Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deployable AI for IoT/IIoT Security: A Systematic Review of Labeling, Transfer, Resource, and Explainability Constraints

Anass Rharif

anass.rharif@usms.ac.ma

Sultan Moulay Slimane University

Ziad Charafi

Sultan Moulay Slimane University

Mounia Zaydi

Université Catholique de Lille

Sofia Belkhala

Sultan Moulay Slimane University

Yassine Maleh

Sultan Moulay Slimane University

Systematic Review

Keywords:

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9929399/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Deployable AI for IoT/IIoT Security: A Systematic Review of Labeling, Transfer, Resource, and Explainability Constraints

Anass Rharif[1,*], Ziad Charafi[1], Mounia Zaydi[2], Sofia Belkhala[1], and Yassine Maleh[1]

[1]Laboratory LaSTI, ENSAK, Sultan Moulay Slimane University, Khouribga, 54000, Morocco
[2]ICL, Junia, Université Catholique de Lille, LITL, Lille, 59000, France
[*]anass.rharif@usms.ac.ma

ABSTRACT

Artificial intelligence (AI) is increasingly applied to secure Internet of Things (IoT) and Industrial Internet of Things (IIoT) systems, yet evidence from benchmark experiments may not transfer to operational settings where labels are scarce, device behaviour changes, and deployments face latency, resource, explainability, and auditability constraints. This systematic literature review synthesises 154 peer reviewed studies published between January 2021 and November 2025 to examine security task coverage, AI family adoption, evaluation rigour, deployment limitations, and future research directions. The evidence is concentrated in intrusion and anomaly detection, while identity assurance, insider threat detection, threat response, and root cause localisation remain comparatively underexplored. Across AI families, reported accuracy is less informative than external validity, adaptation effort, runtime cost, robustness, and operator trust. We identify a recurring need for standardised train-here deploy-there evaluation protocols, drift aware validation, edge oriented model design, robustness testing, and explainability validated with operational users. These findings support a shift from benchmark centred performance reporting toward deployment grade assurance for AI enabled IoT/IIoT security.

Introduction

The Internet of Things (IoT) has become a foundational layer of modern digital infrastructure, connecting consumer devices, medical systems, industrial controllers, transportation platforms, smart city services, and payment environments into large scale cyber physical ecosystems. This connectivity enables automation, sensing, monitoring, and data-driven operations, but it also expands the attack surface and increases the potential consequences of compromise. Many IoT and Industrial Internet of Things (IIoT) devices operate with limited processing capacity, memory, power, and update mechanisms, and often remain deployed for years in heterogeneous, partially managed, and safety sensitive environments. As a result, persistent weaknesses such as hard coded credentials, weak authentication, outdated firmware, insecure default configurations, fragile protocols, and limited monitoring continue to enable botnets, distributed denial of service attacks, data exfiltration, fraud, lateral movement, and disruption of critical operations[1–5].

Traditional security controls remain necessary, but they are not sufficient for the scale and complexity of IoT/IIoT environments. Signature based intrusion detection systems, static firewall rules, and conventional authentication mechanisms are largely reactive and are typically designed around known attack patterns. They often struggle with the volume, velocity, and heterogeneity of IoT telemetry, the diversity of device protocols, and adversaries that adapt through stealthy, low and slow, or multi stage behaviours. In operational settings, these challenges are amplified by incomplete asset visibility, legacy devices, multi vendor platforms, fragmented ownership, and limited security operations dedicated specifically to IoT and IIoT environments[4,7,10,13].

Artificial intelligence (AI), including machine learning, deep learning, federated learning, reinforcement learning, adversarial learning, explainable AI, and hybrid approaches, has therefore become a prominent research direction for IoT/IIoT security. AI-based methods are increasingly used to model device behaviour, detect anomalous traffic, classify malicious flows, support device fingerprinting, identify malware or botnet activity, and enable privacy preserving analytics across distributed environments[5,6]. However, the maturity of these methods cannot be assessed only through reported detection accuracy. For deployment, AI-based security methods must also be evaluated in relation to robustness, resource cost, transferability, explainability, privacy, governance, and integration into operational security workflows.

Strong benchmark performance does not automatically imply deployment readiness. Many AI-based IoT/IIoT security studies report high accuracy, recall, or F1-score on curated datasets, yet operational environments are noisy, incomplete, non-independent and identically distributed, and continuously changing[8,9]. Device fleets evolve, firmware versions change,

traffic patterns drift, and new applications or adversarial behaviours emerge over time. Edge devices may not support large models, while distributed learning environments introduce additional constraints related to communication overhead, privacy, trust, poisoning resistance, and governance[8,11]. The central question is therefore no longer only whether AI can improve detection performance, but whether AI-based methods provide evidence strong enough to support real-world deployment decisions.

Existing surveys provide valuable insights, but the literature remains fragmented across task-centric, paradigm-centric, and domain-centric perspectives. Task-centric reviews often focus on intrusion detection or anomaly detection; paradigm-centric reviews examine federated learning, blockchain, edge intelligence, explainable AI, or adversarial learning; and domain-centric reviews focus on healthcare IoT, smart homes, Internet of Vehicles, smart cities, or industrial systems[13]. These perspectives are useful, but they rarely provide an integrated view of how AI technique families map to IoT/IIoT security tasks, how evaluation practices affect the credibility of reported results, and which deployment constraints limit practical adoption. A detailed comparison of representative prior surveys and the positioning of this review is provided in Supplementary Table S1.

This systematic literature review addresses this gap by synthesizing 154 peer-reviewed Q1/Q2 journal articles published between January 2021 and November 2025. Journal quartile was used only as a corpus-boundary criterion, while study-level quality was assessed separately. The 2021–2025 window captures the recent acceleration of AI based IoT/IIoT security research, including growth in federated learning, edge intelligence, adversarial robustness, explainability, and privacy-preserving security analytics. The review analyzes AI-based IoT/IIoT security across technique families, security tasks, datasets, evaluation practices, quality evidence, and deployment-oriented constraints. Rather than treating AI performance as an isolated benchmark outcome, the review interprets the evidence through operational pressures including labeling burden, cross-context transfer, resource feasibility, false-positive behaviour, robustness, explainability, privacy, and governance.

Figure 1 summarizes the review logic and shows how the five research questions structure the synthesis.

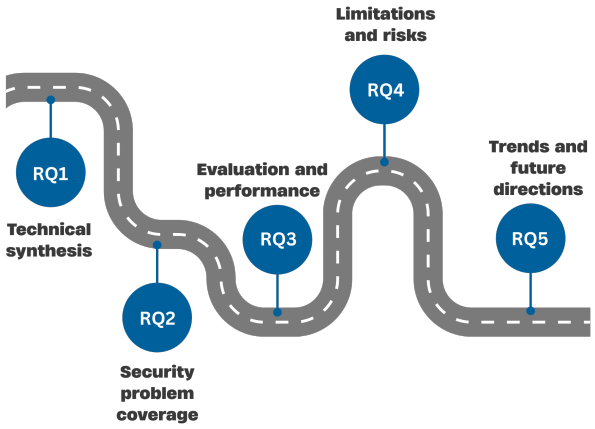


Figure 1. Review roadmap linking the five research questions to the technical synthesis, security-problem coverage, evaluation and performance analysis, limitations and risks, and future research directions.

This review addresses five research questions:

- *RQ1: Technical synthesis.* Which AI technique families are used for IoT/IIoT security, and how do they operate under IoT-specific constraints such as resource limitations, decentralization, heterogeneity, and streaming telemetry?
- *RQ2: Security-problem coverage.* Which IoT/IIoT security problems are well covered by AI-based approaches, and which operationally important areas remain comparatively underexplored?
- *RQ3: Evaluation and performance.* How are AI-based IoT/IIoT security systems evaluated in terms of datasets, metrics, baselines, and validation strategies, and what do the reported results imply for operational reliability and risk reduction?
- *RQ4: Limitations and risks.* What practical limitations and risk mechanisms hinder real-world adoption, including computational overhead, dependence on labeled data, non-independent and identically distributed data, concept drift, adversarial vulnerability, explainability gaps, and business or regulatory constraints?
- *RQ5: Trends and future directions.* Which research directions are most likely to improve deployability, including trustworthy and explainable AI, drift-aware learning, lightweight edge models, adversarial robustness, privacy-preserving collaboration, and governance-aligned security operations?

This paper makes five contributions. First, it provides a PRISMA-guided evidence base of 154 Q1/Q2 journal articles on AI based IoT/IIoT security. Second, it develops a unified taxonomy linking AI technique families to IoT/IIoT security functions, including intrusion detection, anomaly detection, malware and botnet detection, device authentication, privacy preserving analytics, industrial protection, and adversarial robustness. Third, it maps the distribution of AI families across security problems to identify dominant research concentrations and underexplored operational areas. Fourth, it evaluates deployment-oriented evidence by examining datasets, metrics, validation practices, quality indicators, and constraints related to labeling, transfer, drift, real-time feasibility, false-positive behaviour, explainability, privacy, and governance. Fifth, it synthesizes an actionable maturity agenda for deployable AI-based IoT/IIoT security, emphasizing transfer-aware validation, drift-aware learning, edge feasibility, privacy-preserving collaboration, adversarial robustness, explainable decision support, and governance-aware lifecycle management.

Methods

This systematic literature review was conducted in accordance with PRISMA principles to ensure transparency, reproducibility, and traceability across identification, screening, eligibility assessment, inclusion, quality appraisal, and synthesis. The review focused on peer-reviewed Q1/Q2 journal articles published between January 2021 and November 2025. The final corpus consisted of 154 studies that explicitly applied artificial intelligence (AI), machine learning, or related data-driven techniques to Internet of Things (IoT) or Industrial Internet of Things (IIoT) security problems and provided sufficient methodological and experimental detail for synthesis. In this review, deployability refers to the extent to which an AI-based IoT/IIoT security method provides evidence relevant to operational adoption, including realistic validation, false-positive behaviour, latency, memory or energy consumption, robustness to distribution shift, explainability, and implementation feasibility.

Review scope and protocol

The review scope was defined around three concepts: IoT/IIoT environments, security objectives, and AI-based methods. A study was considered eligible only if it addressed an IoT or IIoT security problem using AI, machine learning, or a related data-driven technique. Security problems included intrusion detection, anomaly detection, distributed denial-of-service detection, malware or botnet detection, device authentication, privacy preserving analytics, adversarial robustness, and adaptive security response.

A review protocol was finalized before database searching and before screening decisions were made. The protocol defined the databases, date range, Boolean search strings, inclusion and exclusion criteria, screening procedure, quality assessment rubric, data extraction schema, and synthesis strategy. Establishing the protocol before the search and screening stages was intended to reduce selection bias and ensure that all records were assessed consistently. The protocol was not externally registered. However, it was finalized before database searching and is provided as Supplementary Information, together with the search log, list of included studies, screening decisions, extraction sheet, and quality assessment scores.

Search strategy

Searches were conducted in six scholarly databases: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, ScienceDirect, and SpringerLink. These databases were selected because they provide broad coverage of computer science, cybersecurity, networking, IoT, and engineering journals. Searches were restricted to peer-reviewed journal articles written in English and published between January 2021 and November 2025. The final database search was completed on 17 November 2025. No additional database search was conducted after this date. This cut-off was selected to allow completion of deduplication, full text assessment, quality appraisal, data extraction, and synthesis before manuscript preparation. Consequently, studies published after this date are outside the review scope.

The search strategy prioritized sensitivity over narrow precision because AI-based IoT/IIoT security research is distributed across several research communities and uses heterogeneous terminology. The search combined three groups of terms: IoT/IIoT context terms, security objective terms, and AI technique terms. The general search logic was:

(*IoT/IIoT terms*) AND (*security terms*) AND (*AI terms*)

The IoT/IIoT terms included “IoT”, “Internet of Things”, “IIoT”, and “Industrial Internet of Things”. The security terms included “security”, “cybersecurity”, “intrusion detection”, “anomaly detection”, “malware”, “botnet”, “DDoS”, “authentication”, “privacy”, and “attack detection”. The AI terms included “artificial intelligence”, “machine learning”, “deep learning”, “federated learning”, “reinforcement learning”, “adversarial learning”, “self-supervised learning”, and “explainable AI”. Common abbreviations and spelling variants, including IDS, NIDS, IoMT, CPS, edge AI, and TinyML, were considered during database-specific adaptation and manual screening.

Database-specific syntax was adapted to the search interface and indexing rules of each source. The exact query executed in each database, including fields searched, filters, execution dates, and number of records retrieved, is provided in Supplementary

Table S2. The database search identified 1,250 records. To reduce the risk of missing relevant studies due to indexing limitations and terminology variation, backward citation chaining, forward citation chaining, and manual checks of relevant journals were also conducted, producing 200 additional records. These additional records consisted of 76 records from backward citation chaining, 84 records from forward citation chaining, and 40 records from manual journal checks. In total, 1,450 records were identified before deduplication.

Eligibility criteria

Eligibility criteria were defined before database searching and were applied consistently during title and abstract screening and full-text eligibility assessment. Studies were included if they were peer-reviewed Q1/Q2 journal articles published between January 2021 and November 2025, addressed an explicit IoT or IIoT security context, used AI or a related data-driven method, targeted a security objective, and provided empirical evaluation with sufficient methodological detail. Studies were excluded if they were conference papers, workshop papers, editorials, book chapters, theses, white papers, non-peer-reviewed preprints, non-English publications, non-AI studies, non-security IoT analytics papers, conceptual articles without empirical evaluation, or studies with insufficient methodological detail. The complete inclusion and exclusion criteria are provided in Supplementary Table S4.

The restriction to Q1/Q2 journal articles was adopted as a corpus-boundary decision to support a focused synthesis of mature, peer-reviewed, and sufficiently documented evidence. It was not used as a proxy for methodological quality. AI and cybersecurity research are often disseminated through conferences, workshops, and preprints; however, this review focuses on journal-level evidence because its objective is not to map all emerging ideas, but to assess the strength of evidence supporting deployability claims in AI based IoT/IIoT security. Deployability assessment requires sufficient reporting of datasets, evaluation protocols, metrics, baselines, limitations, and implementation constraints, which are more consistently available in full journal articles than in shorter or preliminary publication formats.

Journal quartile status was checked using SCImago Journal Rank and/or Journal Citation Reports corresponding to the publication year or the closest available ranking year. When a journal appeared in multiple subject categories, the quartile most relevant to computer science, cybersecurity, networking, information systems, or engineering was considered. When quartile information differed across ranking sources, the decision rule and retained classification were documented in the supplementary extraction sheet.

This criterion introduces a scope limitation. Excluding conference papers may underrepresent recent advances in fast-moving areas such as adversarial learning, federated learning, TinyML, edge AI, and security automation. Therefore, the findings should be interpreted as representing mature journal-level evidence rather than the complete universe of AI-based IoT/IIoT security research. To reduce the risk of overinterpreting journal quartile as quality, study-level quality was assessed independently using the predefined quality assessment rubric.

Study selection

The study selection process followed a PRISMA 2020 workflow consisting of identification, deduplication, title and abstract screening, full-text eligibility assessment, and final inclusion. The initial search identified 1,250 records from the six selected databases and 200 additional records through citation chaining and manual checks, producing 1,450 records before deduplication. After removing 300 duplicate records using DOI, title, author, and publication year matching, 1,150 records remained for title and abstract screening.

During title and abstract screening, 850 records were excluded because they did not meet the scope of AI-based IoT/IIoT security or did not satisfy the basic inclusion criteria. The remaining 300 full-text reports were assessed for eligibility. Of these, 146 reports were excluded after full text assessment because they did not meet one or more eligibility criteria. The final corpus included 154 studies, which were retained for qualitative synthesis, descriptive quantitative analysis, quality assessment, and deployment-oriented interpretation. The PRISMA summary and full-text exclusion reasons are provided in Supplementary Tables S5 and S6.

Each record was reviewed by at least two reviewers. Disagreements during title and abstract screening and full-text eligibility assessment were first resolved through discussion. When consensus could not be reached, a third reviewer was consulted. Screening decisions were recorded in the extraction and screening log to preserve traceability. Figure 2 reports the full PRISMA flow.

Quality assessment

A structured quality assessment was conducted for all 154 included studies. The objective was not to exclude studies after inclusion, but to assess the strength of the evidence supporting the synthesis. Each study was evaluated using an eight-criterion rubric covering problem definition, dataset transparency, evaluation protocol, baselines, metric reporting, reproducibility, limitations, and deployability evidence.

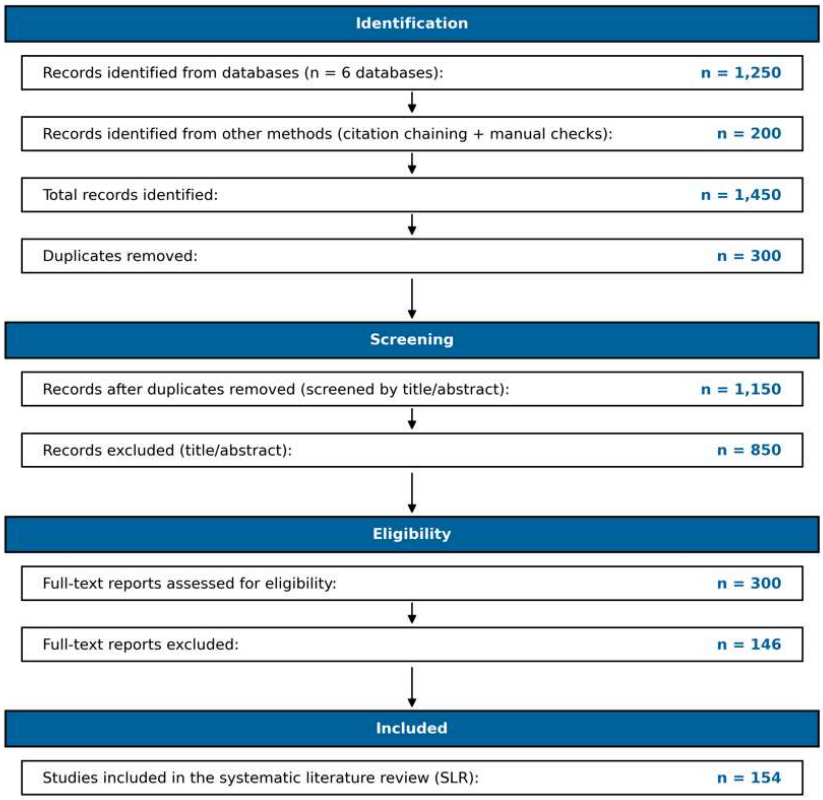


Figure 2. PRISMA flow diagram of the study selection process. The search identified 1,250 database records and 200 additional records through citation chaining and manual checks. After removing 300 duplicates, 1,150 records were screened by title and abstract. Of these, 850 records were excluded, leaving 300 full-text reports for eligibility assessment. A total of 146 full-text reports were excluded, and 154 studies were included in the final synthesis.

Each criterion was scored from 0 to 2. A score of 0 indicated absent or weak evidence, 1 indicated partial evidence, and 2 indicated clear and sufficient evidence. The maximum possible score was 16. Studies scoring between 12 and 16 were classified as high quality, studies scoring between 8 and 11 as medium quality, and studies scoring between 0 and 7 as low quality. The complete rubric and scoring guidance are provided in Supplementary Table S7.

Quality scoring was performed independently by two reviewers using the predefined rubric. Disagreements were discussed criterion by criterion. When disagreement remained after discussion, a third reviewer made the final decision. Inter-reviewer agreement was calculated before consensus resolution. For eligibility screening, Cohen's kappa was calculated on the include/exclude decisions made during full text assessment of the 300 reports, resulting in $\kappa = 0.82$. For quality assessment, agreement was calculated at the quality-band level across the 154 included studies, resulting in $\kappa = 0.79$. After consensus discussion, all quality scores were finalized and recorded in the supplementary quality assessment sheet.

Data extraction

Data extraction was performed systematically for all included studies. The extraction schema captured bibliographic information, methodological features, evaluation practices, and deployment-related evidence. Extracted variables included author information, year, journal, publisher, article type, security task, AI family, dataset, IoT-native status, attack types, evaluation protocol, reported metrics, baselines, quality assessment score, deployment-related constraints, limitations, and inferred methodological risks. The complete data extraction schema is provided in Supplementary Table S9.

Each study was assigned to one primary AI family according to its main methodological contribution and experimental focus. Secondary technique tags were retained in the supplementary extraction sheet. Similarly, each study was assigned to one primary security task based on its main objective, experimental setup, and principal contribution. Secondary task tags were retained to preserve multi-task information without double counting in the main descriptive distributions.

Synthesis approach

The synthesis combined descriptive quantitative analysis and critical narrative interpretation. Quantitative coding was used to summarize the distribution of studies by publication year, security task, AI family, dataset, evaluation protocol, reported metrics, quality score, and deployment-related constraints. This enabled cross-tabulations such as AI family by year, AI family by security task, operational metric reporting, evaluation protocol distribution, dataset usage, and quality assessment distribution.

The narrative synthesis interpreted these quantitative patterns through a deployment-oriented analytical lens. In particular, we assessed whether reported results were supported by realistic datasets, robust validation protocols, operational metrics, transparent quality evidence, and explicit consideration of labeling, transfer, resource, and explainability constraints. Claims about evidence maturity were therefore based not only on the number of studies in a given area, but also on the quality and operational relevance of the supporting evidence.

To reduce overreliance on prior surveys, synthesis claims were anchored primarily in the extracted corpus and supported by representative primary studies. Survey articles were used mainly for background, positioning, and comparison with existing reviews. Corpus-level claims were linked to the quantitative tables and figures produced by this review.

Reproducibility and evidence traceability

To improve reproducibility, all included studies were assigned a unique study identifier. The extraction sheet recorded bibliographic information, eligibility status, quality score, AI family, security task, dataset, evaluation protocol, reported metrics, and deployment-related constraints for each study. The review protocol, complete search log, list of included studies, extraction sheet, quality scores, screening decisions, manually checked journal list, and extended extraction details are provided as Supplementary Information.

The supplementary material allows readers and reviewers to trace each included study from identification to extraction and quality appraisal. This traceability is important because the review not only summarizes the literature, but also evaluates the strength of the evidence supporting deployment claims. Therefore, conclusions are derived from the coded corpus, quality assessment, and quantitative synthesis rather than from narrative interpretation alone.

Results

This section reports the findings extracted from the 154-study corpus. Results are organized around five levels of evidence: publication trends, distribution of artificial intelligence (AI) technique families, distribution of IoT/IIoT security tasks, AI-family-by-security-task mapping, and evidence conditions affecting deployability. Percentages are calculated using the 154 included studies as the denominator unless otherwise stated. Each study was assigned to one primary AI family and one primary security task according to its main methodological contribution, experimental focus, and stated security objective. Secondary technique and task tags were retained in the supplementary extraction sheet to preserve multi-label information without double counting in the main descriptive distributions.

Publication trends

The corpus shows a marked increase in AI based IoT/IIoT security research across the review period. The number of included studies rose from 3 in 2021 to 21 in 2022, 45 in 2023, and 74 in 2024, before reaching 11 in 2025. The smaller 2025 count reflects the November 2025 search cutoff rather than evidence of a clear decline. This trajectory indicates a rapid expansion of journal-level evidence, particularly after 2022, and suggests that the field has moved from early proof-of-concept studies toward broader methodological diversification.

The 2023–2024 increase is not only quantitative. Studies published during this period show greater diversity in modelling strategies and deployment assumptions, including stronger deep learning architectures for intrusion and anomaly detection, growing use of federated or distributed learning for privacy-preserving security analytics, and more explicit attention to adversarial robustness, edge feasibility, and explainability. However, the extent to which these trends translate into operational readiness depends on the quality of validation protocols, dataset realism, operational metric reporting, and evidence of robustness under distribution shift.

Distribution by AI technique family

The corpus is concentrated around three main AI families: classical machine learning, federated, split, or swarm learning, and deep learning. Classical machine learning represents the largest category, with 61 studies (39.6%). Federated, split, or swarm learning accounts for 24 studies (15.6%), while deep learning accounts for 22 studies (14.3%). Together, these three families represent 107 studies, or 69.5% of the corpus.

This concentration indicates that AI based IoT/IIoT security research remains primarily detection-oriented. Classical machine learning is widely used because it supports structured feature-based modelling, comparative benchmarking, and relatively lightweight deployment. Deep learning is mainly used where studies require representation learning, temporal

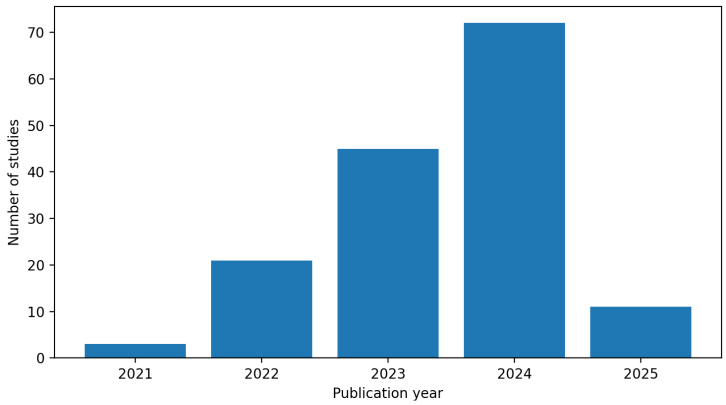


Figure 3. Yearly distribution of the 154 included studies between January 2021 and November 2025.

modelling, or complex traffic-pattern extraction. Federated, split, and swarm learning approaches are increasingly used where privacy, decentralization, or distributed telemetry are central to the security problem.

Other AI families remain less represented. Optimization, metaheuristic, fuzzy, rule-based, or immune-inspired approaches account for 12 studies (7.8%). Graph or knowledge-graph models account for 7 studies (4.5%), reflecting emerging interest in relational and structured security patterns. Explainable AI and privacy, blockchain, or differential-privacy-oriented approaches each account for 6 studies (3.9%). Edge, TinyML, programmable-switch, or knowledge-distillation approaches account for 5 studies (3.2%). Generative models and quantum machine learning each account for 4 studies (2.6%), while reinforcement learning accounts for 3 studies (1.9%). The full distribution is provided in Supplementary Table S11.

Distribution by IoT/IIoT security task

The security-task distribution is strongly concentrated around intrusion detection and anomaly detection. Intrusion detection, including IDS and network intrusion detection systems, is the largest category, with 48 studies (31.2%). Anomaly detection is the second largest category, with 28 studies (18.2%). Together, these two categories account for 76 studies, or 49.4% of the corpus.

This finding shows that the literature is dominated by detection tasks supported by reusable datasets, standardized metrics, and established benchmark pipelines. By contrast, operationally important tasks remain comparatively sparse. Authentication and device identification account for only 2 studies (1.3%), insider threat detection for 2 studies (1.3%), root-cause localization for 3 studies (1.9%), and threat hunting or response for 4 studies (2.6%). Malware and botnet detection accounts for 6 studies (3.9%), while DoS or DDoS detection and adversarial robustness or backdoor defense each account for 7 studies (4.5%). Encrypted traffic classification or threat detection accounts for 11 studies (7.1%), and CPS, ICS, or SCADA protection accounts for 10 studies (6.5%). A broader category of general security, multi-task, or unspecified IoT/IIoT security analytics accounts for 26 studies (16.9%). The full distribution is provided in Supplementary Table S12.

The imbalance between detection-heavy research and less represented operational functions is important. Real-world IoT/IIoT security programs require not only detection, but also identification, explanation, localization, triage, response support, and governance evidence. The limited number of studies addressing authentication, insider threats, threat response, and root-cause localization indicates that several high-value operational security functions remain underdeveloped in the AI based IoT/IIoT security literature.

AI family by security task

The AI-family-by-security-task heatmap shows that the densest areas of evidence occur at the intersection of classical machine learning and detection-oriented tasks. The largest cell is classical machine learning for intrusion detection, with 17 studies (11.0%). Classical machine learning for anomaly detection and deep learning for intrusion detection each account for 12 studies (7.8%). Classical machine learning also appears across encrypted traffic classification, CPS/ICS/SCADA protection, and general security or multi-task security analytics, each with 7 studies (4.5%). Optimization, metaheuristic, fuzzy, rule-based, or immune-inspired approaches also appear prominently in intrusion detection, with 7 studies (4.5%). Federated, split, or swarm learning is most visible in anomaly detection, with 6 studies (3.9%). Selected high-density combinations are reported in Supplementary Table S14.

The heatmap should be interpreted as an evidence-density map rather than a direct measure of operational maturity. A high-density cell indicates that a topic has received substantial research attention, but it does not necessarily mean that the

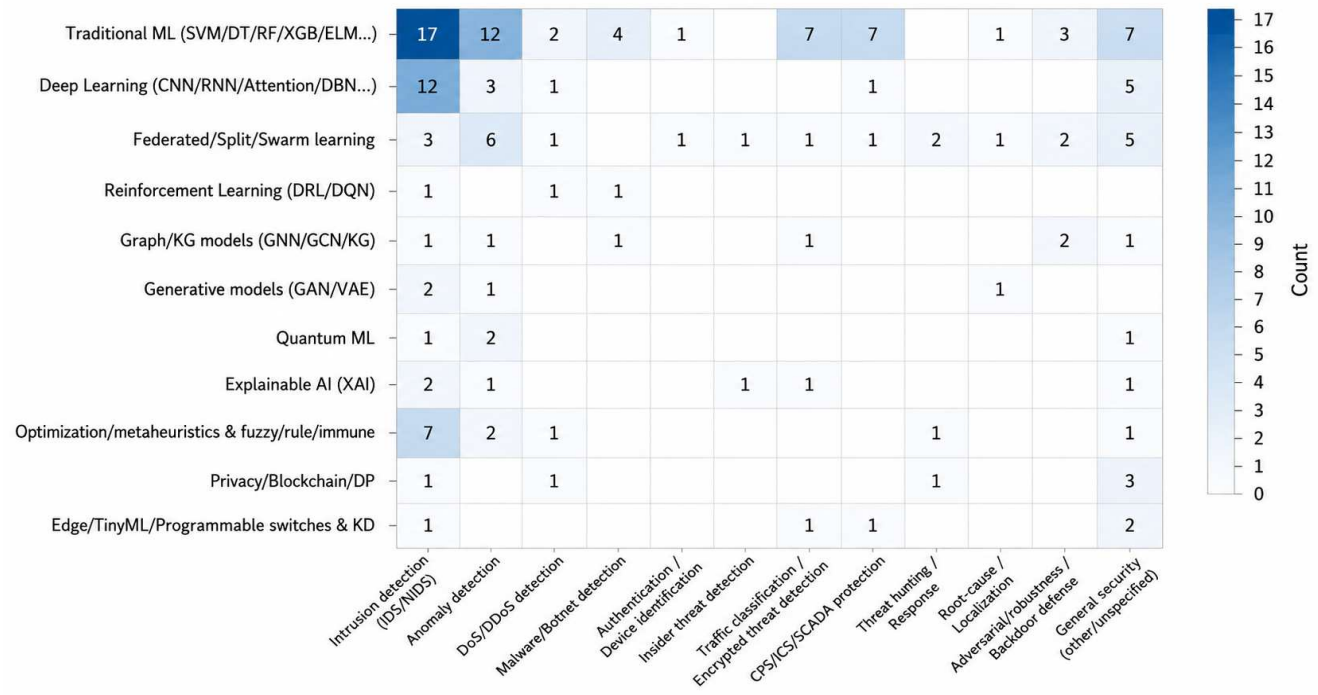


Figure 4. Evidence-density heatmap mapping primary AI technique families to primary IoT/IIoT security tasks across the 154 included studies. Darker cells indicate higher study concentration, but not necessarily stronger deployment maturity.

corresponding methods are deployment-ready. Deployment readiness depends on validation design, dataset realism, cross-context transfer, operational metric reporting, resource feasibility, robustness, and explainability. Conversely, lower-density cells may reflect underexplored but operationally important research areas rather than low practical relevance.

Dataset comparability and benchmarking evidence

The review identified dataset choice as a major determinant of how performance claims should be interpreted. Four widely used benchmarks illustrate this issue: IoT-23, CIC-IDS2017, CSE-CIC-IDS2018, and NSL-KDD. These datasets should not be treated as interchangeable because they differ in domain, data type, realism, scenario design, label structure, and relevance to IoT/IIoT deployment. A detailed comparison is provided in Supplementary Tables S15 and S16.

IoT-23 provides the strongest direct support among these four datasets for IoT malware and behavioural analysis because it includes malware-infected IoT traffic captures and benign traffic from real IoT devices. However, its scenario structure creates a risk of inflated results when records from the same scenario are randomly split across training and testing. Scenario-disjoint evaluation provides stronger evidence for cross-context robustness.

CIC-IDS2017 supports controlled multi-attack intrusion detection analysis through a structured enterprise-style testbed. Its daily attack structure can support temporal robustness checks if day boundaries are respected. However, performance estimates can be inflated when flows from the same scripted attack scenario appear in both training and test sets.

CSE-CIC-IDS2018 extends the enterprise-style evaluation setting by including larger-scale traffic and richer telemetry. This makes it useful for studies that motivate system-aware or multi-telemetry intrusion detection. However, comparability across studies depends on whether authors report the exact subset, artifacts, days, scenarios, feature extraction pipeline, and preprocessing choices used.

NSL-KDD remains useful mainly as a historical baseline because of its predefined train/test structure and high reproducibility. However, it is not IoT-native and does not reflect contemporary IoT protocol heterogeneity, device behaviour, encrypted traffic patterns, or modern deployment constraints. It should therefore not be used as primary evidence for deployable IoT/IIoT security claims.

Overall, benchmark performance is better interpreted as evidence of separability under a specific dataset instrument than as direct evidence of deployability. Dataset realism, split design, attack-scenario independence, device independence, and cross-dataset validation are central to the credibility of AI-based IoT/IIoT security claims.

Evaluation and operational evidence

The reviewed literature shows that evaluation design strongly affects the meaning of reported performance. Random flow-level splits can overestimate generalization when records from the same scenario, device, capture period, or scripted attack sequence appear in both training and testing. Stronger evidence is provided by scenario-disjoint, day-disjoint, device-disjoint, time-aware, and cross-dataset evaluation designs because these protocols better approximate the temporal, contextual, and domain shifts encountered in operational IoT/IIoT environments.

Operational metric reporting is also uneven. Accuracy, precision, recall, and F1-score are widely used, but they are insufficient on their own for evaluating deployment readiness. IoT/IIoT security systems must operate under constraints involving latency, memory, energy, throughput, communication overhead, drift behaviour, false-positive burden, and explanation usefulness. Studies that report only classification metrics provide weaker deployment evidence than studies that also evaluate operational cost, robustness, or integration feasibility.

This review therefore treats evaluation design and operational metric reporting as evidence conditions. The strongest deployment-oriented evidence comes from studies that combine realistic or IoT-native datasets, leakage-resistant validation protocols, strong baselines, operational metrics, and transparent reporting of limitations. Weaker evidence comes from studies that rely on random splits, legacy datasets, narrow metrics, incomplete preprocessing descriptions, or limited discussion of deployment constraints.

Quality assessment distribution

The quality assessment shows that most studies fall into the medium or high quality bands. Of the 154 included studies, 62 studies were classified as high quality, 73 as medium quality, and 19 as low quality. High-quality studies provided clearer problem definitions, stronger dataset transparency, more credible evaluation protocols, more appropriate baselines, richer metric reporting, better reproducibility, explicit limitations, and stronger deployability evidence. Medium-quality studies provided useful evidence but often had limitations related to validation design, baseline selection, metric completeness, or reporting of operational constraints. Low-quality studies were retained for completeness but were assigned lower interpretive weight in the synthesis. The full quality distribution is provided in Supplementary Table S13.

The quality distribution indicates that the field contains a substantial body of useful evidence, but the maturity of deployment claims remains uneven. Many studies provide strong benchmark results, yet fewer provide evidence that directly supports operational adoption under realistic conditions such as distribution shift, concept drift, constrained resources, adversarial manipulation, or analyst-facing explainability.

Summary of results

Five main findings emerge from the corpus. First, AI based IoT/IIoT security research expanded rapidly after 2022, with the strongest publication concentration in 2024. Second, the evidence base is dominated by classical machine learning, deep learning, and federated, split, or swarm learning, which together account for 69.5% of the included studies. Third, the literature is concentrated around intrusion detection and anomaly detection, which together account for 49.4% of the corpus, while authentication, insider threat detection, root-cause localization, and threat response remain comparatively underexplored. Fourth, dataset choice and validation design strongly influence the credibility of reported performance, especially when random splits or non-IoT-native benchmarks are used. Fifth, quality assessment shows that a substantial proportion of the literature provides useful evidence, but deployment readiness remains constrained by uneven reporting of operational metrics, robustness, transferability, resource feasibility, and explainability.

Taken together, the results show that AI has been extensively explored for IoT/IIoT security, but the strength of evidence supporting real-world deployment remains inconsistent. The literature demonstrates clear progress in detection-oriented modelling, yet deployment-grade assurance requires stronger validation under domain shift, more systematic reporting of runtime and resource costs, better treatment of drift and adversarial robustness, and explainability evaluated in relation to operator trust and security decision-making.

Discussion

This systematic review shows that AI based IoT and IIoT security research has expanded rapidly, but that maturity cannot be inferred from benchmark performance alone. The reviewed evidence indicates a clear shift from isolated proof of concept studies toward a broader research landscape that includes classical machine learning, deep learning, federated and split learning, graph models, explainable AI, optimization methods, edge designs, and emerging quantum or generative approaches. This growth is scientifically important, but it also increases heterogeneity in datasets, threat models, validation protocols, metrics, and deployment assumptions. For this reason, publication volume and headline accuracy should not be treated as evidence of operational readiness. A method becomes relevant for deployment only when it is evaluated under conditions that resemble real

IoT and IIoT environments, including scarce or noisy labels, changing device behavior, transfer across contexts, non stationary traffic, constrained resources, adversarial pressure, and governance requirements.

The concentration of the corpus around intrusion detection and anomaly detection is both expected and revealing. These tasks benefit from reusable datasets, established labels, and standard metrics such as accuracy, precision, recall, F1 score, and area under the curve. This has enabled rapid comparison between models and has accelerated methodological progress. However, the same concentration also exposes an imbalance in the evidence base. Operational security programs require more than detection. They require device identification, alert triage, root cause localization, response support, accountability, and integration with human decision making. The limited evidence for authentication, insider threat detection, threat hunting, response support, and root cause localization suggests that several operationally important security functions remain underdeveloped. The field has therefore advanced fastest where benchmark infrastructure is available, but its next phase of maturity requires moving beyond detection centered experiments toward end to end security workflows.

The results also show that the meaning of performance depends strongly on dataset design and validation protocol. Public benchmarks are useful for reproducibility and comparison, but they are not neutral instruments. IoT native datasets provide closer semantic proximity to IoT device behavior and malware activity, while enterprise or legacy intrusion detection datasets support controlled comparison but may introduce domain shift when conclusions are generalized to IoT and IIoT fleets. Performance should therefore be interpreted as evidence of separability under a specific dataset instrument, not as direct proof of deployment readiness. Random flow level splits are especially problematic when records from the same scenario, capture period, scripted attack, or device behavior appear in both training and testing. Stronger evidence comes from scenario disjoint, device disjoint, time aware, and cross dataset validation because these designs better approximate the temporal and contextual shifts encountered in operational environments.

This distinction matters because operational security value is not determined by headline accuracy alone. Classification metrics remain useful for controlled comparison, but they do not fully capture whether an AI based security method can reduce risk in production. In IoT and IIoT environments, deployment readiness also depends on false positive burden, calibration stability, end to end latency, throughput, memory use, energy consumption, communication overhead, drift behavior, robustness to adversarial manipulation, and explanation usefulness. Some studies report model inference time, but real operational latency also includes feature extraction, packet or flow aggregation, buffering, transport, orchestration, alert generation, and response action. A model that appears efficient in isolation may therefore be difficult to operate when embedded in a real security pipeline.

Label scarcity and label quality remain central bottlenecks. Operational ground truth is expensive, delayed, and sometimes ambiguous. Attacks may be rare, incident confirmation often requires investigation, and the meaning of malicious behavior can depend on device role, environment, policy, and business context. This creates a practical gap between curated benchmark labels and operational labeling pipelines. Three risks follow from this gap. First, models may decay under drift when retraining cycles cannot keep pace with changing devices, firmware, workloads, and adversary behavior. Second, benchmark conditions can create overconfidence by hiding missing data, noisy telemetry, inconsistent logging, or partial visibility. Third, public dataset labels may not match operational definitions of maliciousness, causing models to optimize objectives that are only partially aligned with real security decisions.

Transfer across contexts is another major barrier to deployment. IoT and IIoT environments differ across device types, firmware versions, communication protocols, topology, traffic patterns, site policies, and monitoring capabilities. Non independent and non identical behavior is therefore a normal operating condition rather than an exception. A model that performs well in one dataset, site, or time period may degrade when transferred to another unless transfer behavior and adaptation cost are explicitly measured. Future studies should therefore report the training setting, deployment setting, adaptation procedure, required labels or compute, stabilization time, and performance loss under mismatch. This would convert transferability from a narrative claim into a measurable deployment signal.

Resource and real time feasibility should also be evaluated as system properties rather than model properties alone. IoT and IIoT security pipelines involve sensing, collection, preprocessing, feature extraction, inference, communication, orchestration, alerting, and response. Deployment claims are stronger when they report end to end latency, memory footprint, energy use, throughput, overload behavior, and operational trade offs. Lightweight models, pruning, quantization, knowledge distillation, hierarchical inference, gateway assisted analytics, and in network filtering are promising directions, but their value depends on measured system level performance rather than theoretical suitability for constrained devices.

Privacy preserving and distributed learning approaches introduce a different maturity profile. Federated and split learning can reduce the need to centralize raw telemetry, which is important for privacy, data ownership, and regulatory constraints. However, they also move part of the attack surface to model updates, aggregation logic, client behavior, and communication channels. Poisoning, backdoor attacks, gradient leakage, unreliable clients, non independent data, and communication overhead all affect the security and practicality of collaborative learning. These methods should therefore be evaluated not only as learning techniques, but also as distributed security protocols requiring update governance, robust aggregation, client trust

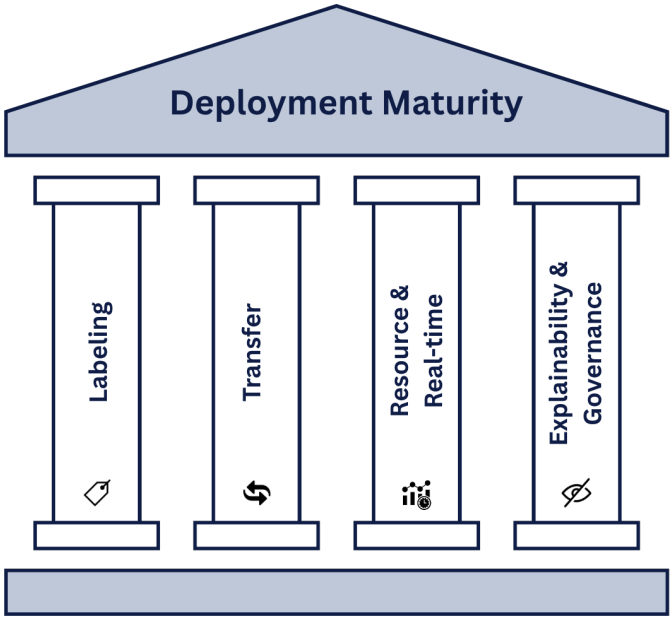


Figure 5. Deployment maturity pillars for AI based IoT and IIoT security. The figure summarizes the four operational constraints used to interpret the reviewed evidence: labeling burden, transfer under domain shift, resource and real time feasibility, and explainability or governance readiness.

management, and monitoring.

Explainability is another deployment condition, not a cosmetic feature. In security operations, explanations are useful only if they help analysts triage alerts, identify root causes, justify escalation, support audit narratives, and reduce unnecessary false escalation. Post hoc feature attributions or visualizations are not sufficient unless their stability, usefulness, and robustness are evaluated in relation to real decisions. Explainable AI for IoT and IIoT security should therefore move from explanation availability toward explanation validation. The strongest evidence would show whether explanations improve analyst decisions, remain stable under drift, support root cause analysis, and align with governance or compliance requirements.

Taken together, the evidence suggests that the most important future work is not simply to design more complex models. The field needs stronger evidence that AI security methods can operate reliably under deployment constraints. Five priorities emerge. First, evaluation should explicitly test transfer from one training setting to another deployment setting, including adaptation cost, required labels, and performance loss under mismatch. Second, long term validation should become standard, including time aware splits, performance decay, recovery behavior, and retraining cost. Third, edge and real time security intelligence should be evaluated as an end to end system property, not only as model compression or inference speed. Fourth, robustness should become a core evaluation axis, including evasion, poisoning, backdoor scenarios, adversarial drift, and robustness versus utility trade offs. Fifth, explainability and governance should be validated as operational controls that support trust, accountability, auditability, and security decisions.

The reviewed evidence suggests that deployment maturity depends on four recurring operational pillars: labeling burden, transfer under domain shift, resource and real time feasibility, and explainability or governance readiness. Figure 5 summarizes these pillars as the conceptual bridge between benchmark evidence and deployment ready AI for IoT and IIoT security.

Using these four pillars, the reviewed AI families can be interpreted according to their deployment strengths, operational risks, and evidence gaps. Figure 6 provides a qualitative synthesis of this relationship. It should not be read as a ranking of AI families, but as a structured interpretation of how each family relates to labeling requirements, transfer behavior, resource feasibility, robustness, explainability, privacy, and governance.

Across AI families, deployment maturity is better understood as evidence completeness than as peak benchmark performance. Classical machine learning remains relevant when resource predictability, simplicity, and interpretability matter, but its maturity depends on validation under changing environments. Deep learning can provide strong representation learning and high benchmark performance, but its deployment readiness depends on label sustainability, calibration, compression, drift behavior, and validation across domains. Unsupervised, self supervised, and online learning align well with label scarce IoT settings, but require stronger evidence on stability, adaptation cost, and safe updating. Federated and split learning address privacy and data ownership constraints, but require realistic non independent data evaluation, communication cost reporting, update security analysis, and defenses against poisoning or leakage. Reinforcement learning is most defensible as a constrained decision

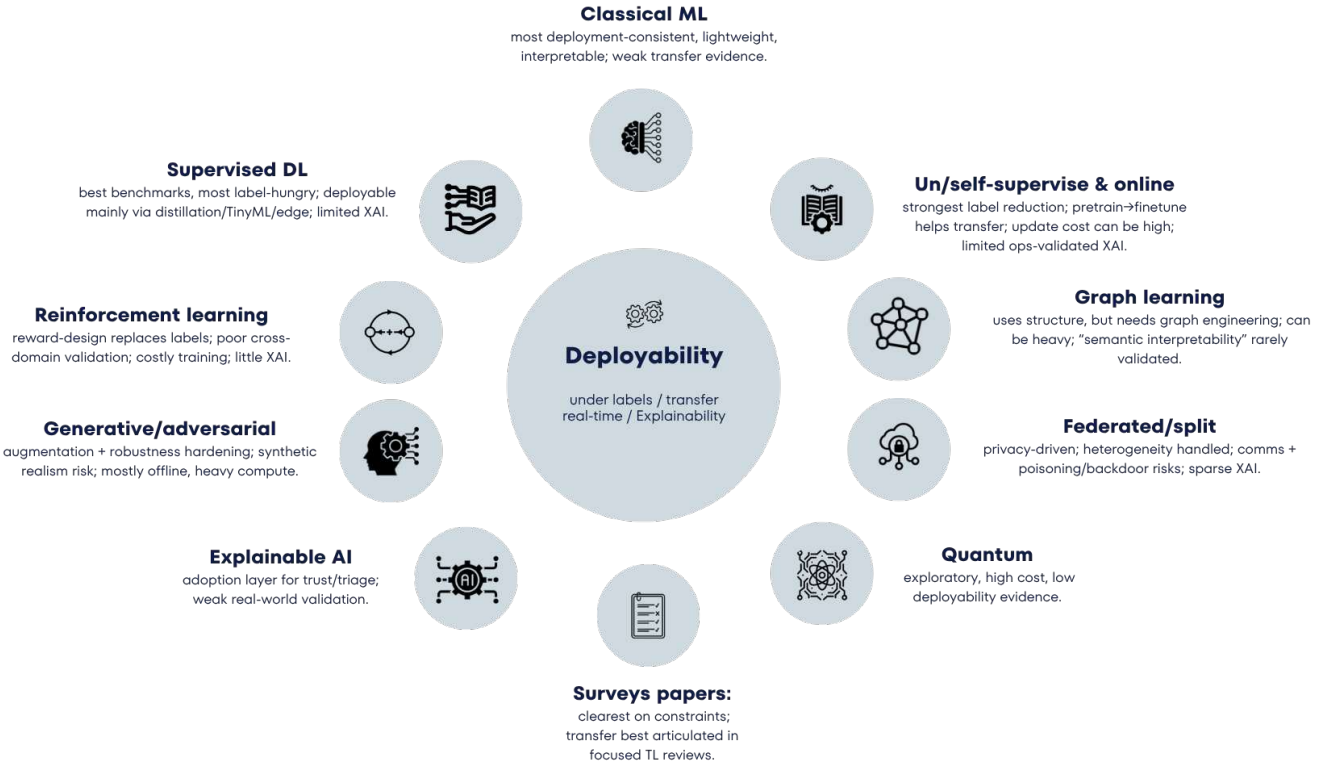


Figure 6. AI family maturity under deployment constraints. The figure summarizes how major AI families relate to labeling burden, transfer, resource and real time feasibility, robustness, explainability, privacy, and governance. It should be interpreted as a qualitative synthesis of evidence conditions rather than a quantitative ranking of model families.

support layer rather than a fully autonomous response system in safety sensitive environments. Generative and adversarial approaches are valuable for augmentation and robustness testing, but their credibility depends on synthetic data fidelity and operational cost. Graph learning is promising for relational threats, lateral movement, and coordinated behavior, but requires evidence on streaming graph construction, maintenance cost, and topology shift robustness. Quantum and quantum inspired approaches remain exploratory and should be framed as forward looking rather than deployment ready.

This review has several limitations. First, the evidence is influenced by dataset availability and reporting practices. Public benchmarks support reproducibility and comparison, but they may not reflect operational IoT and IIoT environments. Second, many evaluations approximate open world security through curated datasets, fixed label sets, limited perturbations, or partial novelty settings. These designs are useful for controlled experimentation but do not fully capture evolving adversaries, unknown attacks, long horizon drift, or deployment specific failure modes. Third, baseline strength, preprocessing, feature access, split design, and reporting detail vary across studies, meaning that some reported gains may reflect dataset artifacts or evaluation design rather than general improvements in security capability. Fourth, publication bias may favor positive benchmark improvements, while negative results, failed deployments, degradation studies, and long term maintenance evidence remain less visible. Fifth, the Q1/Q2 journal focus provides a mature evidence boundary but may underrepresent recent conference driven advances, proprietary industrial evidence, and sensitive operational deployments.

Finally, evaluation protocol and operational metric variables were not fully frequency coded across all included studies. For this reason, deployment related conclusions should be interpreted as evidence conditions rather than corpus level prevalence estimates. This limitation does not weaken the central finding; rather, it clarifies it. The field has generated substantial evidence that AI can support IoT and IIoT security, especially for detection tasks, but deployment ready assurance remains uneven. Future progress depends on moving from benchmark centered performance reporting to evidence frameworks that test label sustainability, transfer, drift, end to end cost, robustness, explainability, and governance. In practical terms, deployable AI for IoT and IIoT security should be evaluated not as a standalone model, but as a maintainable security capability embedded in operational environments.

Conclusion

This systematic literature review synthesized 154 Q1/Q2 journal articles published between January 2021 and November 2025 to assess what artificial intelligence can reliably contribute to IoT/IIoT security and what still limits deployment-grade assurance. The evidence shows rapid growth in AI based IoT/IIoT security research, but also a strong concentration around intrusion detection and anomaly detection. These areas benefit from reusable datasets and standardized metrics, whereas identity assurance, insider threat detection, root-cause localization, threat hunting, and response support remain comparatively underexplored.

The main finding is that deployment maturity cannot be inferred from benchmark accuracy alone. Across AI families, operational relevance depends on whether methods remain useful under scarce or noisy labels, heterogeneous device fleets, cross-context transfer, concept drift, constrained resources, adversarial pressure, and accountability requirements. High benchmark scores should therefore be interpreted as evidence of performance under specific experimental conditions, not as direct proof of operational reliability.

Future research should move from benchmark centered evaluation toward deployment-oriented evidence. This requires standardized "train-here-deploy-there" protocols, explicit reporting of adaptation costs, drift-aware and long-horizon validation, end-to-end latency and resource measurement, robustness testing, and explainability validated with operational users. New public benchmarks and testbeds are also needed for underrepresented security functions. These shifts would help transform AI for IoT/IIoT security from promising models into maintainable, accountable, and deployment-ready security capabilities.

Data availability

NOT APPLICABLE

Acknowledgements

The authors thank their respective institutions for supporting the research environment in which this work was conducted.

Author contributions statement

A.R. and Z.C. conceived the study and designed the review protocol. A.R., Z.C., M.Z., S.B. and Y.M. contributed to the methodology, literature screening, data extraction, quality assessment, analysis, interpretation, and manuscript preparation. All authors reviewed, edited, and approved the final manuscript.

Additional information

Competing interests: The authors declare no competing interests.

Ethics approval and consent to participate: Not applicable. This study is a systematic literature review and did not involve human participants, human tissue, animals, or clinical data.

Consent for publication: Not applicable.

Use of artificial intelligence tools: During manuscript preparation, language-editing and formatting assistance was supported by a large language model. The authors critically reviewed, edited, and approved all content and remain fully responsible for the accuracy, integrity, originality, and scientific validity of the manuscript.

References

1. Park, J. H. *et al.* A comprehensive survey on blockchain for secure IoT-enabled smart city beyond 5G: Approaches, processes, challenges, and opportunities. *Hum.-Centric Comput. Inf. Sci.* **13**, 51, https://doi.org/10.22967/HCIS.2023.13.051 (2023).
2. Park, J. H. *et al.* PoAh enabled federated learning architecture for DDoS attack detection in IoT networks. *Hum.-Centric Comput. Inf. Sci.* **14**, 3, https://doi.org/10.22967/HCIS.2024.14.003 (2024).
3. Li, Q. *et al.* A novel intrusion detection method for imbalanced datasets. *Hum.-Centric Comput. Inf. Sci.* **15**, 19, https://doi.org/10.22967/HCIS.2025.15.019 (2025).
4. Hdaib, M. *et al.* Quantum deep learning-based anomaly detection for enhanced network security. *Quantum Mach. Intell.* **6**, 26, https://doi.org/10.1007/s42484-024-00163-2 (2024).
5. Tychola, K. A. *et al.* Quantum machine learning: An overview. *Electronics* **12**, 2379, https://doi.org/10.3390/electronics12112379 (2023).

6. Oh, H. & Park, D. K. Quantum support vector data description for anomaly detection. *Mach. Learn. Sci. Technol.* **5**, 035052, https://doi.org/10.1088/2632-2153/ad6be8 (2024).
7. Jamalipour, A. & Murali, S. A taxonomy of machine learning-based intrusion detection systems for the Internet of Things: A survey. *IEEE Internet Things J.* **9**, 9444–9466, https://doi.org/10.1109/JIOT.2021.3126811 (2022).
8. Ferrag, M. A., Shu, L., Friha, O. & Yang, X. Cyber security intrusion detection for agriculture 4.0: Machine learning-based solutions, datasets, and future directions. *IEEE/CAA J. Autom. Sin.* **9**, 407–436, https://doi.org/10.1109/JAS.2021.1004344 (2022).
9. Yang, M. & Zhang, J. Data anomaly detection in the Internet of Things: A review of current trends and research challenges. *Int. J. Adv. Comput. Sci. Appl.* **14**, https://doi.org/10.14569/IJACSA.2023.0140901 (2023).
10. Bansal, K. & Singhrova, A. Review on intrusion detection system for IoT/IIoT brief study. *Multimed. Tools Appl.* **83**, 23083–23108, https://doi.org/10.1007/s11042-023-16395-6 (2024).
11. Belay, M. A. *et al.* Unsupervised anomaly detection for IoT-based multivariate time series: Existing solutions, performance analysis and future directions. *Sensors* **23**, 2844, https://doi.org/10.3390/s23052844 (2023).
12. Moustafa, N. *et al.* Explainable intrusion detection for cyber defences in the Internet of Things: Opportunities and solutions. *IEEE Commun. Surv. Tutor.* **25**, 1775–1807, https://doi.org/10.1109/COMST.2023.3280465 (2023).
13. Adhikari, D. *et al.* Recent advances in anomaly detection in Internet of Things: Status, challenges, and perspectives. *Comput. Sci. Rev.* **54**, 100665, https://doi.org/10.1016/j.cosrev.2024.100665 (2024).
14. Trilles, S. *et al.* Anomaly detection based on artificial intelligence of things: A systematic literature mapping. *Internet Things* **25**, 101063, https://doi.org/10.1016/j.iot.2024.101063 (2024).
15. Cruz, M. A. A. D. *et al.* Detecting compromised IoT devices through XGBoost. *IEEE Trans. Intell. Transp. Syst.* **24**, 15392–15399, https://doi.org/10.1109/TITS.2022.3187252 (2023).
16. Yu, J. *et al.* CONGO^2: Scalable online anomaly detection and localization in power electronics networks. *IEEE Internet Things J.* **9**, 13862–13875, https://doi.org/10.1109/JIOT.2022.3143123 (2022).
17. Huertas Celdrán, A. *et al.* Privacy-preserving and syscall-based intrusion detection system for IoT spectrum sensors affected by data falsification attacks. *IEEE Internet Things J.* **10**, 8408–8415, https://doi.org/10.1109/JIOT.2023.3261989 (2023).
18. Rosero-Montalvo, P. D. *et al.* Hybrid anomaly detection model on trusted IoT devices. *IEEE Internet Things J.* **10**, 10959–10969, https://doi.org/10.1109/JIOT.2023.3243037 (2023).
19. Li, R. *et al.* ADRIoT: An edge-assisted anomaly detection framework against IoT-based network attacks. *IEEE Internet Things J.* **9**, 10576–10587, https://doi.org/10.1109/JIOT.2021.3122148 (2022).
20. Shukla, P. *et al.* EIoT-DDoS: Embedded classification approach for IoT traffic-based DDoS attacks. *Clust. Comput.* **27**, 1471–1490, https://doi.org/10.1007/s10586-023-04027-5 (2024).
21. Mothukuri, V. *et al.* Federated learning-based anomaly detection for IoT security attacks. *IEEE Internet Things J.* **9**, 2545–2554, https://doi.org/10.1109/JIOT.2021.3077803 (2022).
22. Sánchez, P. M. *et al.* Studying the robustness of anti-adversarial federated learning models detecting cyberattacks in IoT spectrum sensors. *IEEE Trans. Dependable Secure Comput.* **21**, 573–584, https://doi.org/10.1109/TDSC.2022.3204535 (2024).
23. Nguyen, H. & Kashef, R. TS-IDS: Traffic-aware self-supervised learning for IoT network intrusion detection. *Knowl.-Based Syst.* **279**, 110966, https://doi.org/10.1016/j.knosys.2023.110966 (2023).
24. Fu, C. *et al.* Flow interaction graph analysis: Unknown encrypted malicious traffic detection. *IEEE/ACM Trans. Netw.* **32**, 2972–2987, https://doi.org/10.1109/TNET.2024.3370851 (2024).
25. Sáez de Cámara, X. *et al.* Clustered federated learning architecture for network anomaly detection in large-scale heterogeneous IoT networks. *Comput. Secur.* **131**, 103299, https://doi.org/10.1016/j.cose.2023.103299 (2023).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- RharifAnassSupplementaryInformation.pdf