Atkinson, R., Bayne, E., Bellekens, X.: MQTT-IoT-IDS2020: MQTT Internet of Things Intrusion Detection Dataset. https://doi.org/10.21227/bhxy-ep04.
149. Moustafa, N. New Generations of Internet of Things Datasets for Cybersecurity Applications based Machine Learning: TON_IoT Datasets. In Proceedings of the Proceedings of the eResearch Australasia Conference, Brisbane, Australia, 2019.
150. Garcia, S., Parmisano, A., Erquiaga, M.J.: IoT-23: A Labeled Dataset with Malicious and Benign IoT Network Traffic. https://doi.org/10.5281/zenodo.4743745.
151. Koroniotis, N.; Moustafa, N.; Sitnikova, E.; Turnbull, B. Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset, 2018, [arXiv:cs.CR/1811.00701].
152. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In Proceedings of the International Conference on Information Systems Security and Privacy, 2018.
153. Aubet, F.; Pahl, M. DS2OS traffic traces, 2018.
154. Mirsky, Y.; Doitshman, T.; Elovici, Y.; Shabtai, A. Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection. In Proceedings of the The Network and Distributed System Security Symposium (NDSS) 2018, 2018.
155. Meidan, Y.; Bohadana, M.; Mathov, Y.; Mirsky, Y.; Shabtai, A.; Breitenbacher, D.; Elovici, Y. N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders. *IEEE Pervasive Computing* **2018**, *17*, 12–22. https://doi.org/10.1109/MPRV.2018.03367731.
156. Ring, M.; Wunderlich, S.; Grüdl, D.; Landes, D.; Hotho, A. Creation of Flow-Based Data Sets for Intrusion Detection. *Journal of Information Warfare* **2017**, *16*, 40–53.

 doi:10.20944/preprints202503.0706.v1

157. Jazi, H.H.; Gonzalez, H.; Stakhanova, N.; Ghorbani, A.A. Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling. *Computer Networks* **2017**, *121*, 25–36. https://doi.org/https://doi.org/10.1016/j.comnet.2017.03.018.
158. Rabbani, M.; Rankothge, W. Enhancing Generalizability in DDoS Attack Detection Systems through Transfer Learning and Ensemble Learning Approaches. Webinar presented at the Canadian Institute for Cybersecurity, 2023. Q&A session with Dr. Windhya Rankothge included.
159. Moustafa, N.; Slay, J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1–6. https://doi.org/10.1109/MilCIS.2015.7348942.
160. García, S.; Grill, M.; Stiborek, J.; Zunino, A. An empirical comparison of botnet detection methods. *Computers & Security* **2014**, *45*, 100–123. https://doi.org/https://doi.org/10.1016/j.cose.2014.05.011.
161. Shiravi, A.; Shiravi, H.; Tavallaee, M.; Ghorbani, A.A. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Computers & Security* **2012**, *31*, 357–374. https://doi.org/https://doi.org/10.1016/j.cose.2011.12.012.
162. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009, pp. 1–6. https://doi.org/10.1109/CISDA.2009.5356528.
163. Takakura, H.: Kyoto 2006+ Dataset. https://doi.org/10.23721/100/1478781 . Data collection period: November 2006 to December 2015, Popularity Rank: 56. https://doi.org/10.23721/100/1478781.
164. Repository, U.M.L. KDD Cup 99 Data Set. Online, 1999.
165. Zhang, H.; Huang, L.; Wu, C.Q.; Li, Z. An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset. *Computer Networks* **2020**, *177*, 107315. https://doi.org/https://doi.org/10.1016/j.comnet.2020.107315.