

A Comparative Analysis of Machine Learning and Deep Learning Approaches for Intrusion Detection in IoT/IIoT Networks

Yassine Atmani¹, Omar Chaieb¹, Nabil Kannouf² and Mohammed Benabdellah¹

¹LIPIO Laboratory, SIOSI Team, Faculty of Sciences, University Mohammed Premier, Oujda, Morocco

²LSA Laboratory, SOVAI Team, National School of Applied Sciences, Al Hoceima, Morocco

ABSTRACT

The fast development of the Internet of Things and the Industrial Internet of Things has significantly increased the attack surface of current networks, where heterogeneous and resource-constrained devices work in ever-changing traffic conditions. Intrusion Detection Systems are an important security measure that is usually installed at gateways, edge nodes, or cloud platforms in order to track traffic and detect anomalies or known attack patterns. Although recent machine learning and deep learning techniques have such positive results are reported to be over 99% detection accuracies on benchmark datasets are commonly acquired in controlled experimental conditions and provide black box models that do not transparency is a severe constraint in safety-critical IoT/IIoT implementations where explainability and responsibility is necessary.

*Corresponding author

Yassine Atmani, LIPIO Laboratory, SIOSI Team, Faculty of Sciences, University Mohammed Premier, Oujda, Morocco.

Received: September 01, 2026; **Accepted:** September 09, 2026; **Published:** September 16, 2026

Keywords: Intrusion Detection System, Internet of Things, Industrial IoT, Explainable Artificial Intelligence, SHAP, LIME, Machine Learning, Deep Learning, Systematic Literature Review, PRISMA, Class Imbalance, Feature Selection

Introduction

The Internet of Things (IoT) and Industrial Internet of Things (IIoT) devices have changed the threat landscape of today's networks. Security teams are now working in diverse and resource-constrained environments where perimeter security is insufficient. In response, IDS driven by machine learning and deep learning have emerged as the dominant countermeasure, with reported accuracies approaching 99% across familiar corpora including CICIDS2017, ToN-IoT, CIC-IoT22, UNSW-NB15 and X-IIoT ID. These systems are based on Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and hybrid architectures. Achieving a high accuracy on a benchmark is not always a guarantee of good performance in production, however. Operationally, such systems are deployed in one of three configurations: at the gate-way/router (network level), on the device itself to cut latency (edge level), or on centralized cloud platforms for aggregated monitoring [1,2]. The detection logic can be signature-based, anomaly-based, or hybrid. Two of the most common paradigms for ML/DL research are anomaly-based and hybrid. Based on the literature review, three major weaknesses of the current research in ML/DL-based IDS can be observed. First, there is the lack of representative datasets for training and evaluation, with many datasets being outdated, synthetically generated, or being highly class-imbalanced which results in high overall accuracy but low detection rates for rare but important attack types [2,3]. Second, the intensive pre-processing steps such

as feature selection heuristics, Synthetic Minority Oversampling Technique (SMOTE)-based augmentation and exhaustive hyper-parameter tuning represent a type of experimental over-fitting to controlled settings. Their robustness with respect to real-world traffic dynamics is not well validated [4,5]. Third, no reviewed study trains on one corpus (e.g. CICIDS2017) and then validates on a structurally distinct one such as ToN-IoT or Bot-IoT. If this is not done, generalizability claims cannot be supported [1,2].

The lack of Explainable Artificial Intelligence (XAI) in IoT-IDS design is almost total; this opacity constitutes the deeper structural basis of the three weaknesses identified above. In general, a model with a high detection rate will not give any indication of how a classification decision is related to pattern characteristics, packet-level behavior, or learned temporal patterns. This opacity is a serious impediment to analyst trust and adoption in critical infrastructure, where accountability and regulatory traceability are becoming major requirements [1, 2, 4, 5]. Methods like SHAP, LIME, Grad-CAM, and attention-based visualization might render IDS decisions explainable and actionable, but they have not been extensively studied in the literature reviewed [1, 2].

The key contributions of this systematic review are: (1) a structured comparison of the ML/DL-based IDS architectures used in IoT/IIoT environments; (2) a critical assessment of benchmark datasets, such as aging, class imbalance and cross-dataset generalization issues; (3) an analysis of feature selection strategies and their computational trade-offs; (4) an investigation of the limitations of class imbalance mitigation and evaluation; and (5) the identification of XAI as a critical missing element in current IoT-IDS research, and concrete directions for integration.

The rest of this paper is organized as follows: The research methodology is described in Section 2, which is based on PRISMA guidelines. The taxonomy of IDS methods is introduced in Section 3, and the ML/DL approaches and the XAI motivation are discussed. Benchmark datasets and feature selection techniques are discussed in Section 4. A discussion is provided in Section 5. Section 6 ends with future research directions.

Research Methodology

The study set was assembled under the PRISMA protocol (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), chosen for its transparent and reproducible approach to evidence synthesis. The methodology is divided into three steps: establishing a comparable framework in relation to previous surveys in order to demonstrate the novelty of this work; developing a search strategy that is reproducible across various academic repositories; and applying explicit inclusion and exclusion criteria, as well as quality criteria, to the candidate literature.

To describe the search strategy, Table 1 compares the 13 most representative studies in the 30-article PRISMA corpus that were selected for this review because they were relevant to the eight analytical dimensions described in this paper: the methodological coverage of ML/DL, analysis of the dataset and features, and integration of XAI. None of the above reviews cover all dimensions addressed in this paper, and in particular the XAI dimension, which is systematically not covered in previous IoT-IDS surveys. This is the main impetus of the present contribution.

Research Strategy

The search spanned five repositories covering the relevant disciplines: ScienceDirect, Google Scholar, SpringerLink, IEEE Xplore and the ACM Digital Library. The highest number of selected publications were from ScienceDirect (53%), followed by Google Scholar (20%), SpringerLink (13%), IEEE Xplore (7%), and ACM Digital Library (7%) as is shown in Figure 1. The search query was constructed by combining thematic clusters of keywords through Boolean operators.

The Final Query String used Across All Databases was:

("Internet of Things" OR IoT OR IIoT) AND ("Cy-ber Security" OR "Network Security" OR "Intrusion Detection") AND ("Artificial Intelligence" OR "Ma-chine Learning" OR "Deep Learning" OR XAI OR "Explainable AI").

The search was executed between January and March 2026, with database-specific adaptations applied where necessary. For IEEE Xplore and ACM Digital Library, the query was applied to the full-text and metadata fields. For Science Direct and SpringerLink, the query targeted title, abstract, and keywords. For Google Scholar, the query was restricted to titles and snippets to manage result volume. Search filters were applied to restrict results to: (i) peer-reviewed journal articles and conference proceedings; (ii) publications between 2020 and 2025; and (iii) English-language documents [6,7].

Inclusion, Exclusion and Quality Assessment Criteria

A multi-stage filtering protocol was applied to ensure both methodological rigor and reproducibility.

Inclusion Criteria

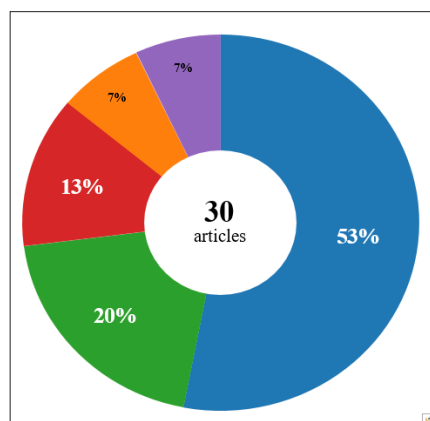
The inclusion criteria were: (i) publication between 2020 and 2025; (ii) the article must be peer-reviewed and published in a Q1–Q4 ranked journal or indexed in international conference proceedings; (iii) the article must be written in English; and (iv) the study must present concrete experimental contributions, such as detection rates, classification accuracy, or latency measurements, tested on well-characterized datasets.

Exclusion Criteria

Exclusion criteria were used to exclude: (i) non-peer-reviewed documents (blog posts, industry white papers, non-reviewed book chapters); (ii) duplicate records found in multiple databases; (iii) studies that were not directly related to cybersecurity in IoT/IIoT environments; (iv) purely theoretical contributions lacking replicable experimental methodology; and (v) surveys or systematic reviews that did not present new experimental findings.

Table 1: Comparative Analysis of Existing Surveys on IDS for IoT/IIoT Networks

Article	IoT	ML	DL	Hybrid	Dataset	FS	Taxonomy	XAI
Ahanger et al. [1] 2025	✓	✓	×	×	✓	×	✓	×
Bakhsh et al. [2] 2023	✓	✓	✓	×	✓	×	✓	×
Malik et al. [3] 2025	✓	✓	✓	×	✓	×	✓	×
Orman [4] 2025	✓	✓	✓	✓	✓	×	✓	×
Ben Henda et al. [5] 2024	✓	✓	×	×	✓	✓	✓	×
Thabit et al. [8] 2024	✓	✓	×	×	✓	✓	✓	×
Rahman et al. [9] 2025	✓	✓	✓	×	✓	✓	✓	×
Saif et al. [11] 2024	✓	✓	✓	×	✓	✓	✓	×
Bensaoud and Kalita [13] 2025	✓	✓	✓	✓	✓	×	✓	×
Sivagaminathan et al. [14] 2023	✓	✓	✓	✓	✓	✓	✓	×
Thomas and A.B.K. [15] 2025	✓	✓	×	✓	✓	✓	✓	×
Elmahfoud et al. [12] 2024	✓	✓	✓	×	✓	×	✓	×
Olanrewaju-George [10] 2025	✓	✓	✓	×	×	×	✓	×
This Review	✓	✓	✓	✓	✓	✓	✓	✓



- ScienceDirect: 16 articles (53%)
- Google Scholar: 6 articles (20%)
- SpringerLink: 4 articles (13%)
- IEEE Xplore: 2 articles (7%)
- ACM Dig. Lib.: 2 articles (7%)

Figure 1: Distribution of Selected Papers Across Academic Databases

Quality Assessment Framework

Each candidate paper was evaluated with a five-item quality checklist: (Q1) clarity of the research problem and objectives; (Q2) description of the datasets, hyperparameters and evaluation protocols that can be reproduced; (Q3) comprehensiveness of the evaluation metrics reported (accuracy, precision, recall, F1-score or operational metrics); (Q4) explicit treatment of dataset limitations or class imbalance; and (Q5) discussion of the limitations and threats to validity of the proposed approach. Only studies that scored 3 or more of the 5 criteria were included, and those that scored 2 or less were excluded as low-quality contributions.

Risk of Bias and Inter-Reviewer Agreement

To minimize the risk of selection bias, the screening and full-text eligibility phases were independently performed by two reviewers. Disagreements about inclusion were settled by discussion and, as needed, by a third reviewer by arbitration. Screening agreement between reviewers reached a Cohen's kappa of $\kappa = 0.82$, a level PRISMA classifies as substantial.

PRISMA Flow Diagram and Final Corpus

The entire selection process is summarized in the PRISMA flow diagram (Figure 2). Initially, 120 records were found in the five academic databases. After discarding 30 duplicate records, 90 records were kept for title and abstract screening. Of these, 42 records were excluded according to the inclusion and exclusion criteria. Of the remaining 48 reports, 18 were not found to be full-text eligible due to the following reasons: Not focused on IoT/IloT security ($n = 6$), No AI/ML-based methodology ($n = 8$), Survey duplicates ($n = 4$). The final corpus consisted of 30 peer-reviewed studies that were included in the present systematic review.

Literature Review and Results

Taxonomy of IDS Approaches

In IoT/IloT, there are two types of intrusion detection paradigms: signature-based and anomaly-based. Signature-based detectors compare live traffic against a stored catalogue of known attack

patterns. While their rule-based nature makes them straightforward to deploy and yields low false-positive rates for well-known threat categories, their drawback is obvious; they are unable to detect previously unseen attacks, and in fast-changing IoT settings, keeping signature databases current carries a high cost [1, 9]. Anomaly-based systems learn or build statistical models of normal traffic and flag traffic that deviates from the model as possible intrusions [9]. This makes them more appropriate for IoT/IloT networks, where the heterogeneity of devices and the ever-changing attack surface make it impractical to perform an exhaustive signature enumeration [11]. In the field of anomaly detection, classical ML classifiers such as decision trees, Support Vector Machines (SVMs), random forests and shallow neural networks have been extensively used and have proven to be effective on standard IoT corpora [8, 12]. More recent proposals include the use of deep learning and hybrid architectures, which employ hierarchical and temporal feature learning to identify multi-stage attack patterns [13]. These models are more likely to generalize well to other, more diverse IloT environments [2, 3]. Federated learning has also been explored as a privacy-preserving method for collaborative anomaly detection without sharing the data at the central location [10]. This review builds on existing surveys of ML and DL for network IDS which offer context on challenges with datasets and feature dimensionality.

ML/DL Techniques

Approaches Based on Machine Learning

Most classical ML-based IoT-IDS methods are based on handcrafted feature pipelines for a specific accuracy on training corpora [1]. In most published work, metrics like accuracy and precision are reported, while recall and interpretability are comparatively less reported [2]. The federated formulations address data distribution issues in decentralized IoT environments, but the aggregated global models are black-box predictors and their inner workings cannot be understood by analysts [3]. Metaheuristics such as PSO, together with Extreme Learning Machines, keep inference cheap on constrained hardware yet leave the link between chosen features and their security meaning unexplained [7,8].

Deep Learning-Based IDS

Near-perfect detection rates are consistently obtained on standard IoT benchmarks with deep learning architectures such as CNNs, Self-Organizing Maps (SOMs), Autoencoders and LSTMs [5,6]. The "trade-off" is well known: these models are black boxes. There is no linkage between an alert and a traffic attribute, packet attribute, or learned temporal dependency. The dataset-centric research tackles the issues of pre-processing rigor and class imbalance but does not go back to the properties of the data that drive model decisions [9]. System-centric work integrates IDS into the larger IoT security system, but the AI inference layer is still obscure [11]. In both cases, detection performance, scalability, and computational efficiency are optimized at the expense of transparency, making it impossible to independently verify operational alerts.

Explainable AI in IoT Intrusion Detection: State of the Art and Identified Gaps

In almost all the IDSs reviewed, high detection rates and low interpretability are found to coexist, and this is the primary motivation for the integration of XAI. Although there are many works on IDS for IoT/IloT, only a few studies have started to integrate explainability mechanisms into their detection pipelines. The following is a critical review of the most prominent XAI techniques used in network intrusion detection, their practical limitations and the gaps that drive the current review.

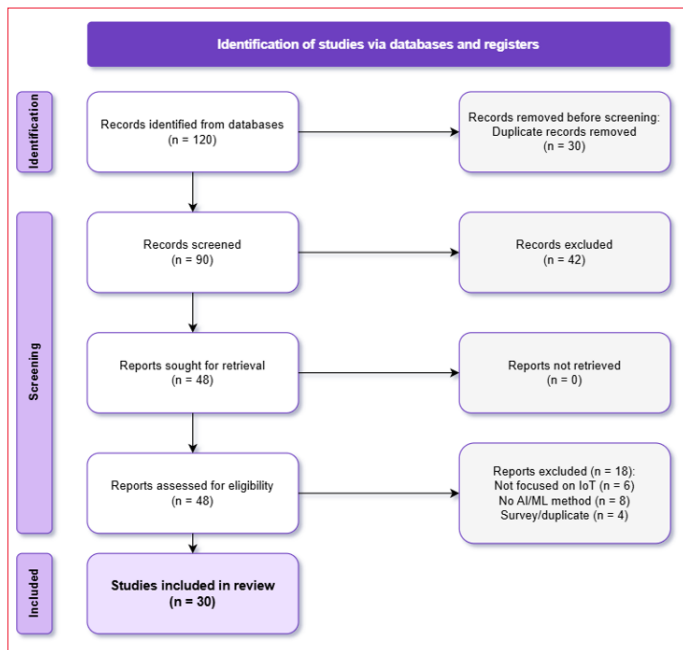


Figure 2: PRISMA Flow Diagram of the Study Selection Process

SHAP-based Explanations for IDS

In the context of network intrusion detection, the most widely used post-hoc explanation methods are SHAP (Shapley Additive explanations). Neupane et al. gave a detailed overview of explainable IDS architectures and introduced a generic human-in-the-loop X-IDS framework, where SHAP is the most popular post-hoc approach because of its theoretical basis in cooperative game theory. Arreche et al. systematically compared SHAP and LIME on three network intrusion datasets, using six metrics (descriptive accuracy, sparsity, stability, efficiency, robustness, and completeness), and found that SHAP outperformed LIME in all of them. SHAP is able to provide a higher degree of global completeness but has an $O(2n)$ computational complexity, which makes it impractical for real-time inference on resource-constrained IoT edge nodes. Ahmed et al. combined feature selection using SHAP with hybrid bagging-boosting classifiers and demonstrated that feature attribution via SHAP not only boosts the accuracy of the model but also enhances its explainability, thus creating a direct link between feature selection and explainability [16-18].

LIME and Local Explanations

LIME (Local Interpretable Model-Agnostic Explanations) has also been used in network security applications. Amarasinghe et al. were among the first to use LIME for anomaly detection using deep neural networks, enabling security analysts to interpret individual decisions made by the anomaly detector in terms of the specific attributes of the packets [19]. As Arreche et al. have shown, however, the inherent instability of LIME, which can give different feature rankings for repeated explanations of the same instance, can make it less reliable in operational deployments where consistency is important [17]. This stability is especially concerning in IoT systems, where automated decision pipelines need to be reproducible for compliance and audit reasons.

Attention-based and Intrinsic Explainability

Attention-based explainability mechanisms are more architecturally integrated, as they are built into the model structure, rather than being added on top of it. In contrast to SHAP or LIME, attention layers are a natural by-product of inference, and require little

extra computation. This is especially relevant for the use of attention mechanisms in IoT-IDS systems, where computational resources are limited. However, explanations based on attention are still architecture dependent and cannot be used as a generic explainability layer, but need to be redesigned based on the model.

Grad-CAM and Visual Explanations

Although popular in computer vision problems, Grad-CAM has been scarcely applied to network intrusion detection because network traffic data is usually tabular or sequential. It is applicable to CNN-based IDS architectures where traffic is represented as image-like data, and is not widely used in the IoT-IDS domain.

Comparative Analysis of XAI Methods

A structured comparison of these XAI methods in the context of IoT-IDS is presented in Table 2. The table summarizes the trade-offs between interpretability, computational cost, and operational suitability for resource-constrained IoT deployments.

Identified Gaps and Practical Implications

The reviewed studies clearly indicate that none of the published works in the IoT-IDS field has been able to combine XAI with the feature selection process and detection pipeline at the same time. Existing methods treat explainability as a mere add-on for models that have already been trained, rather than integrating it as a core design requirement. For example, Achbarou et al. obtained a 99.95% accuracy on CICIDS2017 by using CNN and feature selection based on XGBoost, but they did not offer a way to explain the alert decision in terms of specific traffic-level attributes [20]. This is the prevailing approach: detection performance is maximized without considering the operational transparency requirements.

A practically useful XAI-based IDS would not only detect an intrusion, but also provide a human-readable description of the intrusion. For instance, if a DDoS detection alert is generated due to the packet rate (flow feature), the source IP entropy (network feature), and the flow duration (temporal feature), the analyst can determine that this is a volumetric attack versus an application-layer attack and respond appropriately [21].

The main research gap found in this review is the absence of such integrated XAI mechanisms, and it is the main motivation for the future research directions presented in Section 6.

Benchmark Datasets and Feature Selection Approaches

Intrusion Detection Datasets

Reproducible evaluation of intrusion detection systems relies on the availability of labeled network traffic corpora. Such benchmark datasets encapsulate both benign flows and curated attack scenarios, forming the empirical foundation on which ML/DL classifiers are trained and assessed. To better represent a changing network environment and the current at-tack scenario, several datasets have been developed over time [22-25]. Table 3 presents a structured description of the thirteen main datasets that were identified in the literature reviewed, complemented with information on the method used to generate traffic and on any known imbalance.

Beyond their structural characteristics, the reviewed datasets exhibit four critical limitations that significantly affect the validity and generalizability of IDS models trained on them.

Dataset Aging: KDD99 and NSL-KDD, which are still

Table 2: Comparative Analysis of XAI Methods Applied to IoT-IDS

Method	Type	Advantages	Limitations	Cost	IoT-IDS
SHAP	Post-hoc, global	High interpretability, theoretically grounded	$O(2^n)$ complexity, slow on edge	High	Moderate
LIME	Post-hoc, local	Model-agnostic, flexible	Instability across runs	Medium	Moderate
Grad-CAM	Post-hoc, visual	Intuitive visualization	Limited to CNN	Low	Low
Attention	Intrinsic, local	Embedded in architecture, low overhead	Architecture-dependent	Low	High
SHAP-Tree	Post-hoc, global	Efficient for tree models	Tree models only	Low–Medium	High

Table 3: Structured Characterization of Benchmark Datasets Reviewed in IoT-IDS Literature

Dataset	# Samples	# Feat.	Classes	Attack Coverage	Traffic	Imbalance
KDD99	~5M	41	20+	DoS, Probe, R2L, U2R	Synthetic	High
NSL-KDD	~150K	41	20+	DoS, Probe, R2L, U2R	Synthetic	Moderate
AWID	~2.3M	156	16	Flooding, Injection, Impersonation	Real (802.11)	Moderate
CICIDS2017	var.	80	15	DDoS,	Brute Force, Web	Mixed
CSE-CIC-IDS2018	var.	80	18	Extended	CI-CIDS2017	Mixed
UKM-IDS20	125K	46	4	ARP, DoS, Probing, Exploits	Real (IoT)	Moderate
Bot-IoT	~73M	40	5	DDoS, DoS, Scanning	Real (IoT)	Severe
ToN-IoT	var.	44	9	Ransomware, DDoS, Backdoor	Real (IIoT)	High
WUSTL-IIoT	~1.2M	41	4	DoS,	Injection, Backdoor	Real (SCADA)
IoTID20	~626K	83	3 levels	Binary/Type/ Subcategory	Real (Smart-home)	Moderate
CIC-IoT-2022	var.	var.	18	TCP/UDP Flood, RTSP	Real	(60 dev.)
IoT-23	20 scen.	var.	20	Botnet, Malware, C2	Real (IoT)	High
N-BaIoT	var.	115	2 fam.	Mirai,	BASH-LITE	Real (IoT)

widely used in the literature, were created in 1999 and are not representative of the current IoT device diversity, protocol stacks, or attack vectors like ransomware, supply-chain attacks, or IoT-specific attacks. High accuracy models cannot be regarded as representative of deployable performance in current IoT/IIoT environments [1,11].

Class Imbalance: Severe class imbalance is a frequent problem in all of the datasets analyzed. For example, Bot-IoT has about 73 million samples, but is dominated by DDoS and DoS traffic with less than 1% of the corpus comprising reconnaissance and exfiltration classes [22]. Here a trivial model that labels everything as DDoS still posts above 95% aggregate accuracy, masking the fact that it misses the rare but damaging attack classes entirely. This problem can be partially addressed using oversampling techniques like SMOTE, but synthetic artefacts are created that do not transfer to real operational traffic [3,4].

Cross-Dataset Generalization: None of the 30 studies reviewed performs cross-dataset evaluation, which is the training of a model on one corpus and the testing of the model on a structurally different one. The fact that this is not included makes it difficult to determine if the performance is due to actual intrusion detection or due to memorization of patterns in the dataset. Since the traffic generation methodology, feature distributions, and attack semantics are different in CICIDS2017 and ToN-IoT or Bot-IoT respectively, a model trained on CICIDS2017 would likely suffer

a significant performance degradation when evaluated on ToN-IoT or Bot-IoT [22,24,25].

Concept Drift: As device firmware is updated, new communication protocols are introduced and attack methods are refined, network traffic characteristics are constantly changing. Such temporal evolution cannot be captured in static datasets. Models built on fixed corpora are thus anticipated to lose detection accuracy with time when used in production systems, which is not covered by any of the reviewed studies [4,5].

Feature Selection

IoT network traffic exhibits high dimensionality, heterogeneity, and substantial noise, making principled feature selection an indispensable preprocessing stage for any ML-based IDS pipeline. A good dimensionality reduction can also speed up the calculation on the resource-limited endpoint and even avoid overfitting by eliminating the unnecessary or redundant attributes. The principle of parsimony is key to the development of generalizable security models. The feature selection techniques can be broadly classified into three categories: filter-based methods, wrapper-based methods, and embedded/evolutionary approaches, each with its own set of pros and cons, especially in the context of IoT-IDS, which are characterized by their large-scale nature and the vast number of features. Thakkar and Lohiya [29] have presented a thorough survey of feature selection techniques for IDS, covering more than 50 different techniques used in the field from 2010 to 2022.

Filter-based methods measure feature relevance independently of the subsequent classifier, using statistical criteria such as chi-squared (χ^2) tests, Pearson correlation coefficients (PCC), and mutual information (MI) gain. These methods are computationally efficient and scale well to high-dimensional IoT traffic datasets. In [5], the authors showed that combining correlation-based feature selection with statistical feature ranking reduced the dimensionality of IoT network traffic data while improving SVM-based attack detection performance. The primary limitation of filter methods is that they are classifier-agnostic: they may retain features that are statistically discriminative but suboptimal for a specific learning algorithm, and they provide no inherent mechanism to map selected features to human-understandable security semantics [26-29].

Wrapper-based methods compare candidate feature sub-sets based on the performance of a given target classifier, e.g. Random Forest or ELM. This approach generally results in higher accuracy on the target task compared to filter-based approaches, but is more complex. In high-stakes IDS scenarios, where detection performance can have a significant impact on operational security, the additional cost is often justified [7, 14]. The empirical results from the reviewed literature are consistent in that wrapper methods result in an effective feature dimensionality of 11-17 features on standard bench-mark datasets like NSL-KDD [23] and UNSW-NB15 and offer comparable or better classification accuracy and F-score than the full-dimensional baselines. Chaieb et al. [30] proposed a novel two-stage group-based wrapper approach where they first ranked the feature categories according to their discriminative contribution and then selected the optimal features within each of the categories retained, obtaining a 98.73% feature reduction rate in the cybersecurity domain.

Evolutionary and embedded methods such as Genetic Algorithms (GAs) and PSO, are well suited to a global search over the feature space without exhaustive search, especially when the feature space

is high dimensional (80 to 156 dimensions) for IoT traffic datasets [7, 14]. Embedded methods like feature importance selection using XGBoost [20] embed the selection process into model training, and they achieve competitive computational efficiency and simultaneous optimization of feature relevance and model accuracy.

Table 4 shows a tabular comparison of these categories based on the dimensions most relevant to deployments of IoT-IDS.

However, a major drawback of all the approaches to feature selection reviewed is that none of the 30 reviewed studies offer a human-readable mapping between the selected feature subset and the corresponding semantic security concepts. For instance, the wrapper method might use “Flow Duration”, “Bwd Packet Length Mean” and “PSH Flag Count” as discriminative features for detecting DDoS, but no reviewed study provides any explanation of why these features are semantically relevant to the attack category or how a security analyst should interpret a model alert in terms of these low-level traffic statistics. This semantic opacity is a direct link to the XAI research gap discussed in Section 3.3.

Discussion

The literature reviewed has shown that there is a real technical advancement in the application of ML and DL in IoT IDS. Meanwhile, structural deficiencies remain in almost all of the studies surveyed. The most common issue is the heavy usage of KDD99 and NSL-KDD. These corpora were gathered over 20 years ago and are not representative of the current state of device diversity, protocol variety or attack tools. Newer alternatives like UNSW-NB15 [23] and Bot-IoT [22] are more common in recent publications, but such systematic cross-dataset testing is uncommon, where a model is trained on one dataset and tested on a structurally different one. Generalizability claims are based on an untested assumption without this step [1, 2].

Table 4: Comparative Analysis of Feature Selection Approaches for IoT-IDS

Method	Comp. Cost	Interpretability	Drift Robust.	Scalability	IoT Suit.
Filter (χ^2 , MI, PCC)	Low	High	Low	High	High
Wrapper (RF, SVM, ELM)	High	Moderate	Moderate	Low	Moderate
Evolutionary (GA, PSO)	Medium	Low	Moderate	Moderate	Moderate
Embedded (XG Boost, LASSO)	Medium	Moderate	Low	High	High

The second recurring problem is the almost exclusive focus on accuracy as the sole criterion of evaluation. Accuracy is not a suitable metric for imbalanced data. Categories like U2R, R2L are so rare that even a model with 95% accuracy can fail at detecting them. Oversampling methods, such as SMOTE, partially overcome this imbalance, but they are not a cure for the imbalance or its impact on recall of minority classes [3, 4]. Third, all the reviewed approaches for feature selection are static. Feature sets are not updated when deployed, but are fixed at the time of training. This design does not account for concept drift: the ever-changing nature of network traffic profiles, device behaviors and attack methods. The model that is trained today could gradually become less effective over the course of months of production. The literature does not contain any online, adaptive selection mechanisms that adapt feature relevance based on traffic changes [29].

Fourth and most importantly, the outputs of the IDS models cannot be verified independently because it is not possible to explain the detection pipeline. The relationship between feature selection and XAI is an area of research with a great amount of practical relevance that has been barely explored. A significant advance toward accountable, human-oriented IoT cybersecurity would be to create feature selection processes that would maximize discriminative power and generate analyst-readable attributions between selected features and meaningful security semantics [1,2,4,5].

Table 5 summarizes the five principal research gaps identified in this review, together with their observed manifestations in the literature and the proposed solutions that should guide future research.

Adversarial Robustness: One of the most important aspects that is not covered in the literature reviewed is the adversarial machine learning attacks on models and their evaluation of the robustness of the models [21]. The idea behind adversarial machine learning is to introduce small perturbations into the input traffic that are imperceptible to human analysts yet are deliberately crafted to fool the trained classifiers into making wrong predictions. For IoT-IDS, a sophisticated attacker might be able to create malicious traffic that has a similar statistical signature to that of benign traffic, and yet is still harmful. Given full model knowledge, leading deep-learning IDS see reported accuracy collapse from the 99% range to below 20% once FGSM or Carlini-Wagner perturbations are applied [21]. This is especially problematic in IoT systems where the behavior of the devices is predictable and the traffic patterns are homogeneous, allowing adversarial perturbations to be more easily calibrated. Laboratory benchmarking and real-world deploy ability are separated by a huge margin, as none of the 30 studies reviewed included an adversarial robustness evaluation. Fifth, operational deployment constraints that are essential for IoT/IIoT environments were almost entirely overlooked in the reviewed literature. Table 6 shows the summary of the evaluation metrics reported in the 30 studies reviewed and the systematic lack of operational metrics. Detection accuracy and F1-score are reported in most studies, whereas critical deployment metrics such as inference latency, memory foot-print and energy consumption are reported in less than two studies. The accuracy of 99% and inference latency of 500 ms would not be suitable for an operational IoT-IDS deployment on resource-constrained edge nodes [20, 4].

Conclusion and Future Directions

This work has offered a systematic review of Intrusion Detection Systems (IDSs) for IoT and IIoT environments using ML/DL techniques, including the feature selection strategies, limitations of benchmark datasets, and the inclusion of Explain-able Artificial Intelligence (XAI). Thirty peer-reviewed studies were reviewed from 2020 to 2025.

What this means is that the models are becoming more sophisticated, but continue to rely on outdated data, rigid pre-

processing pipelines, and opaque black-box decision-making. All of these combined makes even the highest performing systems reviewed less valuable for operation in actual IoT/IIoT deployments.

Five directions are suggested to overcome these gaps. (1) Co-design feature selection with XAI from the beginning, rather than tackling it at the end. This implies that the outputs of the feature selection procedures are directly interpretable in security terms, not just in statistical discriminatory terms. (2) Federated learning architectures for IoT-IDS should include explainability elements to reveal the rationale behind the decisions made while maintaining data locality. (3) Adversarial robustness should be a requirement for evaluation dimension; models should be tested with perturbation attacks (e.g. FGSM, CW) prior to being deployed. (4) Online adaptive feature selection should replace static pipelines in all production-oriented IDS designs, enabling real-time tracking of concept drift as traffic profiles and attack patterns evolve. (5) All claims of edge deploy ability must be validated through hardware-level benchmarking on microcontroller-class platforms (e.g., Rasp-berry Pi, Jetson Nano, ARM Cortex-M), reporting inference latency, memory usage, and energy consumption alongside classification metrics.

There are also a few new paradigms that should be noted. The transformer architectures, which were developed for language tasks, have been modified to fit network traffic sequences and demonstrate a greater ability to learn long-range dependencies than typical LSTM models. Graph Neural Networks offer a natural representation of the topology of IoT networks: devices, traffic flows, and attack propagation paths are represented as nodes and edges in a dynamic graph, allowing for detection of coordinated, multi-hop attacks that flow-level classifiers are unable to detect. Tiny ML and edge intelligence push inference directly onto the hardware. On-device deployment would now be possible due to the absence of cloud dependency, as well as latency and privacy concerns, if anomaly detection could be performed on ARM Cortex-M or RISC-V processors with a memory budget of less than 256 KB.

Table 5: Summary of Identified Research Gaps and Proposed Solutions

#	Research Gap	Evidence in Literature	Proposed Solution
1	Reliance on outdated datasets	KDD99 and NSL-KDD still widely used; no IoT-native traffic	Continuously updated IoT-native benchmarks with realistic heteroge-neous traffic
2	Severe class imbal-ance	Bot-IoT: DDoS >99% of samples; U2R/R2L recall <10%	MCC and Balanced Ac-curacy as primary met-rics; advanced resam-pling strategies
3	No cross-dataset gen-eralization	0 of 30 studies test across structurally dif-ferent datasets	Standardized cross-dataset evaluation as mandatory validation protocol
4	Static feature selec-tion	All FS methods are fixed; no adaptive mech-anism reported	Online adaptive feature selection with concept drift detection
5	Systematic absence of XAI	0 of 30 studies integrate SHAP, LIME, or atten-tion	XAI-native IDS design: explainability as archi-tectural constraint, not post-hoc addition

Table 6: Evaluation Metrics Reported Across the 30 Reviewed Studies

Metric	Reported in	Note
Accuracy	30 / 30	Universally reported; misleading under severe class imbalance
Precision	23 / 30	Frequently reported
Recall	21 / 30	Often omitted despite critical security rele-vance
F1-Score	19 / 30	Partially reported
ROC-AUC	8 / 30	Rarely reported
MCC	0 / 30	Never reported; most robust metric under class imbalance

Inference latency	4 / 30	Almost never reported
Memory footprint	0 / 30	Never reported
Energy consumption	0 / 30	Never reported

Adversarial training and certified robustness are not “nice-to-have” improvements; they are essential for any machine that has to perform in the presence of an adaptive adversary. Taken together, these directions point toward a next-generation IoT-IDS that is accurate while also being accountable, lightweight and resistant to deliberate evasion.

References

- Ahanger TA, Ullah I, Algamdi SA, Tariq U (2025) Machine learning-inspired intrusion detection system for IoT: Security issues and future challenges. Computers and Electrical Engineering 123: 110265.
- Bakhsh SA, Khan MA, Ahmed F, Alshehri MS, Ali H, et al. (2023) Enhancing IoT network security through deep learning-powered Intrusion Detection System. Internet of Things 24: 100936.
- Malik J, Akhunzada A, Al Shamayleh AS, Zeadally S, Almogren A (2025) Hybrid deep learning-based threat intelligence framework for Industrial IoT systems. Journal of Industrial Information Integration 45: 100846.
- Orman A (2025) Cyberattack Detection Systems in Industrial Internet of Things (IIoT) Networks in Big Data Environments. Applied Sciences 15: 3121.
- Ben Henda N, Msolli A, Haggui I, Helali A, Maaref H (2024) Attack Detection in IoT Network Using Support Vector Machine and Improved Feature Selection Technique. Journal of Network and Systems Management 32: 92.
- Peterson JM, Leevy J, Khoshgoftaar T (2021) A Review and Analysis of the Bot-IoT Dataset. 2021 IEEE International Conference on Service-Oriented System Engineering (SOSE): 20-27.
- Birhriya SK, Ahlawat P, Jain AK (2025) Detection and prevention of spear phishing attacks: A comprehensive survey. Computers & Security 151: 104317.
- Thabit F, Can O, Abdaljlil S, Alkhzaimi HA (2024) Enhanced an Intrusion Detection System for IoT networks through machine learning techniques: an examination utilizing the AWID dataset. Cogent Engineering 11: 2378603.
- Rahman MM, Al Shakil S, Mustakim MR (2025) A survey on intrusion detection system in IoT networks. Cyber Security and Applications 3: 100082.
- Olanrewaju-George B, Pranggono B (2025) Federated learning-based intrusion detection system for the internet of things using unsupervised and supervised deep learning models. Cyber Security and Applications 3: 100068.
- Saif S, Ansari AA, Biswas S, Giri D (2024) A comprehensive analysis of machine learning-based intrusion detection systems: evaluating datasets and algorithms for internet of things. Journal of Cyber Security Technology 10: 1-27.
- El Mahfoud E, El Hajla S, Yassine M, Mounir S (2024) Machine Learning Algorithms for Intrusion Detection in IoT Prediction and Performance Analysis. Procedia Computer Science 236: 460-467.
- Bensaoud A, Kalita J (2025) Optimized detection of cyber-attacks on IoT networks via hybrid deep learning models. Ad Hoc Networks 170: 103770.
- Sivagaminathan V, Sharma M, Henge SK (2023) Intrusion detection systems for wireless sensor networks using computational intelligence techniques. Cybersecurity 6: 27.
- Thomas L, Anoop BK (2025) An Efficient IoT Based Intrusion Detection System Using Optimization Kernel Extreme Learning Machine. International Journal of Computer Network and Information Security 17: 72-87.
- Neupane S, Ables J, Anderson W, Mittal S, Rahimi S, et al. (2022) Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges and Opportunities. IEEE Access 10: 112392-112415.
- Arreche O, Guntur TR, Roberts JW, Abdallah M (2024) E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection. IEEE Access 12: 23954-23988.
- Ahmed U, Jiangbin Z, Almogren A, Sadiq M, Ur Rehman A, et al. (2024) Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems. Scientific Reports 14: 30532.
- Amarasinghe K, Kenney K, Manic M (2019) Toward Explainable Deep Neural Network Based Anomaly Detection. 12th International Conference on Human System Interaction <https://hsi2019.welcometohsi.org/>.
- Achbarou O, Datsi T, Bourkoukou O, El kiram AM (2026) Enhanced intrusion detection system using feature selection and hybrid learning models for high performance and efficiency in an IoT environment. Journal of Engineering Research 14: 953-965.
- Martins N, Cruz JM, Cruz TJ, Abreu PH (2020) Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review. IEEE Access 8: 35403-35419.
- Koroniotis N, Moustafa N, Sitnikova E, Turnbull BP (2019) Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset. Future Generation Computer Systems 100: 779-796.
- Moustafa N, Slay J (2015) UNSW-NB15: a comprehensive data set for network intrusion detection systems. Military Communications and Information Systems Conference <https://www.sto.nato.int/icmcis/>.
- Alsaedi A, Moustafa N, Tari Z, Mahmood A, Anwar A (2020) TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems. IEEE Access 8: 165130-165150.
- Sharafaldin I, Habibi Lashkari A, Ghorbani AA (2018) Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. Proceedings of the 4th International Conference on Information Systems Security and Privacy <https://www.sto.nato.int/icmcis/>.
- Chaieb O, Kannouf N, Amjoun R, Benabdellah M (2023) Machine Learning-Based Intrusion Detection System: Review and Taxonomy. Lecture Notes in Networks and Systems 625: 10-21.
- Berhili M, Chaieb O, Benabdellah M (2024) Intrusion detection systems in IoT based on machine learning: a state of the art. Procedia Computer Science 251: 99-107.
- Ahmad Z, Khan AS, Wai Shiang C, Abdullah J, Ahmad F (2021) Network intrusion detection system: a systematic study of machine learning and deep learning approaches. Transactions on Emerging Telecommunications Technologies 32: 4150.

29. Thakkar A, Lohiya R (2022) A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. Artificial Intelligence Review 55: 453-563.
30. Chaieb O, Kannouf N, Benabdellah M (2025) A new wrapper feature selection approach for binary ransomware detection. IAES International Journal of Artificial Intelligence 14: 2102-2110.