International Journal of Engineering Research & Management Technology
(Peer-Reviewed, Open Access, Fully Refereed International Journal) ISSN: 2348-4039
Volume 11, Issue-1 January-February- 2024 Impact Factor: 7.09
Email: editor@ijermt.in www.ijermt.in

AI-Driven Intrusion Detection for Industrial IoT Networks: A Critical Review of Machine Learning and Deep Learning Approaches, Benchmark Datasets, and Evaluation Gaps

Mungara Dhaval Vithalbhai
Research Scholar NIILM University, Kaithal, Haryana, India
Dr. M K Agarwal
Professor, NIILM University, Kaithal, Haryana

Abstract

IIoT networks link sensors, controllers and enterprise platforms that were originally built for closed, isolated operation. In such settings, signature-matching intrusion detection systems (IDS) fall behind attacks that are new, polymorphic or staged over time. Artificial intelligence (AI)-driven IDS, built on machine learning (ML) and deep learning (DL), are widely proposed as the remedy, and published studies routinely report accuracies above 99%.

We examine that evidence with a critical eye. It synthesises foundational, survey and empirical literature from 1987 to 2026 across five themes: the security requirements of IIoT, the threat landscape, classical ML detectors, DL detectors (ANN, CNN and LSTM), and the benchmark datasets and evaluation practices on which reported results depend. Eleven empirical studies, twelve surveys and foundational works, and five datasets are compared in structured tables.

The review finds that tree-based learners are consistently strong on tabular flow features, that DL gains are real in some studies but conditional on input construction and data quality, and that feature selection and class-imbalance treatment materially change reported performance. It also finds that much of the headline performance in the field is undermined by three recurring weaknesses: data leakage and near-duplicate records between training and test sets, documented defects in widely used benchmarks such as CICIDS2017, and over-reliance on aggregate accuracy that hides false alarms and minority-class failures. The paper identifies five research gaps and proposes an evaluation agenda — leakage-safe preprocessing, per-class and false-positive reporting, multi-dataset testing and cost measurement — for AI-driven IDS research intended for industrial deployment.

Keywords: IIoT security; network intrusion detection; tree-based classifiers; deep neural networks; false alarms; benchmark data quality

1. Introduction

Across manufacturing, energy, transport and utilities, plants that once ran on isolated control networks now operate as connected digital ecosystems through the Industrial Internet of Things (IIoT). Sensors, programmable controllers, supervisory systems and enterprise platforms now exchange data continuously, enabling predictive maintenance, real-time monitoring and process optimisation. The same connectivity, however, exposes equipment that was engineered for closed networks to a hostile threat environment (Sadeghi et al., 2015). Many industrial components lack authentication or encryption, cannot be patched without halting production, and run on constrained hardware. Successful attacks on such systems can stop production, damage equipment and endanger human safety, which makes IIoT security a problem distinct in kind from enterprise IT security (Stouffer et al., 2023).

Intrusion detection is the control that carries the residual risk when prevention cannot be guaranteed. Traditional signature-based IDS detect only the attacks their rules encode; novel, polymorphic and obfuscated attacks pass unrecognised, and the rule base requires continuous expert maintenance (Khraisat et al., 2019). Anomaly-based detection promised to find previously unseen attacks but historically produced false-alarm volumes that operators could not sustain (Sommer & Paxson, 2010). AI-driven IDS answer both weaknesses by inferring what normal and hostile traffic look like from labelled records, with machine learning (ML) or deep learning (DL) algorithms; they now dominate the research agenda (Buczak & Guven, 2016; Alsoufi et al., 2021).

The published record of AI-driven IDS is large and, on its face, encouraging: accuracies above 99% are common. Yet these results are reached on different datasets, with different preprocessing, class-balancing, data splits and metrics, and several widely used benchmarks have since been shown to contain defects (Engelen et al., 2021; Lanvin et al., 2023). For a plant manager or security leader deciding whether to adopt an AI-based detector, the question is not whether some model achieved high accuracy on some dataset, but whether the evidence supports reliable detection with tolerable false alarms and computational cost in an industrial setting.

We take up that question by critically appraising published work on AI-driven intrusion detection for IIoT networks. The article offers four contributions:

1. It synthesises the theoretical foundations, IIoT security requirements and threat landscape that any industrial IDS must satisfy.
2. It contrasts classical ML and DL detectors: single trees and random forests, SVM classifiers, feed-forward ANNs, CNNs and LSTM networks.
3. It examines the benchmark datasets NSL-KDD, UNSW-NB15 and CICIDS2017, and the evaluation practices whose weaknesses limit what published results can establish.
4. It identifies five research gaps and proposes an evaluation agenda for future studies.

The rest of the article proceeds in this order. Section 2 explains how sources were gathered and analysed. Section 3 presents the background on IIoT architecture, threats and the evolution of IDS. Sections 4 and 5 review ML and DL approaches respectively. Section 6 examines benchmark datasets and evaluation practice. Section 7 tabulates the reviewed studies side by side, Section 8 states the open gaps together with an agenda for further work, and Section 9 closes the paper.

2. Review Methodology

This study is a narrative-critical review. It does not claim the status of a systematic review under formal reporting protocols such as PRISMA; instead, it aims to evaluate the strength and comparability of the evidence on a focused question.

Search and selection. Sources were located by repeated keyword queries in Google Scholar and the major publisher platforms of computing research. Queries combined terms for intrusion and anomaly detection, IIoT and ICS security, ML and DL classifiers, feature selection and class imbalance, together with the names of the benchmark datasets. Backward and forward citation tracing extended the set. Works were included when they (a) proposed or evaluated ML- or DL-based network intrusion detection for IoT, IIoT or industrial control settings; (b) surveyed such work; (c) supplied a foundational model or method used by such work; or (d) introduced or audited a benchmark dataset used in it. Bibliographic details were checked against the publisher's record; works whose identity could not be verified were excluded.

Scope. The final corpus spans 1987 to 2026. It comprises foundational works on intrusion detection and learning methods (e.g., Denning, 1987; Cortes & Vapnik, 1995; Hochreiter & Schmidhuber, 1997; Breiman, 2001; Chawla et al., 2002), surveys and reviews, empirical detection studies, and dataset papers together with their audits.

Analysis dimensions. Each empirical study was examined on six dimensions: the dataset(s) used, the models evaluated, the reported findings, the validation practice, the limitations acknowledged by its authors, and — where the material allowed — whether preprocessing and class balancing were applied in a leakage-safe manner. Accuracy figures from studies with different datasets, splits and preprocessing are reported with their context and are deliberately not ranked against one another.

3. Background: IIoT Security and the Evolution of Intrusion Detection

3.1 Security requirements of IIoT

Security objectives are conventionally expressed through the confidentiality–integrity–availability (CIA) triad. Industrial environments invert the enterprise ordering: operational technology (OT) typically places availability and integrity ahead of confidentiality, because the physical processes under control require continuous, timely and correct operation (Stouffer et al., 2023). Defence-in-depth layers segmentation, access control, monitoring and response so that no single failure compromises the system; within it, intrusion detection is the detective control that identifies what preventive controls have missed.

Two abstractions describe the industrial network. The three-layer IoT model distinguishes perception, network and application layers (Sadeghi et al., 2015). The Purdue reference model, used in NIST OT guidance, stratifies the estate from field devices (level 0) through controllers and supervisory systems to the enterprise zone, with communication restricted to defined conduits (Stouffer et al., 2023). Both imply that network-based IDS are best placed at conduits and demilitarised zones, where cross-zone traffic is security-critical and fully observable.

Three constraints recur throughout the literature. **Heterogeneity** of devices, protocols and firmware frustrates uniform endpoint protection and favours network-based monitoring. **Resource limits** on field and edge devices preclude heavy security processing and constrain where detection can run. **Legacy integration** means many vulnerabilities cannot be patched, so detection must compensate where prevention is impossible (Sadeghi et al., 2015; Stouffer et al., 2023). Industrial protocols add a further difficulty: Modbus transmits commands with no native authentication or encryption, while the security services of OPC UA and MQTT are frequently left disabled in constrained deployments. Where payloads are encrypted or proprietary, detection must rely on flow statistics, timing and header features.

From these sources, four requirements for an industrial IDS follow: it must (i) operate on network-derived features available without payload access or endpoint agents; (ii) respect the computational envelope of the tier on which it runs; (iii) keep false-positive rates low, because alarm fatigue in small

industrial security teams degrades the entire defence (Sommer & Paxson, 2010); and (iv) tolerate distribution shift across heterogeneous sites.

3.2 Threat landscape

IIoT threats fall into three broad groups. **Volumetric availability attacks** — denial of service (DoS) and distributed DoS — are especially consequential in availability-first environments. Kolias et al. (2017) explain how weakly protected IoT devices that still use factory passwords are recruited into botnets that launch such floods. Mirai, which Antonakakis et al. (2017) measured in detail, grew to roughly 600,000 infected devices by scanning for telnet services and trying a small dictionary of factory-default passwords; the compromised devices were then used for large flooding campaigns, among them the 2016 disruption of a major DNS provider. Although Mirai targeted consumer devices, it demonstrated mass compromise of the same class of embedded hardware on which industrial sensing depends.

Reconnaissance and brute-force access attacks produce distinctive connection counts, failure rates and timing patterns. **Zero-day, insider and multi-stage threats** are intrinsically harder: no signature exists for a zero-day exploit, insider traffic originates within the perimeter, and multi-stage campaigns combine individually innocuous steps. Khraisat et al. (2019) analyse how polymorphism and fragmentation defeat signature matching, motivating behaviour-based detection. In the anomaly taxonomy of Chandola et al. (2009), such campaigns are contextual or collective anomalies, detectable only by representations that relate records to one another over time.

3.3 Evolution of intrusion detection systems

Denning's (1987) intrusion-detection expert system proposed detecting intrusions as statistically significant deviations from learned profiles of normal behaviour — the conceptual ancestor of every learning-based detector. Deployed systems then diverged into two traditions. **Signature-based** systems, such as Snort-style rule engines, produce specific and interpretable alarms but detect only what they encode (Khraisat et al., 2019). **Anomaly-based** systems promised novelty detection but encountered two structural obstacles identified by Sommer and Paxson (2010): the *semantic gap* between statistical deviation and actual maliciousness, and the *base-rate problem*, whereby even a small false-positive rate produces overwhelming alarm volumes because benign traffic vastly outnumbers attacks. **Hybrid** designs combine both to cover their complementary weaknesses (Zarpelão et al., 2017).

IDS are also distinguished by vantage point. Host-based IDS monitor individual devices with high fidelity but require an agent on each one; network-based IDS observe traffic without endpoint cooperation, which is decisive in IIoT, where endpoints are heterogeneous, constrained and often sealed (Zarpelão et al., 2017). Machine learning entered this history as the means of realising anomaly detection at practical false-positive rates (Buczak & Guven, 2016). Whether it delivers that promise is the subject of Sections 4 and 5.

4. Classical ML Detectors

4.1 Theoretical basis of the main classical models

The **support vector machine** learns the hyperplane that maximises the margin between classes, possibly in a kernel-induced feature space; margin maximisation gives it a theoretical bound on generalisation error (Cortes & Vapnik, 1995). Its practical limits are equally well understood: kernel training cost grows with the number of examples, multiclass problems require ensembles of binary classifiers, and performance is sensitive to unscaled and irrelevant features. **Decision trees** partition the feature space by recursive splits and yield human-readable rules. Their ensemble extension, the **random forest**, aggregates trees grown on bootstrap samples with randomised feature choice and is comparatively robust to noise (Breiman, 2001). For these reasons, tree models serve as the usual reference point when other IDS learners are assessed.

4.2 Empirical record

Early evaluations established feasibility. Hodo et al. (2016) trained a multi-layer perceptron on packet traces from a simulated IoT network and showed that it recognised denial-of-service traffic, both single-source and distributed. Mondal and Singh (2022) ran several ML classifiers on IoT intrusion data, and tree-based models (random forests and single trees) scored highest — a pattern that recurs throughout the literature. Surveys of the period by da Costa et al. (2019) and Gupta et al. (2021) emphasised real-time operation and adaptability as requirements for IoT deployment.

Feature selection. Flow-derived datasets are high-dimensional, redundant and uneven in scale, and classifier performance depends strongly on how features are treated. Disha and Waheed (2022) proposed a Gini-impurity-based weighted random forest (GIWRF) selector that identified 20 salient features on UNSW-NB15 and 10 on TON-IoT; the single tree achieved its top results on the reduced feature set, and its TON-IoT accuracy rose from 99.50% to 99.90%. Their study addressed binary classification only and did not analyse the cost of the selection stage. Three distinct operations are often conflated under

"feature selection": scaling and encoding, subset selection, and feature construction. Reported gains attach to the specific combination each study used, and the evidence reviewed here leaves open whether a feature core chosen on one dataset carries over to another. A further validity issue is rarely documented: feature selection performed on the full dataset before the train–test split leaks test information into training.

Class imbalance. Attack records are often outnumbered by benign traffic, and within attacks a few families dominate. The Synthetic Minority Over-sampling Technique (SMOTE) creates new minority examples by interpolating between nearest minority neighbours rather than duplicating them (Chawla et al., 2002). On the WUSTL-IIoT-2021 dataset, Eid et al. (2024) compared preprocessing pipelines and six balancing techniques across six classical models and reported systematic improvements, with random forest giving the best performance-to-cost ratio. Other work replaces interpolation with generative models; PWG-IDS, for example, uses a GAN to rebalance IIoT data (Zhang et al., 2021). Three cautions apply: interpolated samples may fall outside the true class region; over-sampling before the split leaks test information; and aggregate scores can conceal minority-class failure. Panahi (2026) found random forest on UNSW-NB15 achieving aggregate F1 above 0.95 while still producing high false-negative rates on minority attack classes.

Interpretability and cost. Decision trees are the most interpretable detectors; SVM decisions occur in transformed spaces that resist explanation. Computational cost is increasingly measured rather than assumed. Jamshidi et al. (2025) evaluated seven ML-based IDS under edge deployment and found statistically significant differences in CPU load and energy consumption under active attack — evidence that model choice has consequences for constrained IIoT platforms, although the differences attach to the configurations tested rather than to model families in general.

Industrial and distributed settings. Gauthama Raman et al. (2021) evaluated ML anomaly detectors for industrial control systems (ICS) and documented recurring obstacles: data scarcity, label quality and the cost of validation. The surveys of Umer et al. (2022) and Termanini et al. (2025) converge on the same unresolved challenges and note that much ICS research still validates on legacy datasets whose attack taxonomies predate modern deployments. For distributed estates, Chatterjee and Hanawal (2022) proposed a federated hybrid ensemble (k-nearest neighbours with random forest) that improved the true-positive/false-positive trade-off on NSL-KDD.

4.3 Synthesis

The classical ML record supports four conclusions. First, tree-based learners are consistently strong on tabular flow features (Mondal & Singh, 2022; Disha & Waheed, 2022; Eid et al., 2024; Panahi, 2026). Second, SVM performance is highly dependent on feature treatment. Third, the effect of class balancing on per-class detection is not settled: improvements are reported under specific protocols, but no reviewed study isolates the effect through a controlled ablation. Fourth, validation practice varies so widely that results across studies cannot be ranked.

5. Deep Neural Detectors

5.1 Architectures and their assumptions

Deep models stack many layers, each building a more abstract description of the input than the one before (LeCun et al., 2015). Each architecture carries an inductive bias from its original domain. **Feed-forward ANNs** learn general non-linear mappings. **CNNs** exploit spatially local, translation-invariant structure. **LSTM networks** use gating to preserve information across long sequences, overcoming the vanishing-gradient problem of earlier recurrent networks (Hochreiter & Schmidhuber, 1997). The key question for intrusion detection is whether network data actually offer this structure. Tabular flow records are unordered attribute vectors — the natural domain of tree ensembles. Temporal structure exists in streams of packets and flows, but exploiting it requires inputs to be built as time-ordered sequences of records.

5.2 Empirical record

Kim et al. (2016) applied an LSTM classifier to the KDD Cup 1999 data, an early demonstration that sequential deep models could perform intrusion detection. Amarasinghe and Manic (2018) built a deep neural network IDS on NSL-KDD for binary and multiclass tasks and found that architectures of different depth reached similar accuracy while relying on different features. They proposed reporting the features most relevant to each decision as a way of supporting operator trust — a point that matters for critical infrastructure, where alarms must be explainable. Idrissi et al. (2021) built a convolutional botnet detector and tested it on Bot-IoT; the CNN outperformed simple recurrent, LSTM and GRU variants, with a reported validation accuracy of 99.94% and a prediction time below 0.34 ms. This was a comparison among deep architectures, not a test against classical baselines. The systematic review by Alsoufi et al. (2021) catalogues the DL approaches applied to anomaly-based IoT detection and their evaluation practices.

Cautionary evidence stands against these results. In a multi-dataset evaluation across UNSW-NB15, CICIDS2017, NSL-KDD and ToN-IoT with over five million records, Panahi (2026) found LSTM-CNN hybrids performing very well on favourable data but prone to overfitting; their performance fell materially once near-duplicate test flows were removed, while random forest retained the best balance of detection, throughput and model size. Jamshidi et al. (2025) add the cost dimension: detector configurations differ significantly in CPU and energy use at the edge.

5.3 The sequential-input problem

The most consequential methodological issue in the DL literature is how inputs are constructed. In some studies, flow or packet records are ordered in time and presented as genuine sequences, engaging the temporal strengths of CNN and LSTM models. In others, a single tabular feature vector is simply reshaped into a matrix and fed to the same architectures — an arrangement with no temporal meaning. The second practice does not test sequential capability, yet studies frequently do not state which practice they used. This ambiguity is one reason DL-versus-ML comparisons in the literature disagree.

5.4 Synthesis

The evidence justifies including ANN, CNN and LSTM models in IIoT detection research, but not the expectation that they will dominate. Their strengths — learning from lightly engineered features, exploiting sequence structure when it genuinely exists, and fast inference in some designs — are balanced by documented costs: data hunger, overfitting on duplicated or favourable data, and computational weight. Whether deep models hold a real advantage for complex, multi-stage attacks under genuine sequence construction remains an open question.

6. Benchmark Datasets and Evaluation Practice

Almost every result reviewed above rests on a small set of public benchmark datasets. Their attack taxonomies, class distributions and documented defects therefore condition every performance figure the field publishes.

6.1 General-purpose benchmarks

NSL-KDD (Tavallaee et al., 2009) refines the KDD Cup 1999 data by removing redundant duplicate records and rebalancing record difficulty. It retains four attack categories — DoS, probing, remote-to-local and user-to-root — derived from the 1998 DARPA evaluation. Its attack repertoire is therefore

dated, but it remains the field's most established common reference, which keeps new results comparable with a large body of prior work.

UNSW-NB15 (Moustafa & Slay, 2015) was generated with the IXIA PerfectStorm tool as a hybrid of real normal traffic and synthetic attacks across nine families: analysis, backdoor, DoS, exploits, fuzzers, generic, reconnaissance, shellcode and worms. Its finer taxonomy makes it useful for per-class evaluation, but attacks are tool-generated and the class composition is heavily skewed. Panahi (2026) also reports near-duplicate flows between its training and test partitions.

CICIDS2017 (Sharafaldin et al., 2018) was captured on a testbed with profiled benign behaviour and attacks including brute force, DoS/DDoS, web attacks, infiltration, port scans and botnet traffic, with features extracted by CICFlowMeter. A later audit literature documented material defects. Engelen et al. (2021) found corrupted protocol and state features, violations of TCP state-machine rules and duplicate records. Lanvin et al. (2023) reported inconsistent labelling and showed that these errors make significant differences to measured detection performance. The dataset remains usable, but results on it require explicit acknowledgement of these defects.

6.2 IIoT-native datasets

Purpose-built industrial datasets show what realistic evaluation would require. Edge-IIoTset (Ferrag et al., 2022) captures traffic from a seven-layer IIoT testbed spanning cloud, fog, edge and perception tiers, with industrial protocols such as Modbus and MQTT and 14 attack types in five threat families. X-IIoTID (Al-Hawawreh et al., 2022) supplies IIoT intrusion records that do not depend on any particular connectivity type or device. Their existence sharpens a central limitation: NSL-KDD, UNSW-NB15 and CICIDS2017 are not industrial datasets, so results on them support conclusions about model behaviour under controlled conditions, not direct claims about operational IIoT performance.

6.3 Three recurring evaluation weaknesses

1. **Representativeness.** Closed-world evaluation on static datasets overstates field performance, because operational traffic is non-stationary and semantically richer (Sommer & Paxson, 2010). The dataset audits add that benchmark data themselves contain generation and labelling errors (Engelen et al., 2021; Lanvin et al., 2023), and industrial surveys note continued reliance on legacy datasets (Gauthama Raman et al., 2021; Termanini et al., 2025).
2. **Protocol validity.** Scaling, feature selection or over-sampling applied before the train–test split lets test information influence training and inflates results. Duplicates have the same effect: once Panahi (2026) stripped near-duplicate records out of the UNSW-NB15 test partition, one deep

model's F1 fell noticeably. Many studies do not document whether their pipeline was leakage-safe.

3. **Metric adequacy.** Under heavy class imbalance, aggregate accuracy is close to uninformative: a detector can score highly while missing most minority attacks. Per-class precision, recall and F1, together with false-positive and false-negative rates, are needed to reveal the failures that matter operationally (Sommer & Paxson, 2010; Panahi, 2026).

The common thread is that headline accuracy, the figure most often reported, is the least informative about industrial readiness. The false-positive rate, which determines analyst workload and the risk of disrupting production, is reported far less consistently.

7. Comparative Summary of Reviewed Studies

Tables 1–3 summarise the reviewed evidence. Where a study did not report a detail in the material reviewed, the table says so rather than inferring it. Accuracy figures from studies with different datasets, splits and preprocessing are not comparable and are not ranked.

Table 1. Empirical studies of ML/DL-based intrusion detection

Study	Dataset(s)	Models	Reported findings	Validation practice	Limitations
Hodo et al. (2016)	Packet traces; simulated IoT network	MLP (ANN)	ANN detected DoS/DDoS attacks	Train/test on captured and simulated data	Early study; limited scope
Kim et al. (2016)	KDD Cup 1999	LSTM classifier	Feasibility of LSTM-based IDS shown	Supervised training on KDD99	Dated dataset; single architecture
Amarasinghe & Manic (2018)	NSL-KDD	DNN (MLP-based)	Similar accuracy across depths; feature-relevance feedback for operator trust	Binary and multiclass	Single dataset
Idrissi et al. (2021)	Bot-IoT	CNN vs. RNN, LSTM, GRU	CNN best: 99.94% validation accuracy, <0.34 ms prediction	Train/validation split	Single dataset; botnet focus
Zhang et al. (2021)	IIoT setting	GAN-based IDS (PWG-IDS)	GAN proposed for class imbalance	Not reported in material reviewed	Preprint
Chatterjee & Hanawal (2022)	NSL-KDD	Federated KNN+RF ensemble	Improved TPR/FPR trade-off	Not reported in material reviewed	Binary classification focus
Disha & Waheed (2022)	UNSW-NB15, TON-IoT	DT, MLP, AdaBoost, GBT, GRU, LSTM + GIWRF	DT best with selected features; TON-IoT accuracy 99.50% → 99.90%	Binary classification	Selection cost not analysed

Study	Dataset(s)	Models	Reported findings	Validation practice	Limitations
Mondal & Singh (2022)	IoT intrusion data	RF, DT, other ML	RF and DT performed effectively	Not reported in material reviewed	Conference-scale evaluation
Eid et al. (2024)	WUSTL-IIoT-2021	RF, DT, KNN, LR, SVM, NB + balancing	Preprocessing and balancing improved results; RF best	Preprocessing/balancing comparison	Single IIoT dataset
Jamshidi et al. (2025)	IoT IDS at the edge	7 ML-based IDS	Significant CPU and energy differences under attack	ANOVA-based comparison	Edge-deployment focus
Panahi (2026)	UNSW-NB15, CICIDS2017, NSL-KDD, ToN-IoT	RF, IF, LOF, OC-SVM, LSTM-CNN, autoencoder	RF best performance–cost balance; DL prone to overfitting; near-duplicates inflated results	Multi-dataset; duplicate analysis	Results depend on dataset properties

Table 2. Surveys, reviews and foundational works

Work	Scope	Contribution
Denning (1987)	Foundational IDS model	Origin of anomaly-based detection
Chandola et al. (2009)	Anomaly detection	Point, contextual and collective anomaly taxonomy
Sommer & Paxson (2010)	ML for network IDS	Semantic gap, base-rate problem, evaluation validity
Buczak & Guven (2016)	ML/data mining for cyber security	Algorithm families and evaluation issues
Zarpelão et al. (2017)	IDS for IoT	Classification by technique, placement and validation
Khraisat et al. (2019)	IDS techniques and datasets	Signature/anomaly limitations; evasion
da Costa et al. (2019)	ML-based IoT IDS	Survey of ML approaches for IoT
Alsoufi et al. (2021)	DL-based anomaly IDS in IoT	Systematic review of DL approaches
Gauthama Raman et al. (2021)	ML for ICS intrusion detection	Industrial challenges; experimental lessons
Gupta et al. (2021)	ML-based IoT IDS	Real-time and adaptability requirements
Umer et al. (2022)	ML for IDS in ICS	Deployment challenges and recommendations
Termanini et al. (2025)	ML-based IDS in ICS	Unresolved challenges in industrial settings

Table 3. Benchmark datasets and their documented limitations

Dataset	Source	Attack coverage	Documented limitations
NSL-KDD	Tavallaee et al. (2009)	DoS, probing, R2L, U2R	Derived from 1998 DARPA simulation; dated attacks
UNSW-NB15	Moustafa & Slay (2015)	Nine families, incl. exploits, backdoor, reconnaissance	Hybrid real/synthetic generation; skewed classes; near-duplicates (Panahi, 2026)
CICIDS2017	Sharafaldin et al. (2018)	Brute force, DoS/DDoS, web, infiltration, port scan, botnet	Feature corruption and labelling errors (Engelen et al., 2021; Lanvin et al., 2023)
Edge-IIoTset	Ferrag et al. (2022)	14 attacks in 5 families on IIoT testbed	Industrial protocols; less established as a common reference

Dataset	Source	Attack coverage	Documented limitations
X-IIoTID	Al-Hawawreh et al. (2022)	Connectivity- and device-agnostic IIoT intrusions	Less established as a common reference

8. Research Gaps and Future Directions

Comparing the reviewed works reveals five gaps. Each is inferred by setting studies against one another, not from any claim about how widespread a practice is.

Gap 1 — No common-protocol comparison of AI and rule-based detection. Strong results exist for both tree ensembles (Mondal & Singh, 2022; Disha & Waheed, 2022; Eid et al., 2024; Panahi, 2026) and deep models (Kim et al., 2016; Idrissi et al., 2021). Because those numbers come from differing datasets, preprocessing, balancing, data splits and metrics, they cannot be set against one another. Few studies compare AI detectors against a conventional rule-based baseline on the same data and protocol, so the core adoption question — does AI detect more intrusions, with fewer false alarms, than the systems it would replace? — remains open.

Gap 2 — False-positive behaviour is under-reported. In industrial environments, false alarms consume scarce analyst time and can trigger disruptive responses. Yet false-positive rates are reported far less consistently than accuracy, and almost never with confidence intervals that would allow a meaningful comparison.

Gap 3 — The transferability of feature-selection gains is unknown. Feature selection improves results in specific studies, but those studies differ in datasets, selection methods and models. Whether the gains hold across model families and datasets under matched conditions has not been tested.

Gap 4 — The DL advantage for complex attacks is unresolved. Theory suggests sequential architectures should detect multi-stage attacks better, but empirical DL studies often do not state whether inputs were genuine time-ordered sequences or reshaped tabular vectors. Whether deep models hold a measurable advantage for complex attacks under genuine sequence construction is unknown.

Gap 5 — The effect of class balancing is unmeasured. Balancing is reported to help under particular protocols (Eid et al., 2024), and aggregate metrics are known to hide minority-class errors (Panahi, 2026), but no reviewed study isolates the effect of balancing through a controlled ablation within a leakage-safe pipeline.

8.1 An evaluation agenda

These gaps suggest six practices that later work ought to treat as routine:

1. **Leakage-safe pipelines.** Fit scaling, encoding, feature selection and over-sampling on training folds only, and check for duplicates and near-duplicates across partitions.
2. **Operationally meaningful metrics.** Report per-class precision, recall and F1, false-positive and false-negative rates and confusion matrices alongside accuracy, with paired confidence intervals for model comparisons.
3. **Multi-dataset evaluation.** Test every model on several datasets within one protocol, report results per dataset, and acknowledge each dataset's documented defects.
4. **A real baseline.** Compare AI detectors against a clearly specified rule-based or signature-based comparator on identical data.
5. **Cost measurement.** Record training time, inference latency and resource use, since edge-constrained IIoT platforms cannot host every model.
6. **Movement towards industrial data.** Validate promising models on IIoT-native datasets such as Edge-IIoTset and X-IIoTID and, ultimately, on real plant traffic.

8.2 Future research directions

Beyond evaluation practice, the literature points to several research directions: deployment of detectors at the edge within the computational budget of industrial gateways; federated training across IIoT sites that are unable to pool their raw data (Chatterjee & Hanawal, 2022); protocol-aware detection for Modbus, OPC UA and MQTT traffic; hybrid architectures that combine signature precision with learned anomaly detection; explainable detection that tells operators why an alarm was raised (Amarasinghe & Manic, 2018); and the development of large, well-labelled, IIoT-native datasets.

9. Conclusion

AI-driven intrusion detection is a well-founded response to the security problem of IIoT networks, where unpatchable, heterogeneous and resource-constrained devices make prevention alone insufficient and signature matching lags behind emerging attacks. The material examined here shows, though, that the published evidence is weaker than its headline numbers imply.

Three conclusions stand out. First, tree-based classical learners are consistently strong on tabular network data and offer the best balance of accuracy, interpretability and cost in several controlled comparisons; deep models are valuable but their advantage is conditional on data quality and on genuine sequence construction. Second, feature treatment and class balancing materially change reported performance, yet their effects are rarely isolated under leakage-safe conditions. Third, the field's

evaluation practice — dated or defective benchmarks, leakage and near-duplicates, and reliance on aggregate accuracy — means that many reported results cannot be compared with one another or taken as evidence of industrial readiness.

For practitioners and managers deciding whether to adopt AI-based detection, the implication is to ask for evidence beyond accuracy: false-positive rates on representative traffic, per-class detection of the attacks that matter to the plant, and the computational cost on the intended hardware. For researchers, the evaluation agenda in Section 8.1 offers a path to results that are comparable, reproducible and relevant to industrial deployment. The limitations of this review should also be noted: it is a narrative-critical rather than a systematic review, and its corpus emphasises studies relevant to network-based, supervised detection on benchmark data.

Declarations

Conflict of interest: Neither author has any financial or personal relationship that could have influenced this work.

Funding: The work was carried out without external financial support.

Ethical approval: Not required, as no human participants or personal data were involved; the work draws only on published sources.

Declaration on the use of AI: An AI writing assistant (Claude, Anthropic) was used to support drafting, language editing and formatting of this manuscript. The authors reviewed, verified and edited all content and take full responsibility for the work.

Author contributions: Mungara Dhaval Vithalbhai: study design, survey of sources and preparation of the first draft. Dr. M K Agarwal: academic guidance and critical revision of the manuscript.

References

1. Al-Hawawreh, M., Sitnikova, E., & Aboutorab, N. (2022). X-IIoTID: A connectivity-agnostic and device-agnostic intrusion data set for industrial Internet of Things. *IEEE Internet of Things Journal, 9*(5), 3962–3977. https://doi.org/10.1109/JIOT.2021.3102056

2. Alsoufi, M. A., Razak, S., Md Siraj, M., Nafea, I., Ghaleb, F. A., Saeed, F., & Nasser, M. (2021). Anomaly-based intrusion detection systems in IoT using deep learning: A systematic literature review. *Applied Sciences, 11*(18), 8383. https://doi.org/10.3390/app11188383

3. Amarasinghe, K., & Manic, M. (2018). Improving user trust on deep neural networks based intrusion detection systems. In *IECON 2018 – 44th Annual Conference of the IEEE Industrial Electronics Society* (pp. 3262–3268). IEEE. https://doi.org/10.1109/IECON.2018.8591322

4. Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J. A., Invernizzi, L., Kallitsis, M., Kumar, D., Lever, C., Ma, Z., Mason, J., Menscher, D., Seaman, C., Sullivan, N., Thomas, K., & Zhou, Y. (2017). Understanding the Mirai botnet. In *26th USENIX Security Symposium* (pp. 1093–1110). USENIX Association. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis

5. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

6. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials, 18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

7. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3), Article 15. https://doi.org/10.1145/1541880.1541882

8. Chatterjee, S., & Hanawal, M. K. (2022). Federated learning for intrusion detection in IoT security: A hybrid ensemble approach. *International Journal of Internet of Things and Cyber-Assurance, 2*(1), 62–86. https://doi.org/10.1504/IJITCA.2022.124372

9. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

11. da Costa, K. A. P., Papa, J. P., Lisboa, C. O., Munoz, R., & de Albuquerque, V. H. C. (2019). Internet of Things: A survey on machine learning-based intrusion detection approaches. *Computer Networks, 151*, 147–157. https://doi.org/10.1016/j.comnet.2019.01.023

12. Denning, D. E. (1987). An intrusion-detection model. *IEEE Transactions on Software Engineering, SE-13*(2), 222–232. https://doi.org/10.1109/TSE.1987.232894

13. Disha, R. A., & Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. *Cybersecurity, 5*(1), Article 1. https://doi.org/10.1186/s42400-021-00103-8

14. Eid, A. M., Soudan, B., Bou Nassif, A., & Injadat, M. N. (2024). Comparative study of ML models for IIoT intrusion detection: Impact of data preprocessing and balancing. *Neural Computing and Applications, 36*, 6955–6972. https://doi.org/10.1007/s00521-024-09439-x

15. Engelen, G., Rimmer, V., & Joosen, W. (2021). Troubleshooting an intrusion detection dataset: The CICIDS2017 case study. In *2021 IEEE Security and Privacy Workshops (SPW)* (pp. 7–12). IEEE. https://doi.org/10.1109/SPW53761.2021.00009

16. Ferrag, M. A., Friha, O., Hamouda, D., Maglaras, L., & Janicke, H. (2022). Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. *IEEE Access, 10*, 40281–40306. https://doi.org/10.1109/ACCESS.2022.3165809

17. Gauthama Raman, M. R., Ahmed, C. M., & Mathur, A. (2021). Machine learning for intrusion detection in industrial control systems: Challenges and lessons from experimental evaluation. *Cybersecurity, 4*, Article 27. https://doi.org/10.1186/s42400-021-00095-5

18. Gupta, P., Yadav, L., & Tomar, D. S. (2021). Internet of Things: A comprehensive study on machine learning-based intrusion detection approaches. In *International Conference on Women Researchers in Electronics and Computing* (AIJR Proceedings, pp. 208–218). AIJR Publisher. https://doi.org/10.21467/proceedings.114.29

19. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

20. Hodo, E., Bellekens, X., Hamilton, A., Dubouilh, P.-L., Iorkyase, E., Tachtatzis, C., & Atkinson, R. (2016). Threat analysis of IoT networks using artificial neural network intrusion detection system. In *2016 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE. https://doi.org/10.1109/ISNCC.2016.7746067

21. Idrissi, I., Boukabous, M., Azizi, M., Moussaoui, O., & El Fadili, H. (2021). Toward a deep learning-based intrusion detection system for IoT against botnet attacks. *IAES International*

Journal of Artificial Intelligence, 10(1), 110–120. https://doi.org/10.11591/ijai.v10.i1.pp110-120

22. Jamshidi, S., Nafi, K. W., Nikanjam, A., & Khomh, F. (2025). Evaluating machine learning-driven intrusion detection systems in IoT: Performance and energy consumption. *Computers & Industrial Engineering, 204*, Article 111103. https://doi.org/10.1016/j.cie.2025.111103

23. Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity, 2*, Article 20. https://doi.org/10.1186/s42400-019-0038-7

24. Kim, J., Kim, J., Le, T.-T.-H., & Kim, H. (2016). Long short term memory recurrent neural network classifier for intrusion detection. In *2016 International Conference on Platform Technology and Service (PlatCon)* (pp. 1–5). IEEE. https://doi.org/10.1109/PlatCon.2016.7456805

25. Kolias, C., Kambourakis, G., Stavrou, A., & Voas, J. (2017). DDoS in the IoT: Mirai and other botnets. *Computer, 50*(7), 80–84. https://doi.org/10.1109/MC.2017.201

26. Lanvin, M., Gimenez, P. F., Han, Y., Majorczyk, F., Mé, L., & Totel, É. (2023). Errors in the CICIDS2017 dataset and the significant differences in detection performances it makes. In *Proceedings of the 17th International Conference on Risks and Security of Internet and Systems (CRiSIS 2022)*. Springer. https://doi.org/10.1007/978-3-031-31108-6_2

27. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

28. Mondal, B., & Singh, S. K. (2022). A comparative analysis of network intrusion detection system for IoT using machine learning. In *Internet of Things and Its Applications: Select Proceedings of ICIA 2020* (pp. 211–221). Springer. https://doi.org/10.1007/978-981-16-7637-6_19

29. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. In *2015 Military Communications and Information Systems Conference (MilCIS)* (pp. 1–6). IEEE. https://doi.org/10.1109/MilCIS.2015.7348942

30. Panahi, U. (2026). Assessing the strength of machine learning and deep learning models for network anomaly detection: Insights from multi-dataset cybersecurity analysis. *The Journal of Supercomputing, 82*, Article 399. https://doi.org/10.1007/s11227-026-08556-9

31. Sadeghi, A.-R., Wachsmann, C., & Waidner, M. (2015). Security and privacy challenges in industrial Internet of Things. In *Proceedings of the 52nd Annual Design Automation Conference (DAC '15)* (pp. 1–6). ACM. https://doi.org/10.1145/2744769.2747942

32. Sharafaldin, I., Habibi Lashkari, A., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy* (pp. 108–116). SciTePress. https://doi.org/10.5220/0006639801080116

33. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy* (pp. 305–316). IEEE. https://doi.org/10.1109/SP.2010.25

34. Stouffer, K., Pease, M., Tang, C. Y., Zimmerman, T., Pillitteri, V., Lightman, S., Hahn, A., Saravia, S., Sherule, A., & Thompson, M. (2023). *Guide to operational technology (OT) security* (NIST Special Publication 800-82 Rev. 3). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-82r3

35. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)* (pp. 1–6). IEEE. https://doi.org/10.1109/CISDA.2009.5356528

36. Termanini, A., Al-Abri, D., Bourdoucen, H., & Al Maashri, A. (2025). Using machine learning to detect network intrusions in industrial control systems: A survey. *International Journal of Information Security, 24*, Article 20. https://doi.org/10.1007/s10207-024-00916-x

37. Umer, M. F., Shah, M. A., & Khan, A. (2022). Machine learning for intrusion detection in industrial control systems: Applications, challenges, and recommendations. *International Journal of Critical Infrastructure Protection, 38*, Article 100516. https://doi.org/10.1016/j.ijcip.2022.100516

38. Zarpelão, B. B., Miani, R. S., Kawakani, C. T., & de Alvarenga, S. C. (2017). A survey of intrusion detection in Internet of Things. *Journal of Network and Computer Applications, 84*, 25–37. https://doi.org/10.1016/j.jnca.2017.02.009

39. Zhang, L., Jiang, S., Shen, X., Gupta, B. B., & Tian, Z. (2021). *PWG-IDS: An intrusion detection model for solving class imbalance in IIoT networks using generative adversarial networks* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2110.03445