Journal of
Cyber Security

Tech Science Press

https://doi.org/10.32604/jcs.2026.080477

REVIEW

A Systematic Review of Machine Learning Techniques in Intrusion Detection Systems

Darlington Chigozie Okeke*

Department of Computing and Engineering, University of Gloucestershire, Cheltenham, UK

*Corresponding Author: Darlington Chigozie Okeke. Email: okekechigozied2023@gmail.com

Received: 10 February 2026; Accepted: 29 April 2026; Published: 08 May 2026

ABSTRACT: Background: The evolution of modern networked systems in complexity, volume, and diversity has markedly increased the cyber-attack area. Conventional signature-based intrusion detection systems (IDS) will no longer be adequate for identifying advanced threats. A data-driven, adaptive approach that can identify malicious network activity is provided by machine learning (ML) techniques. This review aims to study, compare, and analyze ML-based approaches in IDS and improve the security defense mechanism. Methods: This systematic review followed the PRISMA 2020 guidelines. ML-based IDS peer-reviewed papers were identified from five scientific databases. Abstracts, full texts, and titles were filtered using predetermined inclusion and exclusion criteria, resulting in a sample of 53 primary studies. Data extraction included the algorithms used, the data used, and the metrics used to evaluate. Findings: The data show that most supervised ML techniques, such as decision trees, support vector machines, ensemble models, and deep learning systems (e.g., convolutional and recurrent neural networks), are predominant. In the majority of studies, high detection accuracy was obtained in controlled experimental settings. Conclusions: ML is a significant addition to intrusion detection, especially for anomaly detection and zero-day attack detection. However, the actual implementation is still limited due to the lack of detailed assessment systems and strict robustness testing. Future studies can focus on reproducibility, the use of diverse datasets, adversarial robustness, and the development of explainable ML methods.

KEYWORDS: Intrusion detection systems; machine learning; cybersecurity; deep learning; network security; adversarial attacks; systematic review

1 Introduction

Computer networks have become a core part of modern societies and organizations due to the rapid growth of digital communications, cloud computing, the Internet of Things (IoT), and data-driven infrastructures. These technologies improve efficiency, connectivity, and scalability, but they also increase the attack surface available to cyber adversaries. As a result, cyberattacks have increased and become more sophisticated [1,2]. Threats, such as malware, denial-of-service attacks, unauthorized access, and advanced persistent threats, compromise the confidentiality of data, the integrity of systems, and the usability of systems in both the public and in-house sectors [3,4].

Intrusion Detection Systems (IDSs) are essential to modern cybersecurity, as they enable continuous monitoring of system and network activities and detect malicious and anomalous behavior. Unlike firewalls and antivirus software, which predominantly use known signatures, IDS can identify known and new attacks, including those involving zero-day exploits [1,2]. This is an important skill because the attackers are evolving

continuously to circumvent fixed defenses. It is necessary to have IDS to guard enterprise networks, cloud, industrial control systems, and IoT ecosystems in which the attacks can initiate cascading failures and cause enormous economic and societal damages [2,5]. The current networked systems are very large and multifaceted, which changes the cybersecurity threat surface. Conventional IDS, particularly fixed signature systems and hard-coded rules, are unable to recognize more advanced, encrypted, or new attacks. AS network traffic increases in volume and diversity, efficient detection is based on the flexible mechanisms able to detect the slightest behavioral variations instead of recognizing explicit attack patterns [6,7].

Machine learning (ML) offers data-driven solutions. ML enables IDS to be trained with distinguishing patterns that are derived from network traffic. ML normalized and malicious behavior of ISI models using statistical, probabilistic, and computational tools, and the detection process was adjusted to account for variations in the network environment and attack strategies. This makes it easy to use anomaly-based detection, which is a massively effective way to detect zero-day attacks and advanced persistent threats that lack any known signature [7,8]. IDS uses both classical and deep learning methods. The classical ML algorithms, i.e., decision trees, support vector machines, k-nearest neighbors, and ensemble-based methods, can be used to detect structured traffic features in an interpretable and efficient manner [6,7]. Deep learning is a generalization of these features, in which a hierarchical representation of raw or least processed traffic data is learned. This reduces the need for handcrafted functionality and augments skills in high-dimensional settings [7,8].

ML can also facilitate the use of IDS in distributed and resource-constrained environments, such as IoT and smart city systems. In these aspects, ML-based IDS can be used with edge or fog computing to deliver scalable, low-latency detection with an acceptable amount of resources consumed [9]. Concurrently, adversarial learning methods may make IDS more robust by training on evasive traffic examples, enabling it to be resistant to adaptive attackers [8]. Regardless of the promise, there are several technical and operational problems that limit the performance of operations with ML-based IDS. It is essential to train data quality and representativeness. ML models involve the utilization of labeled traffic to differentiate between benign and malicious traffic. The real networks always present highly skewed data: the attack instances are infrequent as opposed to normal traffic. This disproportion will favor learning of benign classes, leading to false negatives and missed threats [10,11]. Additionally, features and parameter tuning are never a good choice. With hyperparameters and dimensionality of features, the ML classifiers are likely to have problems, such as the support vector machine and the neural network. Poor tuning or redundancy may negatively affect detection performance and raise computing cost, particularly in high-speed networks [12]. Optimization algorithms can also prove useful, but less so when models are exposed to new attack behavior that is not similar to the one they were trained on.

The generalization of models is a significant issue. ML-based IDS may be quite effective in controlled environments, but cannot easily maintain accuracy in real, more diagnostic environments with diverse traffic distributions, protocols, and device constraints. This problem is sharp in critical infrastructure and industrial control systems, where the functionality of the networks is strongly connected to the process. The deviations may be a normal variation in the operation of their work or maliciousness; there is a problem with the learning models in differentiating the two [13,14]. Even the security of the ML models themselves is complicated. In adversarial attacks, malicious traffic is generated to exploit the model's vulnerabilities and misclassify legitimate traffic. These attacks question the stability of detection based on ML algorithms, particularly on structured data (network flows) [15]. They express questions regarding the durability of IDS, which is based entirely on predictive power. Field deployments are also limited by pragmatic factors of the cost of implementation, the delay of decision implementation, and the understandability of the

application of the machine learning-based IDS in an environment that is limited by resources and mission-critical applications. High-tech models enhance controls on detection, but impair rapid response and lower comprehension of these models on the part of security experts. This, therefore, necessitates the design of a balanced IDS that can be accurate, efficient, and explainable [10,11].

Although research on ML-based IDS has increased, the literature is still quite disproportionate and difficult to extrapolate. Studies are now reflecting the direction in which the field is moving beyond simple classifier benchmarking. For instance, Sajid Farooq et al. [16] proposed an explainable federated learning model for detecting cyber intrusion in smart cities. Their view combines privacy-friendly feature selection with SHAP and LIME explainability, which underscores the increased focus on transparency and privacy, and decentralized deployment in IDS studies. Similarly, Ragab et al. [17] proposed a federated learning system for detecting cyber threats in IoT-enabled smart cities while preserving privacy. They combined feature selection, deep learning, and hyperparameter tuning to scale threat detection in a distributed environment. Ahmed et al. [18] demonstrated that machine learning, deep learning, fuzzy clustering, and signature-based intrusion detection are becoming stronger, indicating a growing trend toward the creation of a hybrid intelligent detection architecture.

Despite these developments, most of the IDS reviews discuss only a limited number of methods, including classical machine learning, deep learning, and attack domains. Most do not have a coherent framework for evaluating algorithms, data, feature-selection schemes, evaluation measures, attack types, and deployment concerns. The systematic review fills that gap through a more comprehensive and informed review of ML-based IDS research. It presents the distinctions and comparisons among traditional machine learning, deep learning, ensemble, hybrid, and federated learning methods, the datasets and feature-engineering methods to train and test models, and the performance measures utilized in studies. The review also covers the widespread methodological limitations of reproducibility, generalization, interpretability, and deployment. In this way, the review provides an analytical framework of the current trends in methodology, gaps in research, and prospects in the field of ML-based intrusion detection.

1.1 Objectives

This is a systematic review whose overall objective is to offer a comprehensive, organized, and critical analysis of the ML methods being used in IDS to enhance the intrusion security defense mechanisms in modern times. As cyber threats keep becoming more complex, sophisticated, and diverse, data-driven and smart IDS are becoming more and more important in detecting malicious activity that conventional rule-based systems do not discriminate in their scope. In this regard, ML has become a focal paradigm in enhancing the accuracy of detection, flexibility, and responsiveness. However, the extensive growth in research in this field of study has led to disjointed conclusions, varied methodologies, and different approaches in the evaluation exercise. Since there is an increasing tendency towards using ML methods in IDS, and the existing literature offers a wide range of approaches, it needs to be synthesized systematically to identify the current tendencies, merits, and shortcomings of the existing methods, and be used to create more efficient and successful solution-based approaches in intrusion detection. The primary objective of the systematic review is to study, compare, and analyze ML-based approaches in IDS and improve the security defense mechanism. Instead of suggesting a new detection model, the review aims to organize the existing evidence, identify prevailing research trends, and place emphasis on the methodological strengths and limitations that have been reported in the literature.

Specifically, the objectives of this review are to:

1. Define and distinguish the ML methods applied in the IDS, which include traditional, deep learning, ensemble, and hybrid methods.

2. Study the most prevalent datasets and feature-engineering algorithms available to serve the purpose of training and testing ML-based IDS.

3. Evaluate the performance indicators—accuracy, precision, recall, F1-score, and false-positive rate, which are utilized to measure the performance of the models of the IDS.

4. Discuss the main challenges and limitations of ML-based IDS.

5. Present the recommendation on future research directions and improvements to create a higher and adaptive ML-based IDS.

1.2 Research Questions (Review Questions)

This systematic review will be guided by the following research questions in accordance with the PRISMA guidelines:

RQ1: *What are the most widely used machine learning methods in intrusion detection systems?*

RQ2: *Which publicly available datasets and feature-engineering/feature-selection techniques are most commonly utilized in machine learning-based intrusion detection system literature?*

RQ3: *How do we evaluate the performance of intrusion detection systems made by machines by using measures and validation strategies?*

RQ4: *What are the main challenges and limitations that are reported in developing, implementing, and deploying intrusion detection systems that are built using machine learning?*

2 Methods

There are different types of literature review, including narrative, systematic, and scoping reviews. Narrative literature reviews provide an in-depth overview of trends, while a systematic literature review provides a systematized analysis with adherence to protocols [19].

2.1 Eligibility Criteria

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, which emphasize transparency, reproducibility, and rigor. When selecting studies, the researcher has set clear eligibility criteria that have led to the development of consistency in study selection and the mitigation of bias. The following were the criteria that studies were supposed to meet to be eligible for inclusion.

- Only peer-reviewed original research studies that investigate exactly ML-based IDS were selected. This will ensure a study does not report summaries of what was already known, but new empirical discoveries.
- Research was required to describe, test, or validate machine-learning methods of intrusion detection and make them relevant to our tasks.
- The review was restricted to English-language articles to preserve consistency and enable the accurate methodological evaluation.
- The review was limited to papers published between 2016 and 2025.

2.2 Exclusion Criteria

Studies were excluded if they failed to satisfy any of the predefined eligibility conditions, as shown in Fig. 1. In particular, the following exclusion criteria were used:

- Non-peer-reviewed (e.g., review articles, survey articles, editorials and book chapters, dissertations and theses, white papers, and conference abstracts that do not contain full articles) were excluded.
- Other works that were not centered on ML-based IDS.
- Language constraints were also used as the exclusion criterion since non-English publications could be difficult to understand, and the methods and findings could be interpreted inconsistently.
- Articles that dated to earlier than the period of time (2016–2025) were eliminated to ensure that the sources relate to existing ML methods and the current state of networks.
- Redundant research or long extensions of studies that had been published were disqualified, and only the full and latest version was included.

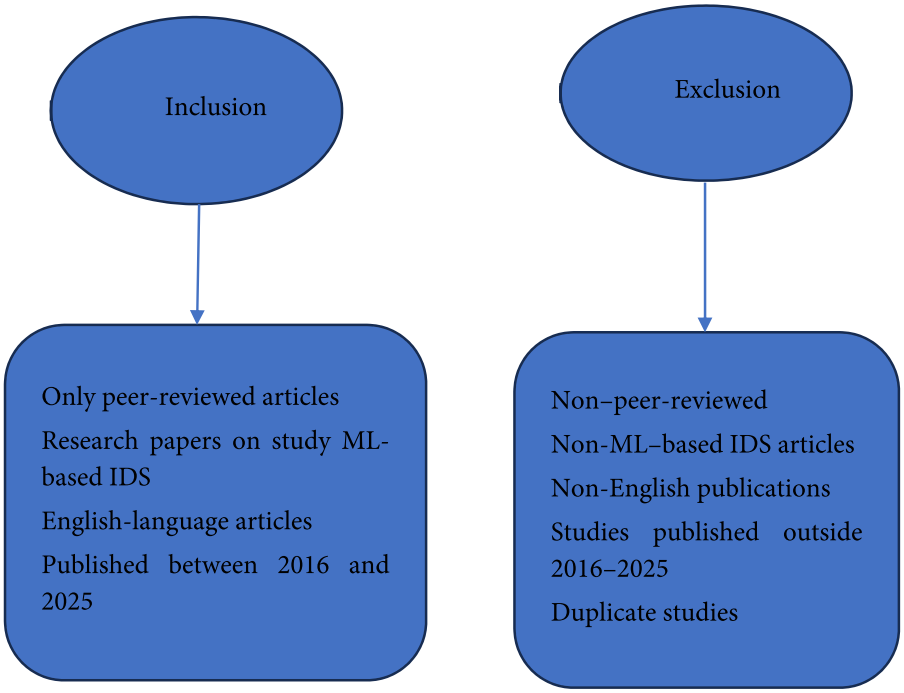

Figure 1: Exclusion and inclusion criteria for the selection of studies. The figure summarizes the eligibility criteria utilized during research selection, which included publication type, topic relevance to ML-based IDS, language, publication era (2016–2025), and duplicate removal.

2.3 Information Sources

A systematic literature search was carried out on five electronic databases, namely IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and ScienceDirect. The selection of these databases was done because they provide depth in computer science and breadth in other associated fields. IEEE Xplore and ACM Digital Library were chosen because they cover peer-reviewed journals and conferences in computer science, cybersecurity, network engineering, and AI. Scopus and Web of Science were selected to broaden coverage, minimize omissions, and retrieve high-quality studies indexed across various disciplines. ScienceDirect was chosen for the availability of full-text journal information in cybersecurity, data science, AI, and networked systems.

The reference lists of the included studies were also manually checked to identify additional relevant articles that the database searches were unable to retrieve. This type of backward citation search reduces the risk of overlooking powerful or most frequently cited work and enhances the completeness of the evidence base. Methodological rigor and empirical validation were the criteria of eligibility, which resulted

in the exclusion of institutional websites, technical reports, and non-peer-reviewed repositories. Using pre-designed search terms, all the databases and other additional sources were searched, and the final search was undertaken in December 2025. The search date was selected to keep the review up to date with recent trends in ML intrusion detection studies.

2.4 Search Strategy

A structured search strategy was developed to identify studies that used ML methods in IDS. The strategy focused on three main search strings: (1) intrusion detection, (2) machine learning and associated tools of artificial intelligence, and (3) cybersecurity or network security. Keywords were: "intrusion detection system", "intrusion detection", intrusion detection IDS, network intrusion detection, machine learning, deep learning, "artificial intelligence", "neural network", "ensemble learning", "cybersecurity", "network security", "cyber attack detection", and "anomaly detection". The search strategy has a Boolean structure as illustrated by Fig. 2.

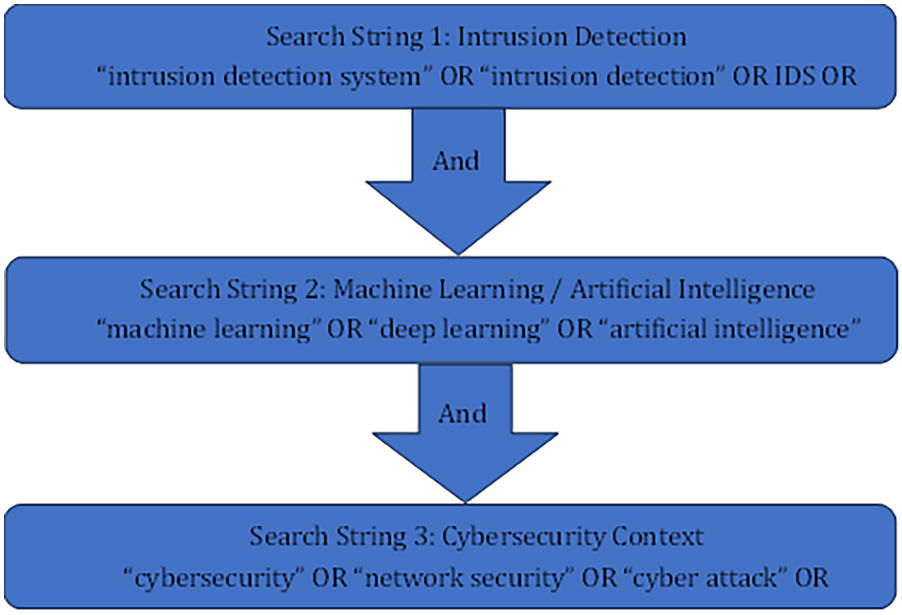

Figure 2: Core Boolean search strategy used to identify studies. It depicts the principal keyword categories used in database searches, including intrusion detection, machine learning approaches, and cybersecurity/network security terms.

The search query was tailored to each database as needed to accommodate variations in the indexing terminologies, interface structure, the use of wildcards, and the search syntax. Title, abstract, and keyword searches were conducted where search capability was possible in the database. Filters were used to narrow the search to publications in the English language that are peer-reviewed, published between 2016 and 2025, and are not geographically limited. Upon retrieval, duplicate records were removed, and the remaining studies were filtered according to the inclusion and exclusion criteria.

2.5 Selection Process

To make studies transparent and reproducible, the PRISMA guidelines are used to select the studies, as illustrated in Fig. 3. The preliminary search in the five scientific databases resulted in 115 records. All the records were added to Zotero, a reference management tool, where duplicates (n = 35) were removed, and

additional records were excluded for other reasons (n = 18) before screening. The 62 entries were preliminarily screened for inclusion and exclusion criteria based on title and abstract. Non-empirical studies (n = 5) whose role in machine-learned intrusion detection was clearly irrelevant, or studies not within the scope of the problem, were filtered out.

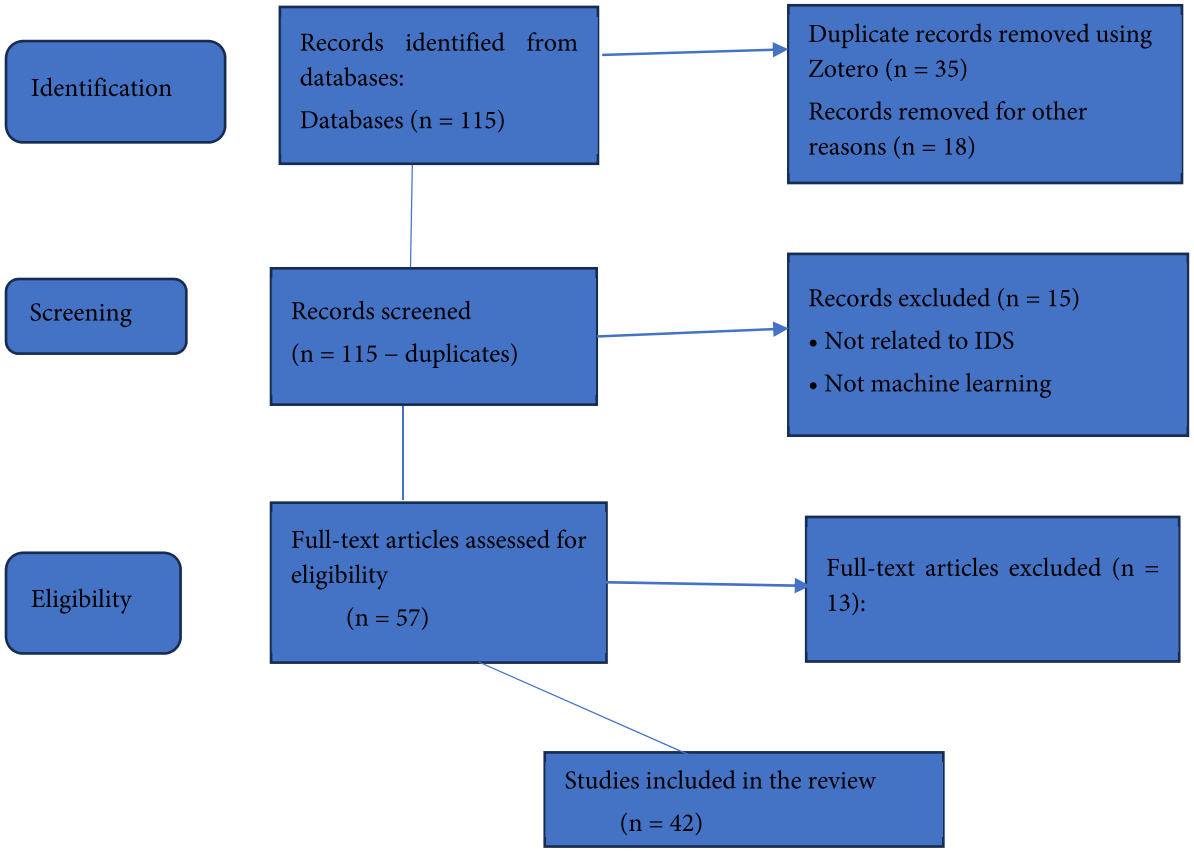

Figure 3: PRISMA flow diagram. It depicts the final set of included research and provides the PRISMA-based filtering process.

The remaining 57 articles were accessed in full text and filtered according to their eligibility. The studies (n = 4) that lacked technical specifics, did not follow a non-machine-learning model, or failed to report evaluation metrics were excluded at this stage. The screening process was done independently. Of the full-text evaluation, 53 studies were found to meet all the eligibility criteria and were included in the final synthesis. The general selection process is described in a PRISMA flow diagram.

2.6 Data Collection Process

A standardized, piloted extraction form was used to extract data, to make it consistent, complete, and rigorous. To minimize biases and ensure minimal errors in data entry, the reviewer used extraction techniques to extract the necessary information for each study separately. Extracted variables had information regarding the publication, machine-learning methods, data, feature-engineering methods, evaluation metrics, and principal findings. Upon extraction, the researcher checked the results of reviewing, resolving discrepancies or ambiguities via discussion and agreement. In case of any need for clarification, the entire

text was re-examined. The data were obtained manually to ensure accuracy and transparency, and were not collected with the help of any automation tools.

2.7 Data Items

The primary data items that were cut out of each of the studies were centered on the outcomes that are directly related to the research questions. The variables that could be extracted were ML type of technique (traditional, deep learning, ensemble, or hybrid), training and evaluation data, feature-engineering or selecting methods, evaluation performance, and validation modes. All the reports of the reported measures of performance in each study, such as accuracy, precision, recall, F1-score, and false-positive rate, regardless of experimental conditions or classification tasks, were collected in case of such availability. Other variables included the year of publication, the field of application, the categories of attacks that were studied, and the limitations of the studies reported. No assumptions were made that some data might be missing or unclear, as these entries were not specified, which would maintain the transparency and prevent bias.

2.8 Study Risk of Bias Assessment

The quality and risk of bias of the included studies were assessed using the Critical Appraisal Skills Program (CASP) checklist. The checklist assisted in contextualizing the rigidity of the methodology used, considering if the objectives were clear, data collection, data analysis, and the soundness of the conclusions. Each study was assessed separately, and only the ones that met the quality criteria were included in the final synthesis.

2.9 Effect Measures

Since the studies were considerably different in designs, datasets, machine-learning models, and evaluation procedures, quantitative measures of effects that can be fitted in a meta-analysis were difficult to use. Hence, effect measures were primarily descriptive and comparative in nature. The synthesis examined performance measures, usually used in the field of intrusion-detection research, including detection accuracy, precision, recall, F1-score, and false-positive rate. The same number reported by the authors for binary and multi-class tasks was used when available. Performance outcomes were interpreted using the datasets and proper feature-engineering techniques, and a validation process. This allowed the qualitative ranking of model performance rather than simply summing the numbers together statistically.

2.10 Synthesis Methods

To structure the synthesis, the studies were categorized by common characteristics of the analysis. The studies were first ranked by the ML method used, including traditional models, deep learning methods, ensembles, and hybrids. They were then clustered in terms of evaluation attributes like the type of the dataset, feature-engineering or feature-selection strategies, and validation. These groupings have been achieved by tabulating the features of each study and aligning them with the research questions.

To prepare the synthesis, the extracted data were prepared by standardizing the language, aligning the descriptions of the model, and summarizing the performance measures that were reported, thereby making each of the studies to be evaluated fairly. In cases of missing and inconsistency of the summary statistics, they were stored in their descriptive form and left as they were. The results were tabulated and organized in structured formats, allowing comparisons of methods. No sensitivity analyses have been prepared, as no statistical pooling was carried out.

2.11 Reporting Bias Assessment

Since it is a systematic review, it addresses reporting bias when including the studies and synthesizing the results, but not when estimating the outcomes. To reduce the chances of publication and reporting bias, a thorough and systematic search of large scientific databases was conducted. Backward search was also searched back through the reference lists of all included studies to identify any possible relevant papers that were not included during the database search. During the data extraction, the differences in the findings of the studies were contrasted to determine whether there was selective reporting, such as unfair use of assessment measures or lack of performance outcomes. In cases where selective or incomplete reporting was identified, it was specifically noted and considered in the narrative synthesis.

2.12 Certainty Assessment

The certainty of the evidence in this systematic review was assessed qualitatively, which corresponded to the purpose and design of the review. Rather than calculating effect sizes, the reliability of the findings was based on the research methodology, the similarity of the findings, and the openness of the experiment. The studies that were given more weight had in common the utilization of popular datasets, clarity in their models and strategies of validation, as well as the ability to report an array of performance indices. The extent of difference between experimental states, data, and evaluation processes in selecting the strength of the evidence, and we package our results in ways that indicate these inadequacies, was also taken into account.

3 Results

3.1 Study Selection

A total of 53 studies that fit the inclusion criteria were included in the synthesis. The selection process was based on the PRISMA 2020 guidelines, which are summarized in the PRISMA flow diagram in Fig. 1. The exclusion of studies was primarily done on the basis that they were not pertinent/relevant to the research questions, they were not empirically evaluated, or they failed to meet the quality criteria.

3.2 Characteristics of Included Studies

The 53 studies span the years 2016 to 2025, and there is an increasing trend in ML-based IDSs. Table 1 indicates that the studies were mostly published in peer-reviewed journals in computer security and artificial intelligence, and some of the high-impact conferences. The majority of the studies involved experimental research, and they tested the performance of IDS using publicly obtainable benchmarks, as shown in Table 1. Minorities used simulation-based or domain-specific datasets, especially in IoT and industrial control systems. Location of evaluation across the studies was very diverse, including variable selection of the dataset, feature-engineering design, and model validation.

Table 1: Characteristics of included studies (n = 53). The table summarizes the publication dates, study designs, machine learning approaches, dataset types, IDS types, and common evaluation techniques of the publications included.

Characteristic	Description
Publication years	2016–2025
Study design	Experimental/comparative evaluation
ML approach	Traditional ML, DL, Ensemble, Hybrid
Dataset type	Benchmark, synthetic, limited real-world

(Continued)

Table 1 (continued)

Characteristic	Description
IDS type	Network-based IDS (NIDS) is dominant.
Evaluation	Accuracy-based metrics with cross-validation

Table 2 indicates that supervised machine learning, deep learning, hybrid, and ensemble-based IDS techniques are given a strong emphasis in the reviewed literature. The most commonly used datasets include NSL-KDD, UNSW-NB15, CICIDS 2017, CICIDS 2018, and Bot-IoT. Various literature also employs application-specific or custom datasets, particularly in the context of IoT, operational technology, and smart cities. The most used evaluation metrics are accuracy, precision, recall, F1-score, and false-positive rate. Nevertheless, the table also shows the increasing interest towards robustness, interpretability, and deployment-oriented validation. Broadly, the literature examined indicates that although most models claim high levels of experimental performance, variations in datasets, attack types, and validation approaches prevent consistent comparisons across studies.

Table 2: Comparative summary of the 53 studies included in the final review of ML-based IDS.

Authors	Method/Focus	Dataset(s)	Metrics	Attack Type(s)	Key Finding
[20]	Review of ML-based cybersecurity/IDS methods	Multiple	Comparative review	General cyber threats	Summarizes trends in ML and DL for IDS
[21]	Review of IDS datasets and ML methods	Multiple IDS datasets	Comparative review	General intrusions	Compares datasets and ML approaches
[22]	Benchmarking ML for anomaly IDS	CICIDS2017	Accuracy, classification metrics	Anomaly intrusions	Reports strong benchmark performance
[23]	Comparative IDS analysis using DT/RF	Multiple	Comparative review	General intrusions	Highlights DT-based comparative findings
[24]	ML for cybersecurity incident detection	Not explicitly stated	Classification metrics	Intrusions/ incidents	Reports effective ML-based detection
[25]	Systematic review of ML/DL IDS	Multiple	Comparative review	General IDS attacks	Shows the growth of hybrid and DL methods
[26]	Review of IDS attacks, methods, and challenges	Multiple	Comparative review	General attacks	Highlights hybrid IDS directions

(Continued)

Table 2 (continued)

Authors	Method/Focus	Dataset(s)	Metrics	Attack Type(s)	Key Finding
[27]	Review of ML for DDoS detection in SDN	Multiple DDoS/SDN datasets	Comparative review	DDoS	ML/DL is promising for SDN DDoS detection
[28]	Review of ML and transfer learning for 5G IDS	Multiple	Comparative review	5G intrusions	Transfer learning is emerging
[29]	Review of ML and feature selection for DDoS	Multiple	Comparative review	DDoS	Emphasizes ensembles and feature selection
[30]	Systematic review of federated IDS	Multiple	Comparative review	General intrusions	Federated IDS supports privacy-preserving deployment
[31]	Customized ML for IDS	NSL-KDD	Recall, F1, FPR, FNR	Benchmark attacks	Reports strong NSL-KDD performance
[32]	Review of ML-based IDS	KDD Cup 99, NSL-KDD	Accuracy, FPR, error rate	General intrusions	Benchmark datasets dominate the field
[33]	Survey of ML and DL IDS methods	NSL-KDD, UNSW-NB15	Accuracy, error rate, FNR	Benchmark attacks	DL is promising, but dataset issues remain
[34]	Overview of ML-based IDS design	NSL-KDD, CICIDS2017	Validation-focused comparison	General attacks	Modern datasets reduce legacy bias
[35]	Review of IDS datasets and challenges	Multiple	Micro metrics, temporal/cross-dataset discussion	General IDS attacks	Stresses dataset limits and generalizability
[36]	Smart DL model for IoT IDS	IoT traffic dataset	Recall, classification metrics	IoT intrusions	Improves IoT intrusion detection
[37]	Least-square SVM-based IDS	Custom network dataset	Precision, recall, FPR	General intrusions	Reports effective ML-based detection

(Continued)

Table 2 (continued)

Authors	Method/Focus	Dataset(s)	Metrics	Attack Type(s)	Key Finding
[38]	Multiclass anomaly detection using ML	CICIDS2017	Accuracy, precision, F1	Network anomalies	Effective multiclass anomaly detection
[39]	Feature selection vs. extraction for IDS	UNSW-NB15, KDDCUP99	Accuracy, runtime	Benchmark attacks	Feature selection often outperforms extraction
[40]	Feature engineering performance analysis	IDS benchmark datasets	Performance comparison	General intrusions	Embedded methods balance performance and efficiency
[41]	Feature engineering for OT IDS	Modbus/TCP, OT traffic	Anomaly detection performance	OT/industrial anomalies	Protocol-specific features improved detection
[42]	Feature engineering in ML/DL IDS	Network IDS datasets	Detection comparison	General intrusions	Entropy/information features reduced redundancy
[43]	Review of feature selection for IDS	Multiple	Comparative review	General intrusions	MI, chi-square, and filter methods are common
[44]	ML-based feature selection for large-scale IDS	Large-scale IDS context	Comparative performance	General intrusions	RF-based and embedded selection performed well
[45]	Hybrid feature selection and stack ensemble	–	Accuracy, precision, recall, F1, FPR/FNR	General intrusions	Multi-metric evaluation improves assessment
[46]	Experimental ML models for IoT IDS	IoT device datasets	Accuracy, minority-class metrics	IoT attacks	Accuracy masked minority-class weaknesses
[47]	IDS evaluation with explainable AI	–	Standard metrics + explainability	General intrusions	XAI improves transparency
[48]	Progressive dataset evaluation of ML IDS	Progressive/temporal dataset	Progressive validation	Evolving intrusions	Performance declines under temporal drift

(Continued)

Table 2 (continued)

Authors	Method/Focus	Dataset(s)	Metrics	Attack Type(s)	Key Finding
[49]	ML classification model for IDS	UNSW-NB15	Precision, recall, accuracy	DoS, DDoS, botnet, ransomware	Supports multiclass attack detection
[50]	DL with hyperparameter optimization	NSL-KDD, CSE-CIC-IDS2018	Classification metrics	DoS, Probe, U2R, R2L, brute force	Improved attack classification
[51]	Data-driven DL IDS for WSNs	NSL-KDD, CICIDS2017, UNSW-NB15, CTU-13	Classification performance	DoS, botnet, infiltration, adversarial	Efficient IDS for WSN settings
[52]	Ensemble IDS for IoT-edge platforms	Bot-IoT, CICIDS2018, NSL-KDD, IoTID20	Classification metrics	Botnet, flooding, reconnais-sance	Supports scalable IoT-edge IDS
[53]	Hierarchical classification for IDS	10 IDS datasets	Empirical hierarchy analysis	Hierarchical attack families	Reduced benign/attack misclassification
[54]	Systematic study of ML and DL IDS	Multiple	Comparative review	General IDS attacks	Data quality and scalability remain major issues
[55]	Comprehensive survey of ML/DL intrusion detection	Multiple	Comparative review	General attacks	Real-time deployment remains difficult
[56]	Review of challenges and future directions	Multiple	Comparative review	General threats	Explainability and zero-day detection are key issues
[57]	Review of DL-based IDS taxonomy and challenges	Multiple	Comparative review	General attacks	DL is effective but computationally expensive
[58]	Review of IDS using ML and DL	Multiple	Comparative review	General intrusions	Adversarial robustness and data scarcity remain concerns
[59]	Review of ML IDS for IoT security/privacy	IoT-focused studies	Comparative review	IoT intrusions	Notes privacy, security, and scalability issues

(Continued)

Table 2 (continued)

Authors	Method/Focus	Dataset(s)	Metrics	Attack Type(s)	Key Finding
[60]	Review of maintainability challenges in ML	ML systems literature	Comparative review	Not attack-specific	Retraining and lifecycle maintenance are significant
[61]	Survey of DL applications in NIDS	Multiple NIDS studies	Comparative review	General NIDS attacks	DL improves NIDS, but dataset realism is limited
[62]	Hybrid PCA-Transformer IDS	Combined dataset system	Accuracy, efficiency	General intrusions	Improved dimensionality reduction and detection
[63]	DL framework for feature extraction/classification	Network IDS datasets	Classification performance	General intrusions	Strong performance with DL feature extraction
[64]	Adversarial attacks on supervised ML NIDS	Adversarial evaluation setting	Accuracy degradation	Evasion, poisoning	Adversarial attacks reduce IDS accuracy
[65]	Robust ensemble adversarial ML framework	IoT traffic	Robustness evaluation	Adversarial IoT attacks	Ensemble defense improved robustness
[66]	Defense for DL-based NIDS against adversarial attacks	DL-based NIDS setting	Robustness /defense evaluation	Adversarial attacks	Adversarial training improved resilience
[67]	Explainable AI for IDS using LIME/SHAP	MLP-based IDS setting	Explainability assessment	General intrusions	LIME and SHAP improved interpretability
[18]	Signature-based IDS using ML/DL with fuzzy clustering	Signature-based IDS datasets	Comparative classification	Signature-based intrusions	ML remained practical; DL captured complex patterns
[16]	Fused ML approach for IDS	Not explicitly stated	Validation accuracy, miss rate	General intrusions	95.18% validation accuracy; 4.82% miss rate
[68]	IoT-SecureFusion for smart cities	IoT sensor and network data	Threat detection evaluation	Smart city/IoT threats	Demonstrates smart-city-oriented detection

(Continued)

Table 2 (continued)

Authors	Method/Focus	Dataset(s)	Metrics	Attack Type(s)	Key Finding
[69]	Hybrid AI framework for large-scale cyber-physical systems	Large-scale data-driven systems	Framework evaluation	Cyber-physical threats	Example of a real-world deployment context
[70]	Cascaded IDS using ML	Enterprise/cloud network context	Deployment-oriented evaluation	Known and zero-day attacks	Suitable for real-network deployment

3.3 Risk of Bias within Individual Studies

Risk of bias was assessed using the CASP framework. The methodological quality was moderate-high; however, there are some sources of bias that were commonly present, as illustrated in Table 3. The majority of studies formulated their goals and reported their machine-learning models and data transparently. Nevertheless, it was easy to find bias in the datasets as most studies were based on older benchmark datasets, such as NSL-KDD, that might not indicate the current attack patterns. An evaluation bias was also present, where the only dataset or a simple validation split was employed (no external validation) in the studies.

Table 3: Risk of bias assessment summary.

Bias Domain	Observed Pattern
Dataset representativeness	Moderate risk
Feature selection transparency	Low–moderate risk
Evaluation methodology	Moderate risk
Reporting completeness	Low risk
Reproducibility	Moderate risk

3.4 Risk of Bias across Studies

In all the research that led to every synthesis, homogeneity of data, lack of uniformity in the use of evaluation measures, and poor reporting of false-positive rates constituted the primary risks of bias. Studies that employed modern data, e.g., CICIDS2017 and the UNSWNB15, tended to have a lesser risk of bias compared to those that have used NSL-KDD exclusively. The patterns of bias varied within the syntheses and revealed that these limitations were systemic and not the problem of selected studies. This impacts the generalizability of the synthesized results.

3.5 Synthesis of Results

3.5.1 Theme 1: ML Technique Categories in IDS (RQ1)

Analysis of 11 studies indicates that ML methods in IDS can be divided into traditional machine learning, deep learning, ensemble, and hybrid approaches. This grouping is a direct response to RQ1 by answering

what the most prevalent types of ML methods are used to conduct research dealing with IDS. Traditional ML techniques are highly eminent in the 11 reviewed studies. The most common algorithms, as shown in Table 4, include Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), k-Nearest neighbors (k-NN), and Naive Bayes (NB), which are commonly applied, especially in IDS implementations based on anomalies [20–22]. Several comparative studies and benchmarking participants also specify RF and SVM as the most adopted methods of classic ML practices in terms of their excellent classification ratings, as well as rather low computation expenses [23,24]. The techniques can be assessed by benchmark datasets, e.g., CICIDS2017 and NSL-KDD, in particular when the focus is on identifying denial-of-service and probing attacks.

Table 4: Classification of ML techniques used in IDS.

Study	Technique Category	Algorithms Reported
Shaukat et al. [20]	Traditional ML/DL	SVM, DT, RF, CNN
Arqane et al. [21]	Traditional ML/DL	SVM, NB, k-NN, Autoencoders
Maseer et al. [22]	Traditional ML/Ensemble	RF, DT, k-NN
Azam et al. [23]	Traditional ML/Ensemble	DT, RF
Raja et al. [24]	Traditional ML/DL	SVM, DT, CNN
Jacob & Sultana Habibullah [25]	Traditional ML/DL	SVM, CNN, LSTM, Autoencoders
Gutiérrez-García et al. [26]	Hybrid	ML + DL
Ali et al. [27]	Traditional ML/DL	SVM, DT, CNN, RNN
Noor et al. [28]	DL/Hybrid	CNN, LSTM, Transfer Learning
Roopesh et al. [29]	Ensemble	RF, Bagging, Boosting
Hernandez-Ramos et al. [30]	Hybrid	Federated Learning + ML

The deep learning methods represent an important segment of the literature reviewed that indicates the growing sophistication of network traffic and threats to cyber-attacks. The most widely used are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks, and autoencoders [25,26]. CNN-based systems are often employed in automated feature removal, and LSTM and RNN systems are used in the detection of temporal relationships within the network traffic information. Autoencoders are most commonly used for unsupervised anomaly detection. Recent systematic reviews emphasize that deep learning is becoming increasingly popular in large-scale, dynamic networks [27,28].

Ensemble learning methods, such as bagging, boosting, and stacking, are also less popular but equally important methods of studying IDS. The most popular ensemble technique mentioned is RF, an ensemble of decision trees [22,23]. The other ensemble methods will aggregate multiple classifiers to enhance detection resilience and minimize false positives [29]. Though not as widespread as the traditional or deep learning, the use of ensemble techniques indicates that researchers continue working on the improvement of the IDS performance by means of model aggregation.

The hybrid IDS methods combine various detection paradigms, such as traditional ML with deep learning and ML-based approaches with rule-based systems. Several sources consider hybrid models as the new solutions meant to capitalize on the complementary advantages of separate methods [25,26]. The federated learning-based IDS models build upon the idea of hybridization even more by allowing decentralized training of models using a decentralized environment [30]. Compared with hybrid approaches in general, more hybrid approaches are being reported in recent literature, suggesting that more adaptive

and scalable IDS architectures are being used. The frequency of the major ML technique categories identified across the reviewed studies is summarized in Table 5.

Table 5: Frequency of ML technique categories across studies.

Technique Category	Number of Studies (n = 11)
Traditional ML	9
Deep Learning	7
Ensemble Methods	4
Hybrid Approaches	4

The performance, computational needs, and deployability of ML methods used in IDS have characteristic performance trade-offs as illustrated in Table 6. DT models are known to be very interpretable and quick to execute, thus can be appropriate in a real-time setting, but also can be easily overfit and cannot be applied to unseen attacks in generalization [29]. SVMs offer excellent classification and achieve good results, particularly in binary intrusion detection tasks, but the major weakness of the methods is their inability to scale to high-dimensional and large velocity network-traffic data [24]. Ensemble learning techniques like Random Forests are much more effective at improving detection accuracy and robustness by averaging multiple learners, but at the cost of increased computational and resource costs [22,25,29].

Table 6: Table summarizes the trade-offs between common ML approaches for IDS, including interpretability, computational cost, and deployment practicality.

Technique	Advantages	Limitations	Computational Complexity	Real-World Challenges
Decision Tree (DT)	Interpretable, fast training/testing	Prone to overfitting	Low	Poor generalization in dynamic environments
Support Vector Machine (SVM)	High accuracy in binary classification	Poor scalability, sensitive to high-dimensional data	High	Not suitable for large-scale real-time IDS
Random Forest (RF)/Ensemble	High accuracy, reduces overfitting	Increased computational cost	Medium–High	Resource-intensive deployment
Naïve Bayes (NB)	Fast, computationally efficient	Lower accuracy due to the independence assumption	Very Low	Limited effectiveness for complex attacks
K-Nearest Neighbors (KNN)	High detection accuracy	Slow prediction time	High (inference)	Not scalable for real-time IDS

(Continued)

Table 6 (continued)

Technique	Advantages	Limitations	Computational Complexity	Real-World Challenges
Deep Learning (CNN, LSTM)	High detection of complex patterns	Requires large datasets, lacks interpretability	Very High	Deployment complexity, explainability issues

Light models like Naive Bayes are highly efficient in computational capability; hence, they are effective when resources are limited, but less effective when detecting accuracy is due to simplifying assumptions regarding feature independence [22,25]. K-Nearest neighbors have good detection capacity with a high inference time, making them inapplicable in real-time intrusion detection systems [28]. Deep-learning models such as CNNs and Long Short-Term Memory (LSTM) networks are good at detecting complex and changing attack patterns. However, they require large-scale labeled datasets, are computationally expensive, and pose several issues of interpretability and deployment issues in real-world settings [21,23,27,28].

Although the reviewed studies assess the IDS performance mainly based on such benchmarks as NSL-KDD and CICIDS, some of the studies suggest the increasing relevance of real-world deployment cases. As shown in Table 7, practical applications of IDS are focused on cloud systems, enterprise networks, and IoT systems, especially in smart-city systems. For example, the current studies have suggested IDS models aimed to protect the smart cities of IoT-gated ones by overseeing sensor networks, cloud communication possibilities, and real-time data flows of the city environment model [69]. Here, these systems combine ML models and distributed network monitoring to identify unusual activity in massive cyber-physical systems [68]. There are hybrid IDS designs that integrate both signature-based and anomaly-based methods, which tend to enhance detection performance while allowing these methods to operate efficiently in real network conditions, where zero-day attacks and volume of traffic are widespread phenomena [70].

Table 7: Examples of real-world IDS deployment contexts.

Deployment Environment	IDS Application	Key Characteristics	Study
Smart Cities/IoT Infrastructure	Anomaly detection for sensor networks and urban infrastructure	Monitors IoT sensors, network traffic, and real-time urban data streams	Lilhore et al. [68]
Large-scale Cyber-Physical Systems	ML-based anomaly detection frameworks	Handles high-volume network traffic and heterogeneous devices	Mia et al. [69]
Enterprise/Cloud Networks	Cascaded IDS combining signature and anomaly detection	Detects known attacks and zero-day intrusions in enterprise networks	Ahamed & Karim [70]

3.5.2 Theme 2: Datasets Used in ML-Based IDS (RQ2)

This theme is relevant to addressing RQ2 by examining the publicly accessible datasets and the feature engineering techniques that are widely utilized in the study of ML-based IDS. The review of the eight studies

indicates the existence of three overwhelming data-related sub-themes, as follows: legacy datasets, up-to-date datasets, the real vs. synthetic datasets, and general tendencies in terms of feature engineering and feature selection.

Older data sets remain in focus in the research of IDS, as indicated in Table 8. Among them, the NSL-KDD dataset is the most commonly used in the examined articles [31–33]. NSL-KDD is also preferred since it is better structured than the original KDD Cup 99 dataset, in terms of eliminating unnecessary records. Some researchers use NSL-KDD to compare both traditional ML and deep learning models concerning their usability and standard format. Nevertheless, other reviews admit intrinsic drawbacks, such as archaic example attack profiles and the lack of coverage of current network traffic [34,35]. Despite the identified weaknesses, NSL-KDD remains widely applicable, particularly for comparative and proof-of-concept research.

Table 8: Commonly used datasets in ML-based IDS research.

Study	Dataset(s) Used	Dataset Type
Zakariah et al. [31]	NSL-KDD	Legacy
Kok et al. [32]	KDD Cup 99, NSL-KDD	Legacy
Liu & Lang [33]	NSL-KDD, UNSW-NB15	Legacy/Modern
Dini et al. [34]	NSL-KDD, CICIDS2017	Legacy/Modern
Tripathy & Behera [35]	Multiple benchmark datasets	Legacy/Modern
Alsubaei [36]	IoT traffic dataset	Modern/Synthetic
Waghmode et al. [37]	Custom network dataset	Synthetic
Gunupusala & Kaila [38]	CICIDS2017	Modern

However, compared to legacy datasets, there is a growing number of reports about modern datasets that include CICIDS2017, UNSW-NB15, and Internet-of-Things-specific datasets. These datasets are meant to represent more closely the modern network settings and attack patterns. The IoT intrusion detection studies are often performed on specialized or domain data to acquire the heterogeneous traffic patterns [36,37]. CICIDS2017 is widely mentioned as having fresh attack scenarios and realistic traffic generation, which is why it is appropriate to test deep learning-based IDS models [38]. However, the literature reviewed shows that modern datasets are not often used as compared to legacy datasets, in part because they are more complex and require preprocessing.

One common trend across the reviewed studies is the lack of sufficient real-world deployment data. The majority of the research, as indicated in Table 9, in the field of IDS is based on synthetic or benchmark datasets, which are produced under controlled settings. Although synthetic datasets are easier to replicate and have comparative evaluation, they are not fully associated with the real-world dynamics of network interactions. Some of them cite a lack of practical operational data as a limitation, noting the disconnection between the test and the field [34,35]. When utilized, real-world datasets are usually proprietary or domain-specific and thus do not present a wide range of availability to comparative research.

Both feature engineering and selection are important elements of IDS when used on different datasets. ML-based IDS normally uses manual dimensionality reduction methods, including correlation-based methods, principal component analysis (PCA), and information gain to select the sample [31,38]. Conversely, deep learning-based models gain more and more strength in the automatic learning of features, which does not require active feature engineering [33,36]. Despite these developments, the reviewed studies state that there

are no standardized feature engineering pipelines, practices of which differ greatly based on the choice of dataset and model architecture.

Table 9: Dataset usage patterns across reviewed studies (n = 8).

Dataset Category	Number of Studies
Legacy datasets (NSL-KDD, KDD Cup 99)	5
Modern datasets (CICIDS2017, UNSW-NB15, IoT datasets)	4
Real-world operational datasets	1
Synthetic/Benchmark datasets	7

Table 10 shows that popular datasets like NSL-KDD, UNSW-NB15, and CICIDS2017 have been very instrumental in the fields of IDS, and they are recognized to have various constraints, which influence the ability to scale their influence in the area of generalizability and practical application. One key issue is that legacy datasets used still (KDD-cup 99 and NSL-kdd), though having better data quality, are still used and rely on the outdated data structure of the traffic and cannot reflect the current attack practices and network configurations, which may be the reason why they are not as effective in identifying contemporary threats as the models [31–35,61]. More recent datasets (UNSW-NB15 and CICIDS2017) are more realistic but still fail to adequately simulate real-world traffic because they lack class imbalance, and benign traffic cases prevail in reality. This bias decreases the capability of models in identifying infrequent yet significant attacks [35,36]. The current datasets are also not representative of dynamic and realistic setups, especially in newly emergent situations like IoT ecosystems, encrypted network traffic, and adaptive adversarial behaviors, which also limit the functional dependability of IDS models in a real deployment context [36,61].

Table 10: Limitations of common IDS datasets.

Dataset	Key Limitations	Impact on IDS Performance
NSL-KDD	Outdated traffic patterns and based on legacy attacks	Poor generalization to modern cyber threats
UNSW-NB15	Class imbalance and limited real-world diversity	Bias toward normal traffic and reduced detection of rare attacks
CICIDS2017	Semi-realistic but still partially synthetic	Limited representation of evolving attack behaviors
KDD99	Redundant and obsolete data	Inflated accuracy and misleading model evaluation

In addition to the limitations of the traditional benchmark datasets, the current research indicates the significant prospects of the development of IDS research by introducing the IoT-oriented and more context-aware datasets. As shown in Table 11, IoT environments exhibit extremely dynamic, heterogeneous, and large-scale traffic patterns, which are vastly unlike the traditional network data and demand IDS frameworks that have the capacity to identify the multigenic behavioral anomalies in a collection of devices and communication protocols [36]. Here, IoT-oriented datasets like Bot-IoT and TON-IoT offer more detailed descriptions of the attack surface in the modern context, which allows for assessing the performance

of the IDS more realistically [46]. Resource utilization across these datasets is also constrained by class imbalance between classes and high false-positive rates, and limited ability to identify minority attacks, because of which a more adaptive and situational assessment model is necessary.

Table 11: Emerging dataset directions in IDS research.

Dataset Type	Key Characteristics	Contribution to IDS Research
IoT-Specific Datasets (e.g., Bot-IoT, TON_IoT)	Heterogeneous device traffic, dynamic communication patterns	Improve realism and reflect modern attack environments
Encrypted/Evolving Traffic (Emerging Need)	Privacy-preserving, adaptive attack behavior	Enhances robustness and real-world applicability of IDS
Imbalanced Real-World Traffic	Dominance of benign traffic over attacks	Highlights the need for advanced detection and evaluation strategies

3.5.3 Theme 3: Feature Engineering and Feature Selection Methods (RQ2 Continuation)

Across the reviewed studies, feature engineering and feature selection were found to be critical in boosting the performance, efficiency, and scalability of ML-based IDS, as illustrated in Table 12. Publicly available benchmark datasets were often used, particularly UNSW15–NB15 and KDDCUP99. There are numerous traffic features in these datasets, and before model training, they necessitated dimensionality reduction [39,40]. Additionally, Various studies have shown that feature engineering can capture discriminative traffic features that cannot be captured by raw packet data. Howe and Papa [41] concluded that the statistical features and protocol-layer characteristics of the Modbus/TCP traffic were found to greatly assist anomaly detection in an operating technology (OT) setup. These findings indicated that application and protocol-specific attributes were more informative than the undefined metrics of volume, and also for sharing results with Random Forest feature importance. Similarly, Ning et al. [42] showed that engineered features grounded on entropic and information-theoretic measures minimized redundancy and maximized the detection performance of both the ML and DL models. This was done by adopting feature selection techniques to counter high dimensionality and the computational load. The extensive use of filtering, wrapper, and embedded techniques was reported in reviews by Hashmi et al. [43] and Al-Jarrah et al. [44]. Some of the most prevalent methods were mutual information (MI), correlation selection, and Random Forest ranking. Embedded techniques, particularly those combined with ensemble learners, offered a reasonable trade-off between performance and runtime efficiency [40,44].

Table 12: Feature engineering and selection methods.

Technique Category	Methods Reported	Representative Studies
Feature extraction	PCA, LDA, Autoencoders	Ngo et al. [39]
Wrapper methods	Sequential Forward/Backward Selection	Zare & Mahmoudi-Nasr [40]

(Continued)

Table 12 (continued)

Technique Category	Methods Reported	Representative Studies
Statistical & protocol features	Traffic statistics, protocol-layer attributes	Howe & Papa [41]
Filter methods	Mutual Information, Chi-square, Correlation	Ning et al. [42]; Hashmi et al. [43]
Embedded methods	RF-based ranking, MI-embedded selection	Zare & Mahmoudi-Nasr [40]; Al-Jarrah et al. [44]

The comparative evaluations in Table 13 showed the advantages of feature selection and feature extraction. Ngo et al. [39] showed that selection tended to provide more detection accuracy and less inference time when medium-sized features are retained in UNSW-NB15. In comparison, extraction methods like PCA and LDA were more successful when there was a very small number of features left, but they required more interpretation. Zare and Mahmoudi-Nasr [40] reported that LDA is effective with deep neural networks to improve the final result, and with numerous other selection methods, indicating that data and models are dependent variables.

Table 13: Reported advantages of feature selection vs. feature extraction.

Aspect	Feature Selection	Feature Extraction
Detection accuracy	Higher with moderate feature sets	Competitive with small feature sets
Runtime efficiency	Lower training and inference time	Higher computational cost
Interpretability	High	Low
Sensitivity to K	High	Low

Further comparison of dimensionality methods of reduction shows that there are significant trade-offs of feature space reduction and accuracy loss, as reflected in Table 14. Some of the feature extraction methods used, such as PCA, are computationally efficient tools that are also applicable in real-time IDS settings [62]. However, they are too linear to allow for the description of intricate attack patterns. By contrast, autoencoders possess superior non-linear feature learning and tend to be more prone to being detected by more advanced attacks. Conversely, filter-based feature selection algorithms (e.g., chi-square and ANOVA) are very fast and scalable, enhance accuracy by eliminating irrelevant features, but do not consider feature interaction [62,63]. According to recent research, a combination of selection and extraction methods can provide the highest balance between detection and computational efficiency [63].

Table 14: Comparison of feature engineering techniques in IDS.

Technique	Type	Accuracy Impact	Computational Efficiency	Strengths	Limitations
PCA	Feature Extraction (Linear)	Maintains high accuracy with reduced features	High (fast computation)	Reduces dimensionality effectively; interpretable	Cannot capture non-linear relationships

(Continued)

Table 14 (continued)

Technique	Type	Accuracy Impact	Computational Efficiency	Strengths	Limitations
Autoencoders	Feature Extraction (Non-linear, DL)	Very high accuracy for complex attacks	Low (computationally expensive)	Captures complex patterns; handles non-linear data	Requires large datasets; high training cost
Filter Methods (Chi-square, ANOVA)	Feature Selection	Improves accuracy by removing noise	Very high (fast and scalable)	Simple, fast, avoids overfitting	Ignores feature interactions; may select suboptimal features
Hybrid (Filter + PCA/DL)	Combined	Highest overall performance	Moderate	Balances accuracy and efficiency	More complex to design and implement

3.5.4 Theme 4: Evaluation Metrics and Validation Strategies (RQ3)

To answer RQ3, across the covered studies, as shown in Tables 9 and 10, metrics of evaluation and validation strategies were reported in general as the key mechanisms to evaluate the performance, robustness, and reliability of IDS. Two dominant dimensions indicate the use of multi-metric performance frameworks and validation strategies that assess generalizability under realistic, changing network conditions. The studies reported a convergent set of performance measures: accuracy, precision, recall, F1-score, and the false-positive rate were the most commonly used [45–47]. For example, Sajid Farooq et al. [16] obtained 95.18% validation accuracy and 4.82% miss rate with their IDS-FMLT model. Although such results look encouraging, it is challenging to compare studies because performance may depend on the selected dataset, preprocessing pipeline, feature set, and attack distribution. These measures were reported in the general cases, combined and not separately, because single-metric analysis is believed to be inadequate for IDS evaluation, particularly when classes are not balanced. Recall was consistently stressed since it constitutes the capacity of an IDS to come up with the appropriate response to the occurrence of malicious traffic [45,48]. Concurrently, the rate of a false positive was often reported to measure the reliability of the operations, given that the high rates of false alarms are a critical obstacle to the deployment of IDS in the real world [45,47].

Other measures, such as true positive rate, false negative rate, and general error rate, were reported to provide more detailed insight into the IDS behavior across attack and benign traffic classes [45,47]. The combined results suggest that the methodology is converged to the comprehensive metric reporting as opposed to depending on the accuracy only. As illustrated in Table 15, one of the key constraints observed in the studies analyzed is that accuracy was applied as one of the evaluation metrics, which could be confusing in IDS scenarios where the benign traffic is vastly disproportionate to the attack incidents [45]. High accuracy can indicate proper classification on the majority of benign samples, but failure to identify minority attackers, as shown in studies of near-perfect accuracy and poor recall of rare attacks, such as backdoor or high false positive rate in binary systems [46]. Recall, F1-score, and false positive/false negative rates are therefore more realistic measures of the effectiveness of an IDS, especially those in areas that require security, as missed

attacks are expensive. Moreover, the conventional methods of static validation (e.g., train-test or k-fold cross-validation) cannot be used to reflect the actual dynamics.

Table 15: Critical evaluation issues in IDS performance assessment.

Aspect	Key Insight	Supporting Studies
Accuracy Limitation	High accuracy can mask poor detection of minority attack classes in imbalanced datasets.	Govindaram et al. [46]; Alsaffar et al. [45]
Importance of Multi-Metrics	Precision, recall, F1-score, FPR, and FNR provide a more complete evaluation of IDS performance.	Alsaffar et al. [45]; Mohale & Obagbuwa [47]
Class Imbalance Impact	Models struggle with minority attack detection despite strong overall performance.	Govindaram et al. [46]
Temporal/Progressive Validation	Using future or different datasets reveals performance degradation over time.	Chua & Salam [48]
Adversarial Validation	IDS accuracy significantly drops under evasion and poisoning attacks	Alshahrani et al. [64]

Recent studies propose stronger evaluation methods, such as progressive or cross-dataset validation to emulate the temporal dynamics of network traffic, and adversarial validation to assess models' resilience to evasion and poisoning attacks [64]. These newly emerging trends, including online and constant learning, allow IDS models to employ emerging attack trends, which present a more realistic and practical evaluation framework to implement in dynamic environments, as shown in Table 16. Besides the choice of metrics, the validation strategies have become one of the most important distinctions between studies. Even though they were commonly employed as standard train-test splits, various studies reported weaknesses in their long-term characterization of IDS affectivity [48]. This trend indicates that the effectiveness of the IDS can be overestimated when using static validation methods, underscoring the importance of using temporal and cross-dataset validation.

Table 16: Performance evaluation metrics used in ML-based IDS studies.

Evaluation Metric	Purpose in IDS Evaluation	Studies
Accuracy	Measures the overall correctness of IDS predictions	Kok et al. [32]; Liu & Lang [33]; Gunupusala & Kaila [38]
Precision	Measures the proportion of detected attacks that are truly malicious	Waghmode et al. [37]; Gunupusala & Kaila [38]
Recall (Detection Rate/TPR)	Measures the ability to detect malicious traffic correctly	Zakariah et al. [31]; Waghmode et al. [37]; Alsubaei [36]

(Continued)

Table 16 (continued)

Evaluation Metric	Purpose in IDS Evaluation	Studies
F1-score	Harmonic mean of precision and recall	Gunupusala & Kaila [38]; Zakariah et al. [31]
False Positive Rate (FPR)	Proportion of benign traffic misclassified as malicious	Kok et al. [32]; Zakariah et al. [31]; Waghmode et al. [37]
False Negative Rate (FNR)	Proportion of attacks missed by the IDS	Liu & Lang [33]; Zakariah et al. [31]
Error Rate	Overall misclassification rate	Kok et al. [32]; Liu & Lang [33]
Macro-averaged Recall/F1	Averages performance equally across all classes	Gunupusala & Kaila [38]; Zakariah et al. [31]
Micro-averaged Metrics	Aggregates performance weighted by class size	Gunupusala & Kaila [38]; Tripathy & Behera [35]

Also commonly reported were parallel evaluation through a binary classification scheme and multi-class classification schemes. Scenarios that used binary classification tended to achieve a higher recall, whereas multi-class analysis scenarios showed difficulties in identifying the minority attack types, particularly during imbalanced training [46]. One of the new aspects of evaluation that was found throughout the studies was the adoption of explainable artificial intelligence (XAI) methods in the assessment of IDS [47]. These researchers employed explainability alongside conventional performance metrics to justify model decisions and identify significant features and decisions. Instead of substituting the conventional evaluation measures, validation through XAI increased transparency and trustworthiness, instead of replacing them. The summary of the key validation strategies found in the studies reviewed and their implications on the robustness and the generalizability of IDS are presented in Table 17.

Table 17: Validation strategies used to assess IDS robustness and generalizability.

Validation Strategy	Observed Implications	Studies
Train–Test Split	Common risks of overestimating IDS performance	Kok et al. [32]; Liu & Lang [33]
k-Fold Cross-Validation	Improves robustness, but is still limited for temporal drift	Gunupusala & Kaila [38]; Dini et al. [34]
Temporal Validation	Reveals performance decay under evolving attack patterns	Dini et al. [34]; Tripathy & Behera [35]
Cross-Dataset Validation	Exposes generalization challenges and dataset bias	Tripathy & Behera [35]; Liu & Lang [33]
Binary Classification Evaluation	Often yields higher recall and optimistic results.	Kok et al. [32]; Gunupusala & Kaila [38]

(Continued)

Table 17 (continued)

Validation Strategy	Observed Implications	Studies
Multi-Class Classification Evaluation	Highlights the difficulty in detecting minority attack classes	Gunupusala & Kaila [38]; Zakariah et al. [31]
Class-Imbalance–Aware Validation	Improves minority attack detection reliability	Zakariah et al. [31]; Tripathy & Behera [35]
IoT-Specific Dataset Validation	Demonstrates reduced generalizability from traditional datasets	Alsubaei [36]; Tripathy & Behera [35]
Explainability-Assisted Validation	Enhances trust and transparency alongside metrics	Dini et al. [34]; Alsubaei [36]

3.5.5 Theme 5: Attack Categories Addressed

In the studies reviewed, ML-based IDSs focused on a wide scope of attack types. The results were usually based on benchmark datasets where attack taxonomies were predefined, and as such, these taxonomies influenced the range of attacks studied. One overall trend is the shift toward multi-category attack detection rather than simple normal-vs.-attack classification. Indicatively, Agarwal et al.'s [49] study used normal machine-learning classifiers to identify multiple forms of attacks with varying precision and recall and ascertained that it is possible to detect the cases of a multi-class attack in an enterprise network. Also, multi-class classification results were obtained using datasets such as NSL-KDD, CICIDS2017/2018, UNSW-NB15, and CTU-13 [50,51]. As illustrated in Table 18, the attack categories discussed in these datasets include DoS, probing/scanning, user-to-root (U2R), remote-to-local (R2L), brute-force, botnet, and infiltration attacks.

Table 18: Attack categories addressed across reviewed studies.

Study	Dataset(s) Used	Attack Categories Addressed
Agarwal et al. [49]	UNSW-NB15	DoS, DDoS, Botnet, Ransomware-related attacks
Kunang et al. [50]	NSL-KDD, CSE-CIC-IDS2018	DoS, Probe, U2R, R2L, brute-force
Sinha et al. [51]	NSL-KDD, CICIDS2017, UNSW-NB15, CTU-13	DoS, botnet, infiltration, adversarial attacks
Aldaej et al. [52]	Bot-IoT, CICIDS2018, NSL-KDD, IoTID20	Botnet, flooding, reconnaissance, IoT-specific intrusions
Uddin et al. [53]	10 IDS datasets	Hierarchical attack families and subtypes

Research on edge environments and IoT has increased coverage of attacks, as indicated in Table 19. The weaknesses of this test are that the Aldaej et al. [52] study supports the classification that ensemble methods are inherently scalable to datasets with varying attack distributions, an observation that both underscores the increasing attention to categories of concern to IoT. In contrast to flat multi-class methods, Uddin et al.'s [53]

study demonstrates that hierarchical classification minimizes the error of misclassifying an attack as benign, even in the case of power confusion across subtypes. This highlights the significance of attack taxonomy design as a methodological factor that conditions the performance of the IDS, particularly in a high-risk environment where false negatives are severe.

Table 19: Observed patterns in attack category handling.

Pattern	Evidence from Studies
Emphasis on multi-class detection	Agarwal et al. [49]; Kunang et al. [50]
Inclusion of IoT-specific attacks	Aldaej et al. [52]; Uddin et al. [53]
Hierarchical attack modeling	Uddin et al. [53]
Dataset-driven attack taxonomy	All reviewed studies [49–53]

In addition to traditional attack categories, recent research reports that adversarial attacks have become increasingly prominent threats to ML-based IDS, as illustrated in Table 20. These attacks are normally divided into evasion attacks, whereby a well-designed input will result in misclassification during testing, and poisoning attacks, which will distort training data to make the model perform poorly [64]. Evidential research demonstrates that evasion attacks can reduce detection accuracy and that poisoning attacks interfere with the learning process, leading to more false negative reports, thus undermining the reliability of IDS [64]. Several adversarial robustness strategies have been suggested to reduce such risks. Adversarial training that includes adversarial examples in the training of the models has proved to enhance stable training and retention of detection. Also, powerful feature selection algorithms (e.g., combining feature extraction and selection) reduce sensitivity to adversarial noise and are better predictors. Moreover, ensemble defenses are more robust because they pool together a large number of models, which in turn adds challenges to attackers to take advantage of one point of vulnerability and the overall stability of detection in adversarial conditions [65,66].

Table 20: Adversarial attacks and defense techniques in ML-based IDS. The table lists widely reported defenses, such as adversarial training, strong feature selection, and ensemble-based protection methods, as well as evasion and poisoning attacks.

Category	Type	Description	Impact on IDS	Defence Techniques	Supporting Studies
Adversarial Attack	Evasion Attack	Manipulates input data during testing to cause misclassification	Reduces detection accuracy and increases false negatives	Adversarial training Ensemble defences	Alshahrani et al. [64]; Alkadi et al. [65]
Adversarial Attack	Poisoning Attack	Injects malicious samples into training data to corrupt model learning	Degrades model performance and disrupts the training process	Robust feature selection Data sanitization	Alshahrani et al. [64]

(Continued)

Table 20 (continued)

Category	Type	Description	Impact on IDS	Defence Techniques	Supporting Studies
Defence Technique	Adversarial Training	Incorporates adversarial examples during training	Improves robustness while maintaining accuracy	–	Alkadi et al. [65]; Barik & Misra [66]
Defence Technique	Robust Feature Selection	Removes noisy or adversarial sensitive features	Enhances generalization and reduces overfitting	–	Barik & Misra [66]
Defence Technique	Ensemble Defences	Combines multiple models for decision-making	Increases resilience; reduces single-point failure	–	Alkadi et al. [65]

3.5.6 Theme 6: Challenges and Limitations of ML–Based IDS (RQ4)

Several challenges and limitations, as shown in Table 21, have been reported across studies on the creation and application of ML-based IDS, and have been found in its deployment. A major challenge is on data availability, quality, and representativeness. Numerous studies cite the intensive use of benchmark data, including NSLKDD, CICIDS, and UNSWNB15, which can be helpful for comparison, but they might not resemble real-life traffic distributions [54–56]. The Imbalance of the dataset mentioned in the studies, as the instances of attack are underrepresented, or there is a high concentration on the definite classes. Such a bias results in perverted accuracy rates [57,58]. Further, the absence of egregious labeled traffic of operational networks is another requisite bottleneck to supervised and deep-learning models, which depend on annotated information [57,58].

Table 21: Key challenges and limitations identified in ML-based IDS.

Challenge Category	Reported Issues	Studies
Data limitations	Imbalanced datasets, lack of real-world labeled data	Ahmad et al. [54]; Aleesa et al. [57]; Jakotiya et al. [58]
Computational complexity	High training cost, scalability issues	Ahmad et al. [54]; Krishnamoorthy & Sistla [59]
Interpretability	Black-box nature of DL models	Thakkar & Lohiya [56]; Jakotiya et al. [58]
Evolving attacks	Poor generalization to novel attacks	Ahmad et al. [54]; Surakhi et al. [55]
Adversarial robustness	Susceptibility to adversarial manipulation	Jakotiya et al. [58]; Krishnamoorthy & Sistla [59]

(Continued)

Table 21 (continued)

Challenge Category	Reported Issues	Studies
Maintainability	Model drift, retraining complexity	Shivashankar & Martini [60]

From a computational perspective, studies have shown that processing overhead is a significant concern for scalability. The studies reported that deep-learning models are as resource-intensive and need geometry, training time to operate, and they are also difficult to run in large-scale or memory-constrained settings [54,59]. Aleesa et al. [57] and Surakhi et al. [55] reported challenges further exacerbated in a high-speed environment and the IoT, where enormous traffic volumes need to be processed nearly in real-time.

Another challenge reported was the lack of model interpretability and explainability. Classical ML algorithms, such as decision trees and RF, have certain levels of transparency, whereas deep-learning IDS algorithms are frequently discussed as black boxes and they restrict confidence and embrace in the operational security settings [56,58]. Recent comparative studies also signify this difference. Lilhore et al. [68] found that classic models such as SVM and Random Forest were the most promising when applied to real-world IDS use due to their adaptability and influence, but LSTM and ANN were better at intrusion pattern complexities and changes. Additionally, the challenge of identifying new and emerging attacks was also raised in several studies. ML-based IDS will be susceptible to a zero-day attack and concept drift due to variation in network behavior, even though it becomes more accurate in recognizing known patterns [54,56]. Input manipulation to avoid detection in adversarial attacks is also an up-and-coming issue [58,59]. Finally, the studies highlighted an added complexity in the long-term maintenance and lifecycle. Shivashankar and Martini [60] showed ML-based systems present maintainability problems that do not exist in rule-based IDS, including data engineering/model retraining/deployment pipeline dependencies. These issues influence the reliability of a system in the long run and add an extra degree of complexity. Table 22 illustrates the limitations of ML- and DL-based IDSs.

Table 22: Comparison of limitations between ML and DL-based IDS.

Aspect	Traditional ML-Based IDS	Deep Learning–Based IDS
Data dependency	Requires labeled features	Requires large labeled datasets
Interpretability	Moderate to high	Low
Computational cost	Relatively low	High
Novel attack detection	Limited	Improved but inconsistent
Deployment readiness	Higher	Still evolving

To address the explainability limitations of deep-learning IDS, recent studies propose the inclusion of XAI frameworks. Specifically, they apply model-agnostic methods such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive exPlanations). These approaches provide localized explanations of individual predictions by demonstrating the influence of each input attribute on the decision [67]. What will become more transparent and trustworthy is the result. Perturbation analysis can also be used with them to validate the explanations and quantify the feature influence, which contributes to the continued improvement of the model.

3.6 Risk of Reporting Biases

The risk of reporting bias, as assessed, showed that there was a moderate risk because of selective reporting of outcomes and underreporting of the negative ones. Asymmetry of the funnel plot could not be formally tested due to differences in outcome measures.

3.7 Certainty of Evidence

A modified GRADE-based approach, applied in ML studies, assessed the overall level of certainty of the evidence as moderate. The rating of certainty was reduced due to a lack of consistency in evaluation techniques and the likelihood of bias due to the selection of datasets. Nevertheless, the consistency of results across numerous studies supports confidence in the key conclusions.

4 Discussion

4.1 Interpretation of Findings

This systematic review has reviewed the prevalent methods, data collections, feature designs, testing approaches, coverage of attacks, and the challenges that are reported using the ML-based IDS. The findings indicate that the research in the field of IDS is highly dependent on traditional ML and deep learning, it employs benchmark datasets on a large scale, and is characterized by growing methodological complexity in feature engineering and assessment. There are permanent issues that involve data realism, scalability, interpretability, and long-term implementation, which have not been resolved. In all six themes, a common theme appears: a technique of ML can be used to identify things in a dataset better than it can detect using rule-based IDS, but the usefulness of the technique is constrained by the quality of the data it works with, its features, and its operational requirements. These results support previous findings and reveal discrepancies between theoretical work and practical implementation.

The prevalence of classic algorithms, such as RF, Support Vector Machines, and Decision Trees, is comparable to previous research, which highlights their ability to equalize both the accuracy and algorithm speed [71,72]. The use of deep learning algorithms, particularly CNNs, RNNs, and LSTMs, is on the rise. This supports the previous claims that deep models are better in a complex and heavy traffic area [73]. Consistent with Mohammed and Talib [74], the use of deep learning remains limited by the resource requirements and explanations. The hybrid approaches and the ensemble approaches are emerging; they have not become dominant, meaning that IDS research is in a transition state on its way to a standardized architecture. One of the contributions is the synthesis of practices for using the dataset and for feature engineering. The popularity of legacy datasets, such as NSLKDD, is explained by the fact that they are available and can be benchmarked [72,75]. However, such datasets are not able to detect the current tendencies of traffic and attack patterns, and this point is justified by this review.

The findings also show that feature engineering and feature selection are mere dubbing processes in the IDS performance. This aligns with Torabi et al. [76] and Emirmahmutoglu and Atay [77], who believe that dimensionality reduction has a direct effect on accuracy and efficiency. According to the reviewed studies, there is a tendency to use embedded and heuristic-based selection techniques, which, in a high-dimensional environment, are more effective than manual or filter-only techniques. The reason is that the feature pipelines of a single dataset are not standardized, and it is more difficult to cross-study studies, which confirms the thesis that unified benchmarking frameworks are required [78]. There has been an increasing consensus regarding multi-metric evaluation, which is confirmed by the review. Recall, F1-score, and false-positive rate are gradually replacing accuracy. This aligns with previous critiques that accuracy alone is insufficient, particularly when there is disparity between classes [79].

The reported performance is highly influenced by validation strategies. Empirical results in any study involving temporal or cross-dataset validation typically exhibit decreased performance, indicating prior fears that overfitting or other weak generalizability [80]. These findings argue that a substantial proportion of performance benefits were reported in the IDS using optimistic tests in the interest of a given position under conditions of fixed test. Additionally, explainability is an aspect of assessment that covers related directions in the area of ML research and as a manifestation of the numerous calls to explainable IDS models in critical infrastructures [81].

Findings of this review show that the majority of ML-based IDS is focused on a broad range of attacks, though there is an increasing trend of the focus on multi-classification. This aligns with Vinayakumar et al.'s [82] study, who postulate that modern networks need more than binary classification detection. The introduction of IoT-specific attack data can also be considered as a tendency of domain-sensitive IDS research. However, the review reveals that the attack taxonomies are still data-based, which restricts broad applications and makes them challenging to apply to the real world. Despite the potential of the hierarchical model of attack, it has not been adequately exploited compared to flat classification models.

Continual learning and federated learning are two new directions that can enhance the practicality of ML-based IDS in the real world. Recent research indicates that federated learning enables detection that is privacy-preserving in a distributed environment, like an IoT network or smart cities [17]. Federated learning has privacy, scalability, and decentralized governance concerns that are a core part of a contemporary IDS by removing sensitive traffic information from a central node, along with permitting the joint training of a model. Continual learning is also essential. The IDS model is used in a situation where traffic and attack methodology are in a dynamic environment. Experiments that assess performance progressively and temporally are able to determine that model accuracy decays as there is a new or a new distribution of data [48]. This underscores the need for dynamic updates and regular retraining. These difficulties relate to more widespread maintainability issues in ML systems during deployment. The whole lifecycle, such as retraining and evolution of models, is crucial in maintaining the IDS in its strength and up-to-date status.

4.2 Identifying Gaps in the Literature

Despite the substantial volume of ML-based IDS research, this systematic review reveals that the gaps in this area are long-standing, interlinked, and prevent both theoretical development and practical implementation. This review highlights five gaps. One of the biggest gaps is related to the evaluation of the performance of the IDS. Most of the studies provide various metrics using the static train-test splitting in a single dataset. Good models in controlled experiments may perform poorly when confronted with changing traffic or other hidden types of attack. Second, a large number of studies rely upon a small variety of publicly available benchmark data sets, NSL-KDD, UNSW-NB15, and CICIDS variants. People can easily reproduce the findings and compare procedures across these datasets, which leads to methodological bias and reduces ecological validity due to the overuse of these datasets. Proclaimed high detection rates do not take into consideration old traffic patterns, artificial attack generation, and fixed network setups in those datasets. Benchmark collections have been previously criticized as not being a reflection of modern encrypted, cloud-based, or IoT-intensive environments.

Third, this review indicates that although feature engineering is essential to ML in IDS, feature pipelines are not standardized. Research employs heterogeneous sets of features, selection criteria, and dimensionality reduction methods, which are usually specialized to their datasets. This flexibility is helpful in optimization in single studies, but makes cross-studies comparison challenging and cumulative knowledge construction hard. Fourth, it is also observed in the results that there is a lack of connection between high-performance models and explainable decisions. Deep learning models always work better than traditional approaches in

complicated detection processes, but they are mostly viewed as black boxes. A smaller set of studies also explicitly uses XAI techniques. This non-disclosure has real-world consequences. Security analysts need straightforward explanations to verify alerts, justify security responses, and investigate incidents. Without XAI support, it can be concluded that the implementation of ML-based IDS is confined to controlled or high-risk environments.

Finally, the review outlines a lack in areas of operations lifecycle that include model maintenance, retraining strategies, system integration, and resource constraints. It is well understood that computational complexity is a constraint in many studies, yet few propose practical solutions for running in large-scale or resource-constrained environments. The fact that little has been given to maintainability, model versioning, and system evolution is an indication that the research in IDS is more of experimentation as opposed to deployment-based approaches. This is reflective of broader conclusions that maintenance issues are underreported despite having severe performance implications for the system.

4.3 Limitations of This Review

Although our systematic review is a synthesis of ML-based IDS work, it is important to acknowledge several limitations that affect the potential to interpret the results and, in general, their generalizability. First, the review relied on the published literature. Only publicly available and peer-reviewed studies have been included in the review. Unpublished appraisals, proprietary industrial systems, and in-house deployments are not reflected. Since most real-world applications of identity system detection find use in a private or sensitive setting, this omission can lead to biased results pertaining to academic outcomes over effective performance. Second, the review was constrained to dataset-based evidence. As the majority of the studies use benchmark datasets, the synthesis takes the features of the datasets used. Consequently, the conclusions regarding the performance of the models, the usefulness of the features, and the coverage of the attacks are highly dependent on the datasets.

Third, heterogeneity of study designs. There is a high level of heterogeneity in data sets, sets of features, model structures, and experimental procedures in the literature. Although thematic synthesis identifies general trends, it does not allow for a meta-analytic combination of quantitative indicators. Interpretations of comparisons made between studies should be a qualitative interpretation instead of a performance ranking of direct performance. Fourth, the temporal scope of the included studies. Even though the review is focused on the recent literature, intrusion detection has been developing rapidly. New methods are yet to be represented, such as federated learning, continual learning, and zero-trust architecture. Therefore, the results might not be enough to determine the trajectory of future research on IDS.

Finally, the current review is prone to publication bias, as studies with high positive outcomes will receive more publications, which might exaggerate the effectiveness of ML-based IDS. Another issue is the dataset bias. Benchmark datasets generated by popular networks are not necessarily representative of real-world network traffic. Hence, the models that perform well in testing may not do similarly in practice. It is also not very easy to reproduce an IDS research due to the inconsistency in data cleaning, feature engineering, and assessing procedures, and the lack of open code and details. All this may complicate the checking or comparison of the results between studies, and they have to be considered when interpreting this review.

4.4 Impact of These Limitations

Combined, these constraints imply that the results of this review are indicative rather than definitive. The synthesis encompasses the prevailing trend and those issues that are recurrent, but they are unable to encompass those practices that are not published, proprietary innovations, as well as the emergent

methodologies. However, there is some reliability in the gaps and implications given by the consistent patterns between the varying studies.

5 Conclusion

This systematic review has demonstrated that ML-based IDS has become much more efficient, but is still difficult to deploy due to the lack of data and models, which are challenging to understand, and maintenance issues. The six main themes collected helped to identify which steps have been made and what structural barriers continue to be significant in the research of IDS. The primary issue of future IDS should be bridging the gap between academic implementation and practice.

5.1 Implications for Practice

Practically, the results emphasize how IDS practitioners should not restrict their assessment to focus on accuracy. Priorities of deployment decisions should be:

- Reliability of data used in training, where the data should be up to date with the conditions of the network.
- Critical interpretability of features, understanding, and belief of model outputs is necessary. Analysts should be able to understand them.
- Operational robustness demands that the systems be resistant to concept drift and adversarial reforms.
- Lifecycle management with retraining, monitoring, and maintenance policies included.

Practitioners should be cautious when adopting models that were tested only in benchmark datasets and prefer instead the solution tested in a realistic context of traffic conditions.

5.2 Implications for Policy and Standardization

The review identifies the necessity of uniform evaluation models and common data sets at a policy level. The regulatory bodies and industry stakeholders may partner to:

- Creation of privacy-saving infrastructure for sharing actual traffic information.
- Withdraw benchmarking standards with intertemporal and interdomain validation.
- Advance explainability expectation of IDS installed in critical infrastructure.

5.3 Future Research Recommendations

The research gaps identified ought to be filled in future research by considering the following: first, it is necessary to work on creating datasets that capture encrypted traffic, IoT communication, and developing attack patterns, as well as ensuring privacy. Second, there is a need to study representations of features that can be transferred and compare their performance on several datasets and settings. Third, temporal validation, online learning conditions, and post-depletion monitoring should be included in the studies to measure the IDS's performance over a long period. Fourth, the inclusion of XAI methods in the IDS pipelines will reduce the gap between the accuracy of detection and the trust of the analysts. Finally, privacy-conscious and decentralized learning systems hold the potential to address data scarcity and a changing threat environment.

Acknowledgement: Not applicable.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Diana L, Dini P, Paolini D. Overview on intrusion detection systems for computer networking security. Computers. 2025;14(3):87. doi:10.3390/computers14030087.
2. Hozouri A, Mirzaei A, Effatparvar M. A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning, and emerging cybersecurity challenges. Discov Artif Intell. 2025;5(1):314. doi:10.1007/s44163-025-00578-1.
3. Asif MK, Khan TA, Taj TA, Naeem U, Yakoob S. Network intrusion detection and its strategic importance. In: Proceedings of the 2013 IEEE Business Engineering and Industrial Applications Colloquium (BEIAC); 2013 Apr 7–9; Langkawi, Malaysia. p. 140–4. doi:10.1109/beiac.2013.6560100.
4. Abbas SA, Almhanna MS. Distributed denial of service attacks detection system by machine learning based on dimensionality reduction. J Phys Conf Ser. 2021;1804(1):012136. doi:10.1088/1742-6596/1804/1/012136.
5. Khraisat A, Alazab A. A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges. Cybersecurity. 2021;4(1):18. doi:10.1186/s42400-021-00077-7.
6. Haripriya L, Jabbar MA. Role of machine learning in intrusion detection system: review. In: Proceedings of the 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA); 2018 Mar 29–31; Coimbatore, India. p. 925–9. doi:10.1109/iceca.2018.8474576.
7. Singh A, Prakash J, Kumar G, Jain PK, Ambati LS. Intrusion detection system: a comparative study of machine learning-based IDS. J Database Manag. 2024;35(1):1–25. doi:10.4018/jdm.338276.
8. Mari AG, Zinca D, Dobrota V. Development of a machine-learning intrusion detection system and testing of its performance using a generative adversarial network. Sensors. 2023;23(3):1315. doi:10.3390/s23031315.
9. Prazeres N, de C Costa RL, Santos L, Rabadão C. Engineering the application of machine learning in an IDS based on IoT traffic flow. Intell Syst Appl. 2023;17:200189. doi:10.1016/j.iswa.2023.200189.
10. Kumar A, Gutierrez JA. Impact of machine learning on intrusion detection systems for the protection of critical infrastructure. Information. 2025;16(7):515. doi:10.3390/info16070515.
11. Sowmya T, Mary Anita EA. A comprehensive review of AI based intrusion detection system. Meas Sens. 2023;28(4):100827. doi:10.1016/j.measen.2023.100827.
12. Aljanabi M, Ismail MA, Ali AH. Intrusion detection systems, issues, challenges, and needs. Int J Comput Intell Syst. 2021;14(1):560. doi:10.2991/ijcis.d.210105.001.
13. Umer MA, Junejo KN, Jilani MT, Mathur AP. Machine learning for intrusion detection in industrial control systems: applications, challenges, and recommendations. Int J Crit Infrastruct Prot. 2022;38:100516. doi:10.1016/j.ijcip.2022.100516.
14. Ali QI, Qaddoori SL. Machine learning-based intrusion detection and prevention system for IoT smart metering networks: challenges and solutions. Int STEM J. 2025;6(1):40–57. doi:10.22452/stem.vol6no1.4.
15. Ennaji S, de Gaspari F, Hitaj D, Kbidi A, Vincenzo Mancini L. Adversarial challenges in network intrusion detection systems: research insights and future prospects. IEEE Access. 2025;13:148613–45. doi:10.1109/access.2025.3600984.
16. Sajid Farooq M, Abbas S, Atta-ur-Rahman, Sultan K, Adnan Khan M, Mosavi A. A fused machine learning approach for intrusion detection system. Comput Mater Contin. 2023;74(2):2607–23. doi:10.32604/cmc.2023.032617.
17. Ragab M, Ashary EB, Alghamdi BM, Aboalela R, Alsaadi N, Maghrabi LA, et al. Advanced artificial intelligence with federated learning framework for privacy-preserving cyberthreat detection in IoT-assisted sustainable smart cities. Sci Rep. 2025;15(1):4470. doi:10.1038/s41598-025-88843-2.
18. Ahmed U, Nazir M, Sarwar A, Ali T, Aggoune EM, Shahzad T, et al. Signature-based intrusion detection using machine learning and deep learning approaches empowered with fuzzy clustering. Sci Rep. 2025;15(1):1726. doi:10.1038/s41598-025-85866-7.
19. Sataloff RT, Bush ML, Chandra R, Chepeha D, Rotenberg B, Fisher EW, et al. Systematic and other reviews: criteria and complexities. Ear Nose Throat J. 2021;100(6):403–6. doi:10.1177/01455613211025937.

20. Shaukat K, Luo S, Varadharajan V, Hameed IA, Xu M. A survey on machine learning techniques for cyber security in the last decade. IEEE Access. 2020;8:222310–54. doi:10.1109/access.2020.3041951.
21. Arqane A, Boutkhoum O, Boukhriss H, El Moutaouakkil A. A review of intrusion detection systems: datasets and machine learning methods. In: Proceedings of the 4th International Conference on Networking, Information Systems & Security; 2021 Apr 1–2; Kenitra, Morocco. p. 1–6. doi:10.1145/3454127.3456576.
22. Maseer ZK, Yusof R, Bahaman N, Mostafa SA, Foozy CFM. Benchmarking of machine learning for anomaly based intrusion detection systems in the CICIDS2017 dataset. IEEE Access. 2021;9:22351–70. doi:10.1109/access.2021.3056614.
23. Azam Z, Islam MM, Huda MN. Comparative analysis of intrusion detection systems and machine learning-based model analysis through decision tree. IEEE Access. 2023;11(4):80348–91. doi:10.1109/access.2023.3296444.
24. Raja R, Saleem H, Ahmad S, Arslaan M, Khan N. Cybersecurity incident detection (IDs) using machine learning. Int J Innov Res Comput Sci Technol. 2025;13(3):15–25. doi:10.55524/ijircst.2025.13.3.4.
25. Jacob SL, Sultana Habibullah P. A systematic analysis and review on intrusion detection systems using machine learning and deep learning algorithms. J Comput Cogn Eng. 2025;4(2):108–20. doi:10.47852/bonviewjcce42023249.
26. Gutierrez-Garcia JL, Sanchez-DelaCruz E, del Pilar Pozos-Parra M. A review of intrusion detection systems using machine learning: attacks, algorithms and challenges. In: Advances in information and communication. Cham, Switzerland: Springer Nature; 2023. p. 59–78. doi:10.1007/978-3-031-28073-3_5.
27. Ali TE, Chong YW, Manickam S. Machine learning techniques to detect a DDoS attack in SDN: a systematic review. Appl Sci. 2023;13(5):3183. doi:10.3390/app13053183.
28. Noor K, Imoize AL, Li CT, Weng CY. A review of machine learning and transfer learning strategies for intrusion detection systems in 5G and beyond. Mathematics. 2025;13(7):1088. doi:10.3390/math13071088.
29. Roopesh M, Nishat N, Rasetti S, Rahaman MA. A review of machine learning and feature selection techniques for cybersecurity attack detection with a focus on DDoS attacks. Acad J Sci Technol Eng Math Educ. 2024;4(3):178–94. doi:10.69593/ajsteme.v4i03.105.
30. Hernandez-Ramos JL, Karopoulos G, Chatzoglou E, Kouliaridis V, Marmol E, Gonzalez-Vidal A, et al. Intrusion detection based on federated learning: a systematic review. ACM Comput Surv. 2025;57(12):1–65. doi:10.1145/3731596.
31. Zakariah M, AlQahtani SA, Alawwad AM, Alotaibi AA. Intrusion detection system with customized machine learning techniques for NSL-KDD dataset. Comput Mater Contin. 2023;77(3):4025–54. doi:10.32604/cmc.2023.043752.
32. Kok S, Abdullah A, Jhanjhi N, Supramaniam M. A review of intrusion detection system using machine learning approach. Int J Eng Res Technol. 2019;12(1):8–15.
33. Liu H, Lang B. Machine learning and deep learning methods for intrusion detection systems: a survey. Appl Sci. 2019;9(20):4396. doi:10.3390/app9204396.
34. Dini P, Elhanashi A, Begni A, Saponara S, Zheng Q, Gasmi K. Overview on intrusion detection systems design exploiting machine learning for networking cybersecurity. Appl Sci. 2023;13(13):7507. doi:10.3390/app13137507.
35. Tripathy SS, Behera B. A review of various datasets for machine learning algorithm-based intrusion detection system: advances and challenges. SSRN J. 2025;18(4):1153. doi:10.2139/ssrn.5048254.
36. Alsubaei FS. Smart deep learning model for enhanced IoT intrusion detection. Sci Rep. 2025;15(1):20577. doi:10.1038/s41598-025-06363-5.
37. Waghmode P, Kanumuri M, El-Ocla H, Boyle T. Intrusion detection system based on machine learning using least square support vector machine. Sci Rep. 2025;15(1):12066. doi:10.1038/s41598-025-95621-7.
38. Gunupusala S, Kaila SC. Multi-class network anomaly detection using machine learning techniques. Contemp Math. 2024;5(2):37–49. doi:10.37256/cm.5220243723.
39. Ngo VD, Vuong TC, Van Luong T, Tran H. Machine learning-based intrusion detection: feature selection versus feature extraction. Clust Comput. 2024;27(3):2365–79. doi:10.1007/s10586-023-04089-5.
40. Zare F, Mahmoudi-Nasr P. Feature engineering methods in intrusion detection system: a performance evaluation. Int J Eng. 2023;36(7):1343–53. doi:10.5829/ije.2023.36.07a.15.

41. Howe A, Papa M. Feature engineering in machine learning-based intrusion detection systems for OT networks. In: Proceedings of the 2023 IEEE International Conference on Smart Computing (SMARTCOMP); 2023 Jun 26–30; Nashville, TN, USA. p. 361–6. doi:10.1109/smartcomp58114.2023.00086.
42. Ning S, Nguyen K, Bagchi S, Park Y. The study of feature engineering in machine learning and deep learning for network intrusion detection systems. In: Proceedings of the 2024 Silicon Valley Cybersecurity Conference (SVCC); 2024 Jun 17–19; Seoul, Republic of Korea. p. 1–5. doi:10.1109/svcc61185.2024.10637359.
43. Hashmi SK, Dubey GP, Himthani P. A review on feature selection techniques for intrusion detection system. SSRN J. 2022. doi:10.2139/ssrn.4289276.
44. Al-Jarrah OY, Siddiqui A, Elsalamouny M, Yoo PD, Muhaidat S, Kim K. Machine-learning-based feature selection techniques for large-scale network intrusion detection. In: Proceedings of the 2014 IEEE 34th International Conference on Distributed Computing Systems Workshops; 2014 Jun 30–Jul 3; Madrid, Spain. p. 177–81. doi:10.1109/icdcsw.2014.14.
45. Alsaffar AM, Nouri-Baygi M, Zolbanin HM. Shielding networks: enhancing intrusion detection with hybrid feature selection and stack ensemble learning. J Big Data. 2024;11(1):133. doi:10.1186/s40537-024-00994-7.
46. Govindaram A, Thilagavathi P, Anand AJ, Porkodi G, Parameswari D, Geetha R. Evaluating machine learning models for intrusion detection systems in IoT devices: an experimental study. Premier J Sci. 2026. doi:10.70389/pjs.100184.
47. Mohale VZ, Obagbuwa IC. Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability. Front Comput Sci. 2025;7:1520741. doi:10.3389/fcomp.2025.1520741.
48. Chua TH, Salam I. Evaluation of machine learning algorithms in network-based intrusion detection using progressive dataset. Symmetry. 2023;15(6):1251. doi:10.3390/sym15061251.
49. Agarwal A, Sharma P, Alshehri M, Mohamed AA, Alfarraj O. Classification model for accuracy and intrusion detection using machine learning approach. PeerJ Comput Sci. 2021;7(3):e437. doi:10.7717/peerj-cs.437.
50. Kunang YN, Nurmaini S, Stiawan D, Suprapto BY. Attack classification of an intrusion detection system using deep learning and hyperparameter optimization. J Inf Secur Appl. 2021;58(1):102804. doi:10.1016/j.jisa.2021.102804.
51. Sinha P, Sahu D, Prakash S, Rathore RS, Dixit P, Pandey VK, et al. An efficient data driven framework for intrusion detection in wireless sensor networks using deep learning. Sci Rep. 2025;15(1):34046. doi:10.1038/s41598-025-12867-x.
52. Aldaej A, Ullah I, Ahanger TA, Atiquzzaman M. Ensemble technique of intrusion detection for IoT-edge platform. Sci Rep. 2024;14(1):11703. doi:10.1038/s41598-024-62435-y.
53. Uddin MA, Aryal S, Bouadjenek MR, Al-Hawawreh M, Talukder MA. Hierarchical classification for intrusion detection system: effective design and empirical analysis. Ad Hoc Netw. 2025;178(2):103982. doi:10.1016/j.adhoc.2025.103982.
54. 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. Trans Emerging Tel Technol. 2021;32(1):e4150. doi:10.1002/ett.4150.
55. Surakhi OM, Garcia AM, Jamoos M, Alkhanafseh MY. A comprehensive survey for machine learning and deep learning applications for detecting intrusion detection. In: Proceedings of the 2021 22nd International Arab Conference on Information Technology (ACIT); 2021 Dec 21–23; Muscat, Oman. p. 1–13. doi:10.1109/acit53391.2021.9677375.
56. Thakkar A, Lohiya R. A review on challenges and future research directions for machine learning-based intrusion detection system. Arch Comput Meth Eng. 2023;30(7):4245–69. doi:10.1007/s11831-023-09943-8.
57. Aleesa AM, Zaidan BB, Zaidan AA, Sahar NM. Review of intrusion detection systems based on deep learning techniques: coherent taxonomy, challenges, motivations, recommendations, substantial analysis and future directions. Neural Comput Appl. 2020;32(14):9827–58. doi:10.1007/s00521-019-04557-3.
58. Jakotiya KS, Shirsath V, Mishra RG. Review on intrusion detection system using deep learning and machine learning. In: Proceedings of the 2023 International Conference on Integration of Computational Intelligent System (ICICIS); 2023 Nov 1–4; Pune, India. p. 1–4. doi:10.1109/icicis56802.2023.10430240.

59. Krishnamoorthy G, Sistla SMK. Exploring machine learning intrusion detection: addressing security and privacy challenges in IoT—a comprehensive review. J Knowl Learn Sci Technol. 2023;2(2):114–25. doi:10.60087/jklst.vol2.n2.p125.
60. Shivashankar K, Martini A. Maintainability challenges in ML: a systematic literature review. arXiv:2408.09196. 2024. doi:10.48550/arXiv.2408.09196.
61. Anis FM, Alabdullatif M, Aljbli S, Hammoudeh M. A survey on the applications of deep learning in network intrusion detection systems to enhance network security. IEEE Access. 2025;13:185357–73. doi:10.1109/access.2025.3624952.
62. Kamal H, Mashaly M. Combined dataset system based on a hybrid PCA–transformer model for effective intrusion detection systems. AI. 2025;6(8):168. doi:10.3390/ai6080168.
63. Naveed M, Arif F, Usman SM, Anwar A, Hadjouni M, Elmannai H, et al. A deep learning-based framework for feature extraction and classification of intrusion detection in networks. Wirel Commun Mob Comput. 2022;2022(1):2215852. doi:10.1155/2022/2215852.
64. Alshahrani E, Alghazzawi D, Alotaibi R, Rabie O. Adversarial attacks against supervised machine learning based network intrusion detection systems. PLoS One. 2022;17(10):e0275971. doi:10.1371/journal.pone.0275971.
65. Alkadi S, Al-Ahmadi S, Ben Ismail MM. RobEns: robust ensemble adversarial machine learning framework for securing IoT traffic. Sensors. 2024;24(8):2626. doi:10.3390/s24082626.
66. Barik K, Misra S. A comprehensive defense approach of deep learning-based NIDS against adversarial attacks. Multimed Tools Appl. 2025;84(31):37745–91. doi:10.1007/s11042-025-21008-5.
67. Gaspar D, Silva P, Silva C. Explainable AI for intrusion detection systems: lime and SHAP applicability on multi-layer perceptron. IEEE Access. 2024;12(11):30164–75. doi:10.1109/access.2024.3368377.
68. Lilhore UK, Simaiya S, Rahoof PP, Alroobaea R, Baqasah AM, Alsafyani M, et al. Advanced threat detection for smart cities through IoT sensor and network data integration with IoT-secure fusion. J Wirel Commun Netw. 2025;2026(1):20. doi:10.1186/s13638-025-02554-w.
69. Mia MS, Roy S, Ihsan MA, Hossain S, Ahamed MKU. Data-driven financial fraud detection using hybrid artificial and quantum intelligence. BenchCouncil Trans Benchmarks Stand Eval. 2025;5(4):100252. doi:10.1016/j.tbench.2025.100252.
70. Ahamed MKU, Karim A. Cascaded intrusion detection system using machine learning. Syst Soft Comput. 2025;7(10):200182. doi:10.1016/j.sasc.2024.200182.
71. Hossain MB, Hoque K. Machine learning approaches in IDS. Int J Sci Res Arch. 2022;7(2):706–15. doi:10.30574/ijsra.2022.7.2.0303.
72. Ni M. A review on machine learning methods for intrusion detection system. Appl Comput Eng. 2023;27(1):57–64. doi:10.54254/2755-2721/27/20230148.
73. Sharma V. Comparative analysis of machine learning models for intrusion detection systems. Panam Math J. 2025;35(3s):273–85. doi:10.52783/pmj.v35.i3s.3891.
74. Mohammed MS, Ali Talib H. Using machine learning algorithms in intrusion detection systems: a review. Tikrit J Pure Sci. 2024;29(3):63–74. doi:10.25130/tjps.v29i3.1553.
75. Kasongo SM, Sun Y. Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. J Big Data. 2020;7(1):105. doi:10.1186/s40537-020-00379-6.
76. Torabi M, Udzir NI, Abdullah MT, Yaakob R. A review on feature selection and ensemble techniques for intrusion detection system. Int J Adv Comput Sci Appl. 2021;12(5):538–53. doi:10.14569/ijacsa.2021.0120566.
77. Emirmahmutoğlu E, Atay Y. A feature selection-driven machine learning framework for anomaly-based intrusion detection systems. Peer Peer Netw Appl. 2025;18(3):161. doi:10.1007/s12083-025-01947-4.
78. Mahapatra A, Patra PK. A unified and scalable machine learning framework for feature fusion in object classification using weighted PCA with adaptive concatenation and dynamic scaling. Discov Comput. 2025;28(1):114. doi:10.1007/s10791-025-09622-1.
79. Talukder MA, Islam MM, Uddin MA, Hasan KF, Sharmin S, Alyami SA, et al. Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. J Big Data. 2024;11(1):33. doi:10.1186/s40537-024-00886-w.

80. Sweet LB, Müller C, Anand M, Zscheischler J. Cross-validation strategy impacts the performance and interpretation of machine learning models. Artif Intell Earth Syst. 2023;2(4):e230026. doi:10.1175/aies-d-23-0026.1.
81. Khan N, Ahmad K, Al Tamimi A, Alani MM, Bermak A, Khalil I. Explainable AI-based intrusion detection systems for industry 5.0 and adversarial XAI: a systematic review. Information. 2025;16(12):1036. doi:10.3390/info16121036.
82. Vinayakumar R, Alazab M, Soman KP, Poornachandran P, Al-Nemrat A, Venkatraman S. Deep learning approach for intelligent intrusion detection system. IEEE Access. 2019;7:41525–50. doi:10.1109/access.2019.2895334.