SSR Journal of Artificial Intelligence (SSRJAI)

OPEN ACCESS

Homepage: https://ssrpublisher.com/ssrjai/

Volume 3, Issue 8, 2026 ISSN: 3049-0413

Federated Learning for Privacy-Preserving Threat Detection in Massive IOT Ecosystems: A Systematic Literature Review

Peter Wonah Odey, Habibu Danjuma, Muhammad Salma Abidi*

*ORCID: 0009-0007-9250-3084

Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria

Received: *05.07.2026* | ***Accepted:*** *25.07.2026* | ***Published:*** *12.08.2026*

***Corresponding author:** Muhammad Salma Abidi

DOI: *10.5281/zenodo.21901804*

Abstract **Review Article**

Federated learning (FL) has emerged as a principal architecture for privacy-preserving intrusion detection in Internet of Things (IOT) environments, motivated by the impossibility of transmitting raw device traffic to a centralised server at scale. A rapidly growing body of empirical work applies FL to IOT intrusion detection systems (IDS), yet no PRISMA-compliant synthesis of this literature exists. This systematic review addresses that gap. Sixty-one peer-reviewed journal papers, identified through a PRISMA 2020-compliant search of five electronic databases covering 2018–2026, were subjected to full-text extraction and quality assessment on six dimensions. Four research questions guided the synthesis, organised into four themes. On FL architecture and performance (RQ1), standard FEDAVG and its variants govern aggregation in 65% of papers; two papers exceeded their centralised detection baseline, attributing the gain to data-centric rather than aggregation-level mechanisms. On privacy rigour (RQ2), 80% of papers claim privacy on structural grounds only; a twelve-subcategory privacy taxonomy is established, distinguishing formal differential privacy, cryptographic secure aggregation, and homomorphic encryption from structural-only claims; adaptive noise scheduling reduces the differential privacy accuracy cost from 5.77 percentage points to 0.01 percentage points relative to a non-private baseline. On evaluation realism (RQ3), 85% of papers evaluate on simulation only and 80% use independent and identically distributed data partitioning or do not state it; a 36.5-percentage-point accuracy gap between balanced and imbalanced evaluation conditions quantifies the inflation introduced by default evaluation methodology. On threat and domain coverage (RQ4), 57% of papers address generic multi-class intrusion in unspecified IOT deployment types; healthcare, smart grid, and industrial control systems account for five papers combined. Six research gaps are identified, and four prioritised research directions are proposed.

Keywords: federated learning, intrusion detection system, Internet of Things security, differen-tial privacy, systematic literature review, privacy-preserving machine learning.

1. INTRODUCTION

The number of Internet of Things (IOT) devices connected to global networks exceeded 15 billion in 2023 and is projected to surpass 30 billion by 2030 (Nguyen et al., 2021). The scale is unprecedented. Each device is a potential entry point for adversarial traffic, and the aggregate attack surface of a large IOT deployment is orders of magnitude larger than that of a conventional managed network (Diro and Chilamkurti, 2018). Centralised intrusion

detection, in which raw traffic logs are transmitted to a single analysis server, cannot scale to this environment. The communication overhead is prohibitive at thousands of heterogeneous edge nodes, and the transmission of raw data violates the data-minimisation requirements of general data-protection regulations in force across major deployment markets (Nguyen et al., 2021).

Federated learning (FL) offers a structural response to both constraints (McMahan et al., 2017). Each client trains a local model on device-resident data and transmits only model parameters to a central aggregator; raw traffic never leaves the device. The architecture naturally preserves a structural form of privacy and eliminates the single-point communication bottleneck of centralised training. These properties have motivated a substantial and rapidly growing body of research applying FL to IOT intrusion detection systems (IDS) (Chen et al., 2020; Huong et al., 2022; Sáez-de Cámara et al., 2023).

The technical progress in this area has been substantial. Aggregation strategies have diversified from the original FEDAVG algorithm (McMahan et al., 2017) towards trust-weighted, hierarchical, asynchronous, and cryptographically secured variants (Nakayiza et al., 2026; Islam et al., 2025; Swarnalatha et al., 2026). Local model architectures have expanded from convolutional and recurrent networks to graph neural networks (Anjum et al., 2025; Al Tfaily et al., 2025) and, most recently, large language model integrations (AlHayan and Al-Muhtadi, 2026). Privacy mechanisms have advanced from structural claims to formal differential privacy with tight composition accounting (Nishad et al., 2026; Islam et al., 2025). Despite this breadth, several evaluative deficiencies persist across the literature. The majority of studies evaluate on simulation under independent and identically distributed (IID) data conditions (Ranpara et al., 2025; García-Sáez et al., 2026); privacy is claimed without formal mechanisms in most papers (Ruzafa-Alcázar et al., 2023); and threat coverage is concentrated on generic multi-class intrusion, leaving high-consequence domains such as healthcare IOT and industrial control systems poorly served (Sharaf and Nooh, 2025; Huong et al., 2022). No PRISMA-compliant systematic synthesis of the field exists to map these strengths and deficiencies together.

Existing surveys on FL for IOT security address adjacent questions but not the intersection at the scope and method of a systematic review (Nguyen et al., 2021). Informal surveys typically cover a subset of papers without formal eligibility criteria, quality assessment, or a pre-specified protocol. Surveys of IOT IDS that include FL treat it as one method among many rather than as the primary focus. The present review fills that gap. It is, to the best of the authors' knowledge, the first PRISMA 2020-compliant systematic literature review of FL applied specifically to privacy-preserving threat detection in massive IOT ecosystems.

Four research questions structure the review:

- **RQ1:** What FL architectures and aggregation strategies have been applied to IOT intrusion detection, and what detection performance has been reported?
- **RQ2:** How is privacy preservation operationalised in FL-based IOT IDS research, and how rigorously are privacy guarantees established?
- **RQ3:** To what extent do the datasets, deployment scales, and evaluation methodologies used match realistic IOT deployment conditions?
- **RQ4:** What threat types and IOT application domains are covered by the corpus, and which are absent or underrepresented?

The answers to these questions rest on a corpus of 61 empirical papers identified through a systematic search of five electronic databases (IEEE Xplore, PubMed, ScienceDirect, SpringerLink, Scopus) covering publications from 2018 to 2026, screened against pre-specified eligibility criteria under PRISMA 2020 guidelines (Page et al., 2021). Each paper was subjected to full-text extraction and quality assessment on six dimensions. The synthesis proceeds across four themes (T1–T4), each aligned with one research question.

The principal contributions of this review are as follows:

- It provides the first PRISMA-compliant synthesis of the FL–IDS literature for IOT, covering 61 papers published across 22 journals between 2020 and 2026.
- It establishes a twelve-subcategory taxonomy of privacy mechanisms, distinguishing formal from structural-only claims.
- It quantifies three evaluation realism deficits: 85% simulation-only evaluation, 80% IID or unstated data partitioning, and a 36.5-percentage-point accuracy gap between balanced and imbalanced evaluation conditions (Ranpara et al., 2025).
- It maps eleven threat categories against eleven application domains to expose the structural coverage gaps in the corpus.
- Finally, it identifies six research gaps and proposes four prioritised directions for future work.

The remainder of this paper is structured as follows: Section 2 describes the search strategy, eligibility criteria, data extraction procedure, and quality assessment framework. Section 3 presents the synthesis results across T1–T4. Section 4 discusses the cross-theme findings, research gaps, implications, and limitations, and draws conclusions.

2. METHODOLOGY

The review conforms to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 (PRISMA 2020) guidelines (Page et al., 2021). Inclusion and exclusion criteria, search strings, and data extraction fields were specified before screening commenced. No deviations from the pre-specified design were made during execution.

2.1 Search Strategy

Six databases were searched: IEEE Xplore, PubMed/MEDLINE, ScienceDirect (Elsevier), SpringerLink, Scopus, and Web of Science. The search was executed on 04 July 2026 and covered publications from 2018 to 2026, the period in which federated learning moved from theoretical proposal to practical deployment in security contexts. No language restriction was applied at the search stage; the eligibility criteria applied subsequently limited inclusion to English-language publications.

The search string combined three conceptual facets joined by Boolean operators. The first facet addressed the learning method: ("federated learning" OR "federated machine learning" OR "collaborative learning"). The second addressed the security function: ("intrusion detection" OR "anomaly detection" OR "threat detection" OR "cyberattack detection" OR "malware detection"). The third addressed the deployment context: ("Internet of Things" OR IOT OR "Industrial IOT" OR IIOT OR "edge computing"). The full string was applied to title, abstract, and keyword fields. Database-specific syntax adjustments were made where required without altering the conceptual scope.

2.2 Screening and Selection

Retrieved records were deduplicated before screening. Title and abstract screening was conducted independently by both authors against the eligibility criteria described in Section 2.3; disagreements were resolved by discussion. Full-text screening was then applied to all records that passed title/abstract review. At the full-text stage, 15 records were excluded and replaced through the protocol's replacement procedure (Section 2.3); one further record was excluded without replacement. Record counts at each stage are reported in Section 3.

2.3 Eligibility Criteria

Inclusion and exclusion criteria were defined a priori using a Population–Intervention–Outcome–Study Design (PIOS) framework adapted for computer science systematic reviews. Table 1 specifies the criteria applied at each screening stage.

Table 1. Eligibility criteria applied at title/abstract and full-text screening stages.

Dimension	Inclusion	Exclusion
Population	Large-scale or heterogeneous IOT deployments (any domain)	Single-device or non-IOT contexts
Intervention	Federated learning applied to threat, intrusion, or anomaly detection	Non-FL distributed learning; FL applied to non-security tasks only
Outcome	At least one quantitative detection metric reported	Purely conceptual or architectural proposals with no evaluation
Study design	Original empirical studies in peer-reviewed journals	Surveys, editorials, secondary literature, conference proceedings
Language	English	Non-English
Date range	2018–2026	Outside range

Three borderline decision rules were defined in advance to standardise judgements at the boundary of the criteria. A study claiming FL but providing no empirical implementation of the federated training cycle was excluded under the Study Design criterion. A study presenting a privacy mechanism without any federated learning component was excluded under Intervention. Where the dataset scale was ambiguous at title/abstract screening, the record was advanced to full-text review; if the client count was not stated in the full text, the record was excluded under Population.

The protocol included a replacement procedure for records excluded at full-text stage. Grounds for exclusion at this stage included retraction, inaccessibility, ineligibility (such as FL applied only as a conceptual framing with no empirical evaluation), and scope mismatch. Each excluded slot was filled by a replacement paper identified through a targeted search of the same database. Replacement candidates and selection rationale were documented before the replacement paper was retrieved. Fifteen replacements were made. One excluded slot could not be filled and was removed from the corpus without replacement.

2.4 Data Extraction

A structured extraction form was applied to all 61 included studies. The form was aligned with the four research questions (RQ1–RQ4) defined

in Section 1 and organised into five field groups: bibliographic details (authors, title, journal, year, DOI); study design characteristics (dataset, client count, communication rounds, hardware platform, simulation vs. physical testbed); FL architecture parameters (aggregation strategy, local model family, non-IID handling approach); privacy mechanism (type, formal guarantee status, reported privacy budget); and quantitative detection metrics (accuracy, precision, recall, F1 score, AUC-ROC where available). A first-mention tracker was maintained across the full extraction sequence to record the earliest corpus appearance of each dataset, technique, and architectural pattern; accurate priority attribution during synthesis depended on this record.

Extracted data were coded against four pre-defined synthesis themes (T1–T4), each mapped to one research question. This mapping governed the structure of both the extraction form and the Results section.

Where a field could not be extracted from the paper text, the value was recorded as not reported rather than left blank or inferred; client count and dataset size were among the fields most frequently absent. Authors were not contacted for clarification. Recorded absences were themselves treated as extractable evidence: the frequency of missing FL configuration fields directly informs the RQ3 analysis of evaluation transparency. Extraction was performed by one author and independently verified by the second; discrepancies were resolved by re-reading the source passage.

2.5 Quality Assessment

Each included study was assigned an overall quality rating of Strong, Adequate, or Weak. Six dimensions were assessed during full-text extraction:

1. Whether at least one quantitative detection metric was tabulated rather than reported in figure form only;
2. Whether the dataset and its partitioning were fully specified;
3. Whether the FL configuration (client count, communication rounds, aggregation algorithm) was reported;
4. Whether a privacy mechanism was present and, if so, whether a formal guarantee was provided or merely claimed;
5. Whether the hardware or simulation environment was described in sufficient detail for replication; and
6. Whether at least one comparison against a non-trivial baseline was included.

A study was rated Strong if it satisfied five or six dimensions without critical gaps. Adequate was assigned where two to four dimensions were met but one or more significant reporting gaps were identified. Weak was assigned where fewer than two dimensions were satisfied, or where a critical methodological concern (such as no tabulated results, or FL not empirically implemented) was present. Quality ratings were assigned by one author and independently verified by the second; all disagreements were resolved by consensus before synthesis commenced. The quality distribution across the 61 included studies is reported in Section 3.

3. RESULTS

3.1 Characteristics of Included Studies

The 61 included studies were published between 2020 and 2026; 42 of 61 appeared in 2025–2026. Publication venues span 22 journals: Scientific Reports (Nature/Springer), IEEE Access, and the IEEE Open Journal of Communications Society are the most frequently represented. Eight broad application domains are covered; general heterogeneous IOT and IIOT account for the largest share of papers. Federated averaging and its variants dominate as the aggregation strategy; 32 studies received a quality rating of Strong, 25 Adequate, and 4 Weak. Table 8 summarises the bibliographic and methodological characteristics of all included studies (Appendix A, Table 8).

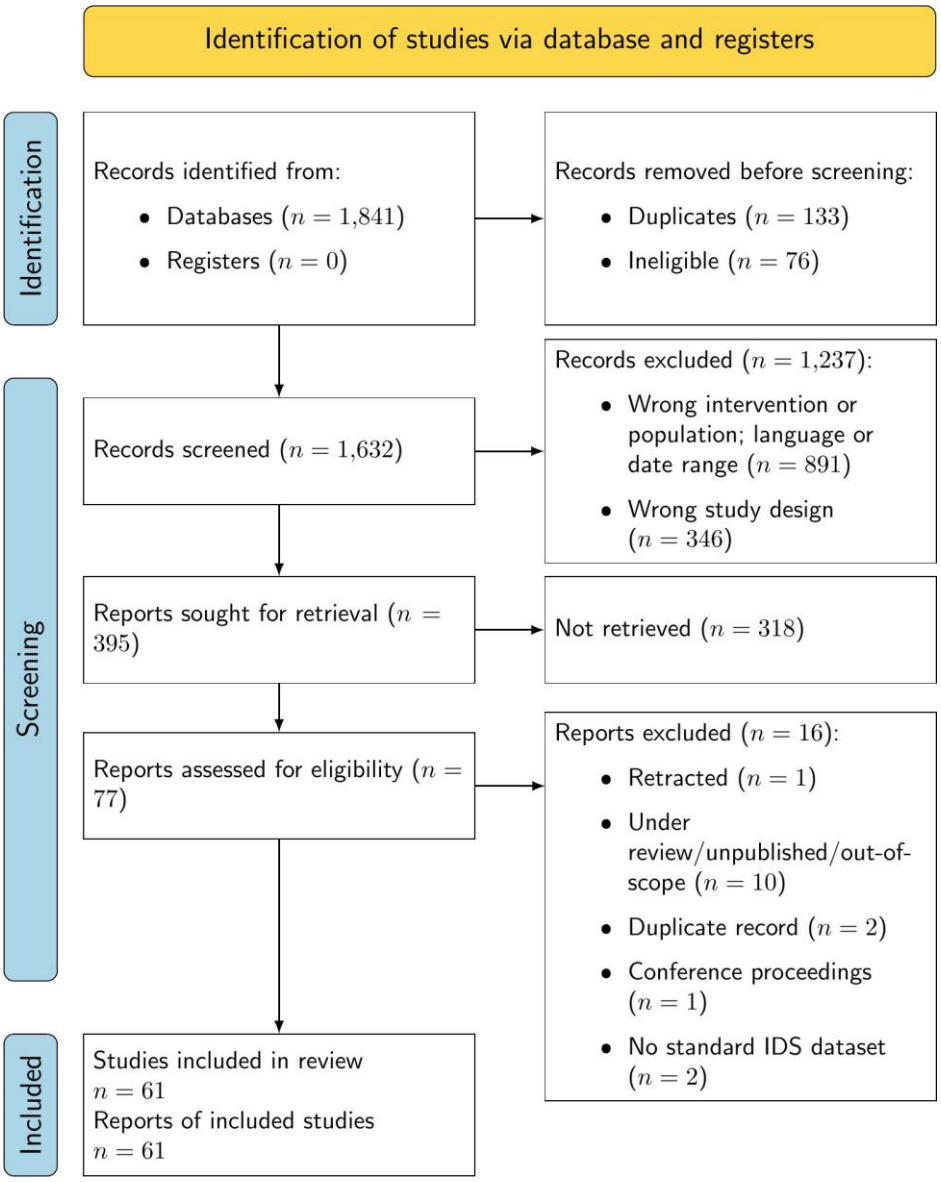

Figure 1. *PRISMA 2020 flow diagram for the systematic search and selection process. Records retrieved from five electronic databases were screened at title/abstract and full-text stages; 15 replacement papers were introduced following full-text exclusions; 61 studies were included in the final synthesis.*

3.2 FL Architectures and Detection Performance (RQ1)

3.2.1 Aggregation Strategies

Standard FEDAVG or a direct extension of it governs aggregation in 40 of the 61 included studies. Table 3 summarises the distribution across strategy families. The dominance of FEDAVG is consistent with the relative maturity of the algorithm and its low communication overhead, but the corpus shows clear movement towards specialised variants as the field has matured: papers published in 2025–2026 account for all hierarchical, asynchronous, and single-paper bespoke variants.

Table 3. Distribution of primary FL aggregation strategies across the 61 included studies.

Aggregation strategy	**n**	**%**
Standard FEDAVG	24	39
FEDAVG variant (personalised, weighted, or trust-based)	16	26
Secure aggregation (HE, DCR+Shamir, or SMPC)	5	8
Hierarchical FEDAVG	3	5
FEDPROX / FEDNOVA / SCAFFOLD	3	5
Asynchronous FEDAVG	3	5
Bespoke single-paper variant (FedAcc, FedDWC, GE2F)	3	5
FEDAVG with knowledge distillation	2	3
Clustered FL	1	2
Non-weight statistic-sharing	1	2
Total	**61**	**100**

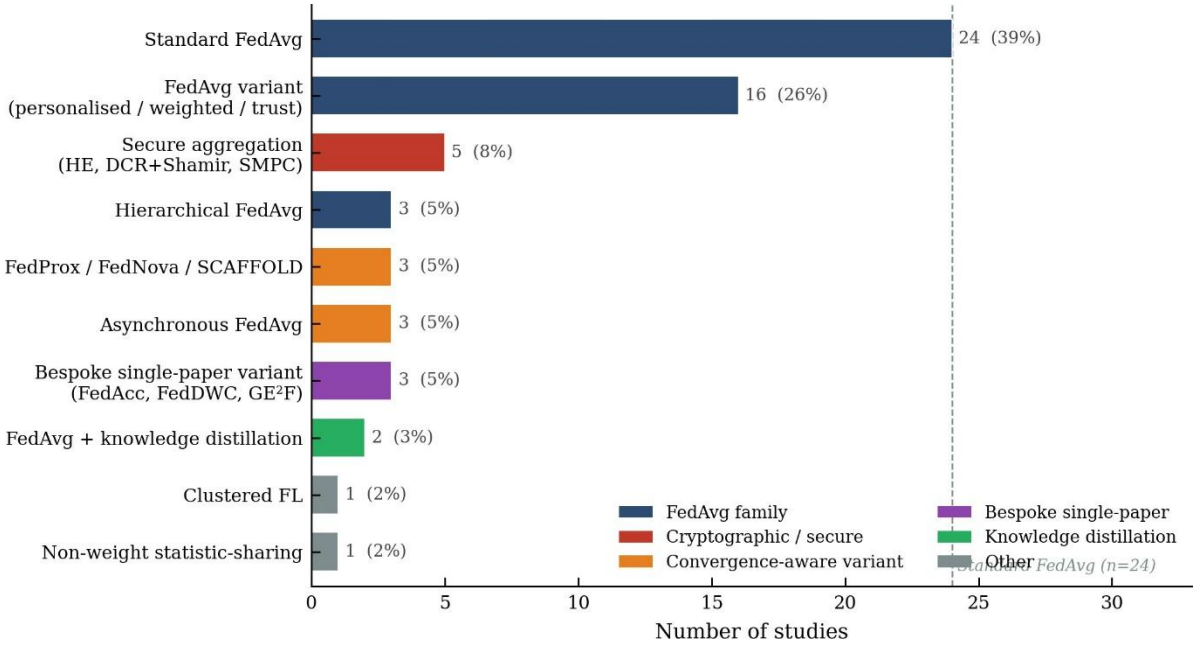

Figure 2. Distribution of primary FL aggregation strategies across the 61 included studies. Percentages are of the total corpus (n = 61). The dashed line marks the standard FEDAVG count (n = 24, 39%). Colour groupings reflect functional families: navy = FEDAVG family; crimson = cryptographic/secure; amber = convergence-aware; plum = bespoke single-paper.

Several notable alternatives to vanilla FEDAVG emerged across the corpus. Sáez-de Cámara et al. (2023) introduced clustered FL with unsupervised model fingerprinting, assigning clients to sub-groups by PCA-reduced weight similarity rather than by topology; this was the first corpus paper to decouple client grouping from any prior knowledge of device types. Attention-weighted aggregation appeared in Chen et al. (2020) (FedAGRU), where a server-side attention mechanism assigned per-client importance scores; the same principle reappeared in later papers as trust-weighted FEDAVG (Alabdulatif, 2025; Babar et al., 2026; Nishad et al., 2026). Asynchronous aggregation, introduced in Bukhari et al. (2024) and Mothukuri et al. (2022), addressed client stragglers without sacrificing convergence; Swarnalatha et al. (2026) extended this by comparing synchronous and asynchronous modes empirically; asynchronous training reduced wall-clock time by 23% on a Kafka-backed streaming pipeline. The most architecturally distinctive aggregation strategy in the corpus is the generalisation-guided GE2F mechanism of Mhawish et al. (2026), which penalises clients whose local validation gap exceeds a threshold before incorporating their update; GE2F is the only corpus paper to exceed its own centralised baseline by a consistent margin across heterogeneous data partitions.

3.2.2 Local Model Families

The corpus draws on a wide range of local model architectures. Convolutional and recurrent networks each appear in roughly a third of papers; CNN variants, LSTM-based models, and GRU architectures are the most common. Autoencoders appear in ten studies, often for unsupervised or semi-supervised anomaly detection where labelled attack traffic is unavailable (Huong et al., 2022; Sáez-de Cámara et al., 2023; Verma et al., 2024). Graph neural networks are adopted by four papers to capture topological relationships among IOT devices (Anjum et al., 2025; Al Mazroa, 2025; Al Tfaily et al., 2025; Sanjalawe et al., 2025). AlHayan and Al-Muhtadi (2026) introduced the first large-language-model integration in the corpus, prepending DISTILBERT embeddings and GPT-2 explanations to a standard LSTM classifier; the pre-mitigation FL failure rate of 74.02% reported in that study is the starkest demonstration in the corpus of how severely standard FEDAVG can degrade under heterogeneous IOT traffic without augmentation.

3.2.3 Detection Performance

Direct cross-paper performance comparison is constrained by dataset fragmentation. No two of the 61 included studies share identical dataset, partitioning scheme, and evaluation protocol. Accuracy figures above 99% appear in 14 papers, but several of these evaluate on near-separable benchmarks or binary classification tasks where high accuracy is expected regardless of the FL design (Hossain et al., 2025; Bhavsar et al., 2024). The most methodologically disciplined performance claims are those evaluated across multiple client scales or data distributions.

Albanbay et al. (2025) swept $N \in \{10, 50, 100, 150\}$ clients on CICIoT2023: CNN maintained F1 > 93% at all scales, while CNN+BiLSTM degraded to 87.82% at 150 clients. Bilal et al. (2026) compared FEDAVG, FEDPROX, and FEDNOVA under consistent conditions; the F1 score dropped by 30 percentage points when models trained on one dataset were evaluated on another. That result is the sharpest cross-domain generalisation gap in the corpus. Ma et al. (2025) reported the widest client-count sweep ($K = 10$–200), quantifying a −1.40-percentage-point accuracy gap between FL and centralised training at $K = 100$ on N-BaIoT.

Two corpus papers exceeded their own centralised baseline: Mhawish et al. (2026) at +0.86 percentage points and Ullah et al. (2026) at +0.22 percentage points. Both attribute the gain to data-centric mechanisms rather than aggregation-level improvements. The ceiling of FEDAVG-class aggregation appears to have been reached; future performance gains in FL-IDS are more likely to come from local data enrichment or model architecture than from aggregation algorithm design. That conclusion carries a caveat: both papers evaluate on a single

dataset, which the 30-pp cross-domain gap reported by Bilal et al. (2026) makes a meaningful constraint on any generalisation claim.

A systemic reporting gap limits T1 synthesis. Fourteen of the 61 papers present results in figure form only, without tabulated numerical values. A further nine do not report all three of client count, communication rounds, and data-partitioning strategy, which prevents reconstruction of experimental conditions. The absence of a standardised FL configuration reporting convention is the corpus's most consequential reproducibility gap.

3.3 Privacy Mechanisms and Rigour (RQ2)

3.3.1 Taxonomy of Privacy Mechanisms

Privacy preservation is operationalised in twelve distinct ways across the 61 included studies. Table 4 presents the full taxonomy. Structural-only privacy (the claim that FL preserves privacy because raw data never leaves client devices) accounts for approximately 49 of 61 papers (80%). Only 12 papers implement a formal or cryptographic mechanism beyond this structural claim.

Table 4. Taxonomy of privacy mechanisms in the 61 included studies.

Sub-category	Papers
Structural-only (FL architectural property)	≈49 studies
Formal DP: fixed ε	S05, S25, S28, S35
Formal DP: Rényi DP via Opacus	S52
Formal DP: Moments Accountant composition	S58
Adaptive gradient noise (no formal bound)	S48
Cryptographic secure aggregation	S07, S12
Homomorphic encryption (Paillier)	S35, S40
HE (CKKS) with dynamic trust scoring	S45
Channel encryption (SSL/TLS)	S06, S17
Blockchain + local DP	S10
Statistic-sharing / minimal disclosure	S08
Quantum-secure authentication	S31

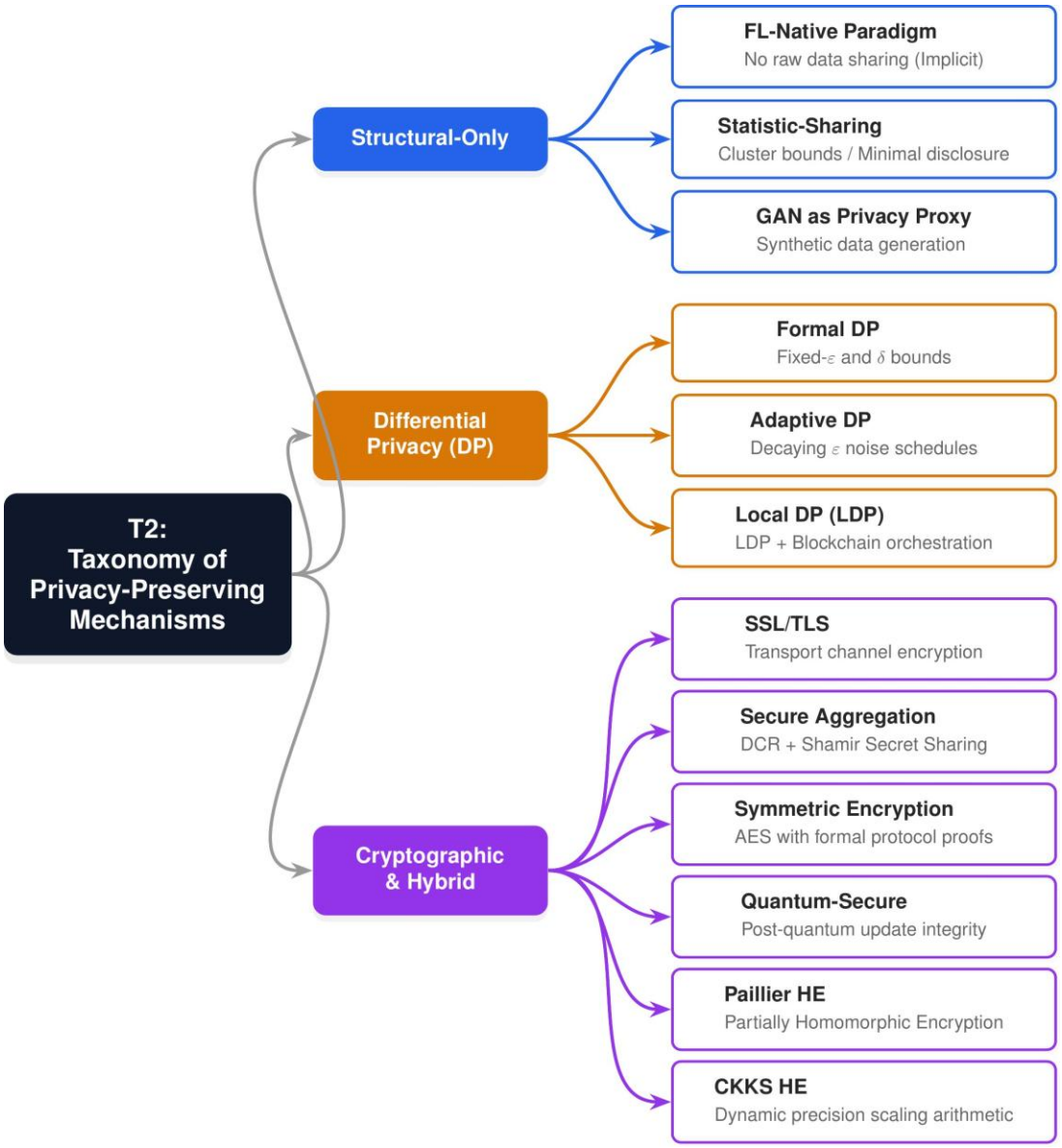

Figure 3. Taxonomy of the twelve privacy-preservation sub-categories identified across the 61 included studies, organised into three mechanism families: structural-only (blue), differential privacy (amber), and cryptographic/hybrid (purple).

3.3.2 Differential Privacy

Differential privacy is the most frequently implemented formal mechanism, appearing in eight papers in some form. The first corpus implementation was Ruzafa-Alcázar et al. (2023), who compared six DP mechanisms applied to logistic-regression weight updates on CIC-ToN-IoT; that paper established both the formal sub-category and the methodological template for subsequent DP evaluations. Six DP mechanisms were compared in that study. Al Mazroa (2025) extended this by sweeping ε ∈ {0.1, 1.0, 10.0} and reporting F1 = 0.963 at all three budgets, the first corpus demonstration that calibrated noise can preserve utility across a wide privacy range.

The accuracy–privacy trade-off is quantified most explicitly by Praveen et al. (2025), whose decaying-ε schedule produced 93.4% accuracy at ε = 1.0 versus 94.7% at ε = 2.0 on TON-IoT; the 1.3-percentage-point gap is the most precisely reported privacy–utility cost in the corpus. Ullah et al. (2026) provided a direct comparison between adaptive and fixed-noise DP: their

GGAP adaptive σ-decay achieved a −0.01% utility loss relative to a non-private baseline, whereas standard fixed-noise DP incurred −5.77%. That comparison carries a caveat: it is the strongest empirical argument against the claim that DP inevitably degrades detection accuracy, yet GGAP provides no formal (ε, δ)-DP bound, which limits its citation rigour in a formal T2 context. Both results are indispensable for any balanced account of DP deployment costs.

The tightest formal DP accounting in the corpus is Nishad et al. (2026), who applied Moments Accountant composition to achieve εtotal = 7.82 over 200 communication rounds (versus ε = 200 under naïve composition). Islam et al. (2025) used Rényi DP via Opacus, reporting an accuracy drop of 1.33–5.78% across ε values on CIC-IoT2023, the only corpus paper to quantify the per-dataset variation in DP overhead. Puviarasu and Sudha (2026) swept nine ε values (0.5–∞), the widest sweep in the corpus, identifying ε ∈ [2, 5] as the optimal operating range for the DP+HE+audit combination evaluated. No other corpus paper evaluates more than three ε values.

3.3.3 Cryptographic and Hybrid Mechanisms

Cryptographic secure aggregation appears in two papers. Tran et al. (2023) implemented DCR-based encryption combined with Shamir secret sharing; the result is a simulation-based security proof against an honest-but-curious aggregator with up to $\tau - 1$ colluding clients; this is the only corpus paper with a formal computational indistinguishability proof. Soomro et al. (2024) applied AES encryption to gradient updates with a formal protocol verification via AVISPA, the only use of an automated security protocol verifier in the corpus.

Homomorphic encryption is implemented in three papers. Alqazzaz (2025a) used Paillier HE with fixed $\varepsilon = 0.7$, $\delta = 10^{-5}$ DP, reporting an F1 of 88.5% on X-IIoTID, approximately 10 percentage points below results from papers on the same dataset without privacy mechanisms, the clearest quantification of the privacy–accuracy cost of HE+DP in the corpus. Nakayiza et al. (2026) extended this to CKKS fully homomorphic encryption with dynamic precision scaling, performing all aggregation in the encrypted domain under RLWE-based IND-CPA security; that paper reported 99.9–100% accuracy, attributing the negligible utility loss to the adaptive precision mechanism. Alqazzaz (2025b) used unnamed lightweight HE with dynamic per-round ε calibration, but neither the HE scheme nor the formal privacy guarantee is stated, which prevents citation for its privacy contribution.

3.3.4 The Formal-Guarantee Gap

The dominant finding of the T2 analysis is the disproportion between privacy claims and formal guarantees. Forty-nine of 61 papers claim privacy structurally; none of those papers quantifies a privacy budget, an adversary model, or an information-theoretic bound. The 12 papers with formal or cryptographic mechanisms are concentrated in later years: all formal DP papers appeared in 2023 or later, and all cryptographic papers in 2023–2026. No paper published before 2023 implemented any mechanism beyond structural claim or channel encryption. The field has moved towards formal rigour, but the movement is slow relative to the volume of publication; at the rate observed in this corpus, structural-only privacy will account for the majority of FL-IDS papers through 2027 at minimum.

3.4 Data and Deployment Realism (RQ3)

3.4.1 Dataset Landscape

Seventy-two dataset appearances are recorded across the 61 included studies; papers using multiple datasets account for the excess over 61. Table 5 summarises the nine most frequently used datasets. CICIoT2023 leads with 11 appearances, followed by ToN-IoT and NSL-KDD at nine each. No single dataset appears in more than 18% of papers. That fragmentation prevents direct cross-paper performance comparison across the corpus even when the same model family or aggregation strategy is used.

Table 5. Dataset frequency across the 61 included studies.

Dataset	**Papers**	**%**
CICIoT2023	11	18
ToN-IoT / TON-IoT	9	15
NSL-KDD	9	15
UNSW-NB15	7	11
N-BaIoT	6	10
X-IIoTID	4	7
CICIDS2017	4	7
Edge-IIoTset	3	5
Other (single-paper use)	11	18
Total appearances	**64**	--

Four corpus papers use unnamed or unidentifiable datasets. Sharaf and Nooh (2025) described an IoMT dataset without naming it; Farooq et al. (2025) used a KDD-schema inferred only from column structure; Ragab et al. (2025) labelled 12 attack classes L-1 to L-12 with no provenance stated; Alqazzaz (2025b) referred to smart city datasets without identification. None of these papers can be reproduced or directly compared against other corpus work. The gap constitutes a reproducibility failure that peer review did not prevent.

A further cross-paper comparability concern arises from inconsistent use of the same dataset. CICIoT2023 appears in 11 papers in at least three distinct configurations: Ullah et al. (2026) used the full 34-class taxonomy, then a 7-class and an 8-class regrouping; Bilal et al. (2026) used an 8-class regrouping; other papers used the default class distribution without stating it. Results reported against different configurations of the same dataset are not directly comparable, which the corpus treats as interchangeable in all but one paper.

3.4.2 Non-IID Data Handling

Twelve of 61 studies address non-IID data distribution in any meaningful way. The remaining 49 use IID partitioning or do not state the partitioning scheme. In an IOT deployment, data at individual nodes is inherently heterogeneous by device type, traffic pattern, and attack exposure; IID evaluation is an optimistic condition that overstates real-world performance.

The most methodologically thorough non-IID analysis in the corpus is García-Sáez et al. (2026), who applied a Dirichlet α-sweep across multiple values and quantified a 2.94-percentage-point accuracy gain under high heterogeneity relative to IID training; that result is the strongest evidence in the corpus that non-IID-aware aggregation can recover accuracy lost to data skew. Mhawish et al. (2026) quantified non-IID heterogeneity via Total Variation Distance before aggregation, the only corpus paper to measure the degree of heterogeneity rather than simply inducing it. TVD-based pre-aggregation assessment is a meaningful methodological advance. Sáez-de Cámara et al. (2023) used an inherently non-IID emulated testbed without artificial partitioning, which is methodologically preferable to Dirichlet sampling but not scalable beyond 78 clients. Al Mazroa (2025) evaluated non-IID performance on 20 clients with Dirichlet $\alpha = 0.3$ and reported cross-domain transfer asymmetry of F1 = 0.39 in one direction versus F1 = 0.99 in the other, the starkest cross-domain generalisation gap measured on heterogeneous data in the corpus.

3.4.3 Deployment Scale and Testbed Realism

Seven papers used physical IOT hardware; two used emulated testbeds; the remaining 52 are simulation-only. Physical hardware deployments include Raspberry Pi 4 (Huong et al., 2022; Jithish et al., 2023; Al Mazroa, 2025), Jetson Nano/Xavier (Al Mazroa, 2025; Agrawal et al., 2024), Arduino Nano 33 BLE (Hajj et al., 2023), and an i5/Xeon workstation cluster (Issa et al., 2026). The most resource-constrained evaluation is Hajj et al. (2023), whose Arduino Nano platform is the smallest device class in the corpus. The most energy-explicit evaluation is Al Mazroa (2025), who reported 0.42 J/inference on Raspberry Pi 4 and 0.18 J/inference on Jetson Nano, the only corpus paper to provide per-inference energy figures across two hardware platforms.

Client count is the most commonly overstated deployment parameter. Twenty-three papers evaluate on ten or fewer clients; only four exceed 100 clients. Ma et al. (2025) swept $K \in \{10, 50, 100, 150, 200\}$, the widest range in the corpus. Peng et al. (2025) scaled to 1,000 simulated clients, the corpus high for simulation scale, but did not evaluate non-IID distribution at that count. The gap between the largest simulation (K = 1,000) and the largest physical deployment (K = 78, Sáez-de Cámara et al. 2023) illustrates the evaluation ceiling imposed by hardware cost: no corpus paper combines both massive scale and physical testbed evidence.

Ranpara et al. (2025) provided the corpus's most explicit controlled-to-realistic degradation figure: 96% accuracy on a balanced partition versus 59.50% on an imbalanced real-world SDN trace, a 36.5-percentage-point drop that quantifies the evaluation inflation introduced by IID and balanced assumptions. No other corpus paper states this gap so directly. Sorour et al. (2025) achieved the most operationally authentic client partitioning, assigning one FL client per device sub-dataset in a six-device heterogeneous suite, but hardware was not specified and the JSO computation cost of 420,000 seconds was not contextualised against a deployment target.

3.4.4 The Realism Deficit

The T3 findings converge on a single gap. Fifty-two of 61 papers (85%) evaluate on simulation only. Forty-nine of 61 (80%) use IID partitioning or do not state the partition scheme. Simulation under IID conditions describes the default evaluation methodology of the corpus and produces results that cannot be taken as evidence of deployability in real IOT environments. The seven physical-testbed papers and twelve non-IID papers supply the corpus's realistic empirical anchor; Figure 4 presents all three dimensions of this realism deficit. The remaining papers establish performance upper bounds, not operational baselines.

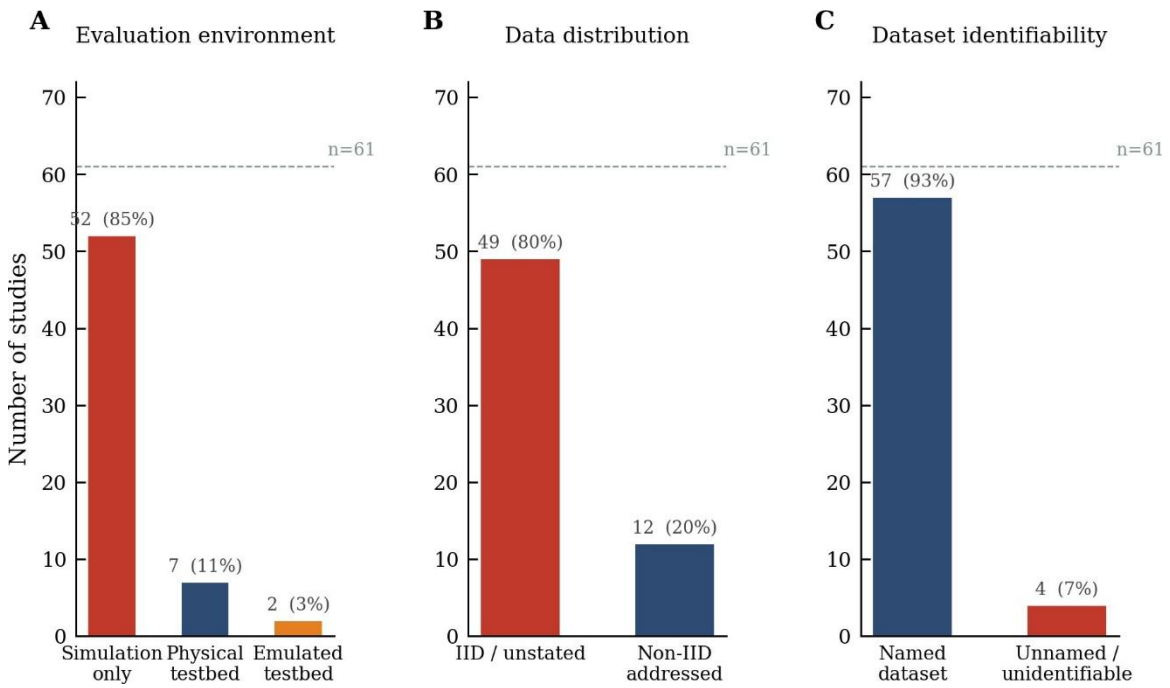

Figure 4. Evaluation realism across three dimensions: (A) testbed environment, (B) data distribution handling, and (C) dataset identifiability. Simulation under IID conditions on named datasets is the dominant evaluation methodology (n values shown with percentage of total corpus).

3.5 Threat Coverage and Domain Spread (RQ4)

3.5.1 Threat Type Distribution

Eleven threat categories are represented across the 61 included studies. Table 6 presents the distribution. Generic network intrusion or multi-class attack detection is the primary focus of 35 papers (57%), reflecting the dominance of benchmark datasets whose class taxonomies define the threat scope rather than any deliberate threat selection by authors. DDOS and botnet/IoT malware each appear as dedicated foci in eight papers; they are the best-represented specific threat classes in the corpus. Generic multi-class intrusion dominates by volume.

Table 6. Primary threat categories across the 61 included studies.

Threat category	**Papers**
Network intrusion / multi-class detection	35
DDOS (dedicated evaluation)	8
Botnet / Mirai / IoT malware	8
FL-process poisoning (adversarial FL)	7
ICS / SCADA / FDIA	3
CAN bus / vehicular attack	3

Threat category	Papers
Zero-day / open-set detection	3
Ransomware / crypto-malware	2
MQTT protocol attacks	1
SDN misbehaviour	1
Anomaly detection (unlabelled)	1

FL-process poisoning (backdoor injection, label-flipping, and scaled gradient attacks) is treated as a distinct threat category in seven papers. This distinction matters. FL poisoning is an attack on the training process, not on the monitored network; papers addressing it evaluate the security of the FL system itself, not the detection of external intrusions. Seven papers is a meaningful count. Sanjalawe et al. (2025) evaluated four simultaneous poisoning types; Nakayiza et al. (2026) reported the highest attack success rate in the corpus under scaled poisoning (ASR = 0.65 at 50% malicious clients), the starkest evidence that trust-weighted aggregation is not immune to coordinated manipulation at high adversarial fractions.

Zero-day and open-set detection appears in three papers. Verma et al. (2024) was the first corpus paper to evaluate zero-day generalisation; Mhawish et al. (2026) subsequently introduced the CATS/R zero-day evaluation protocol, the only structured zero-day test procedure in the corpus. Ransomware and crypto-malware appear in two papers only (Soomro et al., 2024; Hossain et al., 2025), despite constituting a major and growing class of real-world IoT threat. Physical-layer and supply-chain attacks are absent from the corpus entirely.

3.5.2 Application Domain Coverage

The corpus covers eleven application domains. General heterogeneous IOT accounts for 29 papers (48%), a category so broad it provides little domain specificity. IIOT/industrial deployments are the next largest group at 14 papers (23%). Table 7 summarises the distribution.

Table 7. Application domain distribution across the 61 included studies.

Application domain	Papers
General / heterogeneous IOT	29
IIOT / industrial	14
Smart city / WSN	5
Vehicular / IOV	4
Fog / edge IOT	4

Application domain	Papers
Smart grid / energy	2
Healthcare IOT	2
6G / next-generation IOT	2
ICS / SCADA	1
MQTT / protocol-specific	1
SDN	1

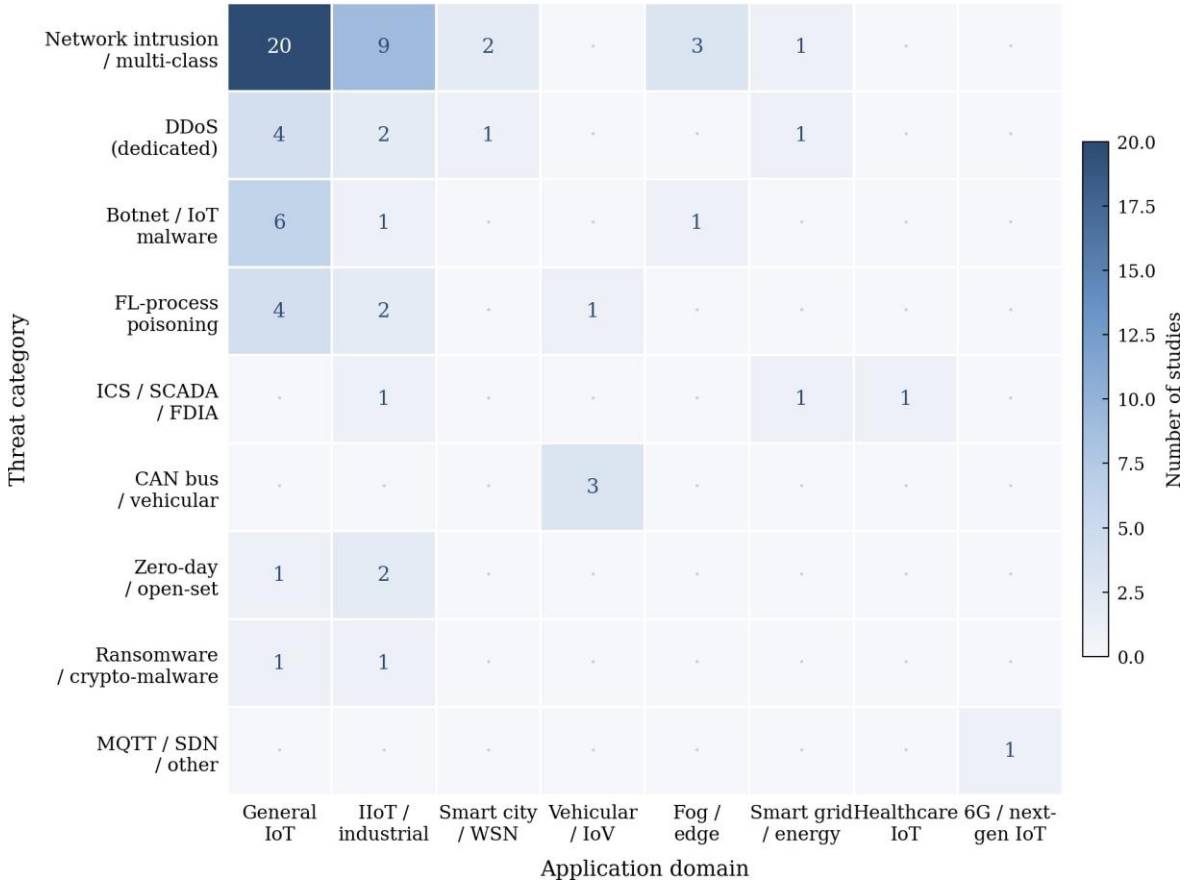

Figure 5. Threat category by application domain co-occurrence across the 61 included studies. Cell values are paper counts; dots mark zero coverage. Concentration on generic network intrusion in general IOT contexts is visible; healthcare, smart grid, and zero-day detection remain sparsely covered.

Several high-risk domains are sparsely represented relative to their real-world attack surface. Healthcare IOT appears in two papers only: Sharaf and Nooh (2025) addressed clinical IoMT security and Babar et al. (2026) targeted 6G-connected healthcare devices. Smart grid security, despite being among the most consequence-critical IOT deployment contexts,

is the primary focus of two papers (Jithish et al., 2023; Tran et al., 2023). ICS/SCADA is addressed in one paper (Huong et al., 2022). Agricultural IOT, water-infrastructure monitoring, and maritime systems are absent from the corpus entirely.

Domain specificity is a further concern. Twenty-nine papers describe their target as "general" or "heterogeneous" IOT without specifying a deployment context. In the absence of a stated domain, the threat model, data distribution, and client heterogeneity assumptions are left implicit; results from such papers are difficult to apply to any concrete deployment. The proportion of domain-unspecified papers (48%) is itself a finding: the majority of FL-IDS research has not yet committed to a target operational context.

3.5.3 Attack Taxonomy Depth

The largest attack taxonomy in the corpus is the 34-class CICIoT2023 configuration used by Ullah et al. (2026). Most papers report far fewer classes: binary classification (normal vs attack) appears in at least 14 papers, and six-to-eight-class taxonomies are the most common multi-class configuration. Binary classification inflates accuracy figures and obscures detection capability at the sub-class level; a system that distinguishes only normal from anomalous traffic cannot be deployed where specific threat identification is operationally required.

Three papers evaluated attack detection latency. Nishad et al. (2026) reported per-attack-class latency ranging from 45 ms for DOS to 189 ms for zero-day, the most detailed latency breakdown in the corpus. Nakayiza et al. (2026) reported end-to-end blockchain consensus latency of 1.79 s at 60 clients. No other corpus paper reports inference or detection latency, despite it being a primary operational constraint for real-time IOT intrusion response. The absence of latency measurement across 58 of 61 papers is a gap between research evaluation and deployment requirement that T4 synthesis cannot bridge.

3.5.4 Coverage Gaps

Three structural gaps define the T4 findings:

1. Threat concentration: 57% of papers address generic multi-class intrusion, while high-consequence specific threats such as ransomware, physical-layer attacks, and supply-chain compromise are marginal or absent.
2. Domain concentration: 48% of papers use a generic IOT framing that does not correspond to any operational deployment, and sectors with the highest security consequence (healthcare, critical infrastructure, smart grid) account for five papers combined.
3. Evaluation depth: binary classification and absence of latency reporting are endemic across the corpus; such a system cannot be deployed where specific threat identification is operationally required.

4. DISCUSSION AND CONCLUSION

4.1 Cross-Theme Synthesis

The four research questions addressed in this review converge on a coherent picture of a field that is technically inventive but evaluatively immature. Four headline findings structure the synthesis:

- **RQ1** finds that FL aggregation for IOT intrusion detection is dominated by FEDAVG and its direct variants; decision diversity is concentrated in the most recent papers.
- **RQ2** finds that privacy is claimed in 80% of papers on structural grounds alone; formal mechanisms appeared only in 2023 and later.
- **RQ3** finds that 85% of papers evaluate on simulation only under IID conditions, producing results that do not carry evidence of deployability.
- **RQ4** finds that threat coverage is concentrated on generic multi-class intrusion in unspecified IOT contexts; high-consequence domains and specific

threat classes are systematically underrepresented.

Three cross-theme tensions define the field's current position. The first is the tension between aggregation sophistication and evaluation simplicity. Papers in the corpus introduce architecturally ambitious mechanisms: CKKS fully homomorphic encryption, PBFT blockchain consensus, trust-weighted FL with dynamic precision scaling. Yet those same papers evaluate them on small client counts, IID data, and single datasets. The sophistication of the mechanism and the realism of the evaluation move in opposite directions. The second tension is between privacy ambition and formal rigour. Every paper in the corpus claims privacy as a contribution; 49 of 61 provide no formal mechanism to justify that claim. The third tension is between breadth of threat coverage and depth of evaluation: papers address an ever-wider range of attack types but evaluate fewer classes per paper; binary classification is the modal configuration.

A structural finding cuts across all four themes: the corpus has no agreed reporting standard. Client count, communication rounds, data partitioning scheme, dataset configuration, privacy budget, and inference latency are all reported inconsistently or not at all. Without a minimum reporting checklist, results from different papers address different implicit experimental conditions and cannot be compared or aggregated.

4.2 Research Gaps

Six research gaps emerge directly from the synthesis findings. Four concern what the corpus does not address; two concern how it addresses what it does. Each gap is stated as a claim that the corpus cannot currently answer.

1. **Massive-scale validation:** The largest physical deployment in the corpus is 78 clients (Sáez-de Cámara et al., 2023); the largest simulation is 1,000 clients (Peng et al., 2025). Neither approaches the device count of a real IOT deployment at city or industrial scale. No corpus paper combines both a client count above 100 and a physical or emulated testbed.
2. **Formal privacy at scale:** All formal DP evaluations in the corpus use ten or fewer clients in the privacy evaluation context. Whether differential privacy guarantees hold under realistic client counts, non-IID data, and heterogeneous device capability remains unanswered by the corpus. The Moments Accountant analysis in Nishad et al. (2026) is the closest approach, but it is conducted on a simulation without physical hardware.
3. **Domain specificity:** Forty-eight of 61 papers (48%) address a generic IOT context. Healthcare IOT, industrial control systems, smart grid, and vehicular networks each have distinct threat models, regulatory constraints, and real-time requirements. Research that does not specify a deployment context cannot produce claims specific enough to be actionable by practitioners in those domains.
4. **Operational evaluation:** Three of 61 papers report detection latency. None reports energy consumption at a client count above 20. No paper reports memory footprint per client alongside detection accuracy and privacy cost. An FL-IDS system that achieves 99% detection accuracy but requires 2.4 GB RAM per edge node (Al Tfaily et al., 2025) is not deployable on standard IOT hardware; the corpus does not yet engage with this constraint systematically.
5. **Cross-dataset generalisation:** Only two papers evaluate cross-domain transfer (Al Mazroa, 2025; Alsaleh et al., 2025); both find large F1 drops when models are transferred. The field has not addressed whether a model trained on CICIoT2023 is useful on ToN-IoT, or whether FL provides any advantage over centralised training in cross-domain settings.
6. **Reporting standard:** The absence of a minimum FL configuration reporting convention is not a research gap in the usual sense: it does not require new experiments. It requires a community norm. The analogous problem in clinical trials was resolved by CONSORT; in

machine learning, to a partial degree, by ML-Reproducibility Checklist initiatives. No equivalent exists for FL-IDS.

4.3 Implications for Practice and Future Research

For practitioners deploying FL-IDS in IOT environments, the synthesis findings support three conclusions. Aggregation algorithm selection is not the primary performance determinant: the corpus shows that FEDAVG and its trust-weighted or hierarchical variants perform comparably on the benchmarks used, and the two papers that exceeded their centralised baseline attribute the gain to data-centric mechanisms, not aggregation design. Privacy mechanism selection does matter: the 1.3-percentage-point accuracy cost of formal DP at $\varepsilon = 1.0$ (Praveen et al., 2025) is small, and the −5.77% cost of fixed-noise DP (Ullah et al., 2026) is avoidable through adaptive noise scheduling; structural-only privacy claims provide no measurable protection against inference attacks and should not satisfy regulatory requirements. Evaluation conditions matter more than any algorithm: results from IID simulation on single datasets overstate deployable performance, and Ranpara et al. (2025)'s 36.5-percentage-point drop from balanced to imbalanced evaluation should be treated as a baseline adjustment factor for any corpus result produced without non-IID and multi-class conditions.

For future research, four directions are identified as highest priority:

1. Physical-testbed evaluation at $K > 100$ clients, combined with formal DP guarantees and non-IID data partitioning, would address gaps one and two simultaneously.
2. Domain-specific FL-IDS benchmarks for healthcare, ICS/SCADA, and vehicular networks would address gap three and unlock domain-appropriate threat modelling.
3. Standardised FL configuration reporting (a four-field minimum of client count, communication rounds, aggregation algorithm, and data partition method) would address gap six without requiring new experiments.
4. Cross-domain transfer evaluation, combining at least two datasets from different IOT sectors, would address gap five and provide the only empirically supported basis for claims of generalisation.

4.4 Limitations of This Review

Several limitations bound the conclusions of this review. The search was conducted on five databases on a single date; studies indexed after 04 July 2026 are not represented. The 2018–2026 date range excludes earlier theoretical foundations of FL that motivate current designs. Fifteen papers were replaced at full-text stage under the protocol's replacement procedure; the excluded papers may contain findings that would alter the synthesis. Quality assessment was conducted by two authors without inter-rater reliability coefficients reported, which is a known limitation of systematic reviews in computer science. The synthesis is qualitative rather than meta-analytic; the dataset fragmentation documented in Section 3.4 precludes pooled quantitative estimates. Finally, the corpus is journal-only; conference proceedings are excluded, which may underrepresent emerging techniques that appear in conferences before journals.

4.5 Conclusion

This review synthesised 61 peer-reviewed studies on federated learning for privacy-preserving threat detection in IOT ecosystems, published between 2020 and 2026. The corpus spans 22 journals and eleven application domains. Across four research questions, the synthesis identifies a field with a strong algorithmic base and a weak empirical foundation. FEDAVG and its variants govern aggregation in 65% of papers, and the literature is converging on trust-weighted and hierarchical extensions for heterogeneous deployments. Privacy formalisation is improving but is still the

exception: 80% of papers claim privacy on structural grounds only. Evaluation realism is the most consequential deficit: simulation under IID conditions describes the default methodology, and the 36.5-percentage-point gap between balanced and realistic evaluation in Ranpara et al. (2025) quantifies what that default costs. Threat coverage is widening but shallow; generic intrusion detection in unspecified IOT contexts accounts for half the corpus, and high-consequence domains are underserved.

The six research gaps identified in Section 4.2 define the immediate agenda. Massive-scale physical testbed evaluation with formal privacy guarantees and non-IID data is the highest-priority experimental direction. A community reporting standard for FL configuration is the highest-priority methodological direction. Progress on either front would substantially improve the field's ability to produce results that transfer from the laboratory to operational IOT deployments.

5. REFERENCES

Mohamed Abd Elaziz, Ibrahim A. Fares, Abdelghani Dahou, and Mansour Shrahili. Federated learning framework for IoT intrusion detection using tab transformer and nature-inspired hyperparameter optimisation. *Frontiers in Big Data*, 8:1526480, 2025. doi:10.3389/fdata.2025.1526480.

Smita Agrawal, Subhrajit Sarkar, Gautam Srivastava, and Uttam Ghosh. Federated learning for decentralized DDoS attack detection in IoT networks. *IEEE Access*, 12:39810–39825, 2024. doi:10.1109/ACCESS.2024.3378727.

Ijaz Ahmad, Asadullah Shaikh, Jawad Ahmad, Salman Alqahtani, Shtwai Alsubai, and Adel Binbusayyis. FFL-IDS: a fog-enabled federated learning-based intrusion detection system to counter jamming and spoofing attacks in IoT. *Sensors*, 25(1):10, 2025. doi:10.3390/s25010010.

Alanoud Al Mazroa. FORT-IDS: a federated, optimized, robust and trustworthy intrusion detection system for IIoT security. *Scientific Reports*, 15, 2025. doi:10.1038/s41598-025-31025-x.

Fouad Al Tfaily, Zakariya Ghalmane, Mohamed el Amine Brahmia, Hussein Hazimeh, Ali Jaber, and Mourad Zghal. Graph-based federated learning approach for intrusion detection in IoT networks. *Scientific Reports*, 15:41264, 2025. doi:10.1038/s41598-025-25175-1.

Abdulatif Alabdulatif. FedCognis: an adaptive federated learning framework for secure anomaly detection in industrial IoT-enabled cognitive cities. *Computers, Materials & Continua*, 2025. doi:10.32604/cmc.2025.066898.

Nurtay Albanbay, Yerlan Tursynbek, Kalman Graffi, Raissa Uskenbayeva, Zhuldyz Kalpeyeva, Zhastalap Abilkaiyr, and Yerlan Ayapov. Federated learning-based intrusion detection in IoT networks: Performance evaluation and data scaling study. *Journal of Sensor and Actuator Networks*, 14(4):78, 2025. doi:10.3390/jsan14040078.

Haya Saleh Alharthi, Suhair Alshehri, and Manal Kalkatawi. Blockchain-enabled hierarchical federated learning framework for anomaly detection in IoT systems. *Applied Sciences*, 15(24):13037, 2025. doi:10.3390/app152413037.

Abdullah AlHayan and Jalal Al-Muhtadi. Federated learning-powered real-time behavioral intrusion detection leveraging LSTM, attention, GANs, and large language models. *Scientific Reports*, 2026. doi:10.1038/s41598-026-40763-5.

Shahad Almansour, Kusum Yadav, Lulwah M. Alkwai, Norah Saleh Alghamdi, Wattana Viriyasitavat, and Gaurav Dhiman. Adaptive personalized federated learning with lightweight depthwise convolutional bottleneck network for novel intrusion detection system in internet of vehicles. *Scientific Reports*, 15, 2025. doi:10.1038/s41598-025-17699-3.

Ali Alqazzaz. SecuFL-IoT: an adaptive privacy-preserving federated learning framework for anomaly detection in smart industrial networks. *Scientific Reports*, 15, 2025a. doi:10.1038/s41598-025-11883-1.

Ali Alqazzaz. Federated learning with homomorphic encryption: a privacy-preserving solution for smart cities. *International Journal of*

Computational Intelligence Systems, 18:304, 2025b. doi:10.1007/s44196-025-00829-0.

Shuroog Alsaleh, Mohamed El Bachir Menai, and Saad Al-Ahmadi. A heterogeneity-aware semi-decentralized model for a lightweight intrusion detection system for IoT networks based on federated learning and BiLSTM. *Sensors*, 25(4):1039, 2025. doi:10.3390/s25041039.

Mohd Anjum, Ashit Kumar Dutta, Ali Elrashidi, Sana Shahab, Asma Aldrees, Zaffar Ahmed Shaikh, and Abeer Aljohani. GraphFedAI framework for DDoS attack detection in IoT systems using federated learning and graph neural networks. *Scientific Reports*, 15:9447, 2025. doi:10.1038/s41598-025-10826-0.

Dinesh Chowdary Attota, Viraaji Mothukuri, Reza M. Parizi, and Seyedamin Pouriyeh. An ensemble multi-view federated learning intrusion detection for IoT. *IEEE Access*, 9:117734–117745, 2021. doi:10.1109/ACCESS.2021.3107337.

Muhammad Babar, Zahid Khan, Sarah Kaleem, Basit Qureshi, and Wadii Boulila. TAWB-SFL: trust-aware blockchain-orchestrated split federated learning for intrusion detection in 6G healthcare IoT. *IEEE Open Journal of Communications Society*, 2026. doi:10.1109/OJCOMS.2026.3667693.

Yonas Kibret Beshah, Surafel Lemma Abebe, and Henock Mulugela Melaku. Dynamic weight clustered federated learning for IoT DDoS attack detection. *Scientific Reports*, 15, 2025. doi:10.1038/s41598-025-13204-y.

Mansi H. Bhavsar, Yohannes B. Bekele, Kaushik Roy, John C. Kelly, and Daniel Limbrick. FL-IDS: federated learning-based intrusion detection system using edge devices for transportation IoT. *IEEE Access*, 12:52215–52230, 2024. doi:10.1109/ACCESS.2024.3386631.

Muhammad Ahmad Bilal, Ihtesham Ul Islam, Sarmad Idrees, Muhammad Qasim, Muhammad Junaid Khan, and Jaleed Khan. Dataset-centric evaluation of federated intrusion detection models in IoT networks. *Scientific Reports*, 2026. doi:10.1038/s41598-025-32567-w.

Wayoud Bouzeraib, Afifa Ghenai, and Nadia Zeghib. Enhancing IoT intrusion detection systems through horizontal federated learning and optimized WGAN-GP. *IEEE Access*, 13:45059–45075, 2025. doi:10.1109/ACCESS.2025.3547255.

Syed Muhammad Salman Bukhari, Muhammad Hamza Zafar, Mohamad Abou Houran, Zakria Qadir, Syed Kumayl Raza Moosavi, and Filippo Sanfilippo. Enhancing cybersecurity in edge IIoT networks: an asynchronous federated learning approach with a deep hybrid detection model. *Internet of Things*, 27:101252, 2024. doi:10.1016/j.iot.2024.101252.

Nisha Chaurasia, Munna Ram, Priyanka Verma, Nakul Mehta, and Nitesh Bharot. A federated learning approach to network intrusion detection using residual networks in industrial IoT networks. *The Journal of Supercomputing*, 80:18325–18346, 2024. doi:10.1007/s11227-024-06153-2.

Zhuo Chen, Na Lv, Pengfei Liu, Yu Fang, Kun Chen, and Wu Pan. Intrusion detection for wireless edge networks based on federated learning. *IEEE Access*, 8:217463–217472, 2020. doi:10.1109/ACCESS.2020.3041793.

Lambert Kofi Gyan Danquah, Stanley Yaw Appiah, Victoria Adzovi Mantey, Iddrisu Danlard, and Emmanuel Kofi Akowuah. Computationally efficient deep federated learning with optimized feature selection for IoT botnet attack detection. *Intelligent Systems with Applications*, 25:200462, 2025. doi:10.1016/j.iswa.2024.200462.

Manu Devi, Priyanka Nandal, and Harkesh Sehrawat. Federated learning-enabled lightweight intrusion detection system for wireless sensor networks: a cybersecurity approach against DDoS attacks in smart city environments. *Intelligent Systems with Applications*, 25:200553, 2025. doi:10.1016/j.iswa.2025.200553.

Abebe Abeshu Diro and Naveen Chilamkurti. Distributed attack detection scheme using deep learning approach for Internet of Things. *Future Generation Computer Systems*, 82:761–768, 2018. doi:10.1016/j.future.2017.08.043.

Maha Driss, Iman Almomani, Zil e Huma, and Jawad Ahmad. A federated learning framework for cyberattack detection in vehicular sensor networks. *Complex & Intelligent Systems*, 8:3569–3579, 2022. doi:10.1007/s40747-022-00705-w.

Muhammad Sajid Farooq, Muhammad Saleem, M.A. Khan, Muhammad Farrukh Khan, Shahan Yamin Siddiqui, Muhammad Shoukat Aslam, and Khan M. Adnan. Interpretable federated learning model for cyber intrusion detection in smart cities with privacy-preserving feature selection. *Computers, Materials & Continua*, 2025. doi:10.32604/cmc.2025.069641.

Luis Miguel García-Sáez, Sergio Ruiz-Villafranca, José Roldán-Gómez, Javier Carrillo-Mondéjar, and José Luis Martínez. Hybrid clustering-guided federated learning for robust intrusion detection in highly heterogeneous IoT environments. *Computer Networks*, 281:112205, 2026. doi:10.1016/j.comnet.2026.112205.

Suzan Hajj, Joseph Azar, Jacques Bou Abdo, Jacques Demerjian, Christophe Guyeux, Abdallah Makhoul, and Dominique Ginhac. Cross-layer federated learning for lightweight IoT intrusion detection systems. *Sensors*, 23(16):7038, 2023. doi:10.3390/s23167038.

Md. Alamgir Hossain, Sadman Saif, and Md. Saiful Islam. A novel federated learning approach for IoT botnet intrusion detection using SHAP-based knowledge distillation. *Complex & Intelligent Systems*, 11, 2025. doi:10.1007/s40747-025-02001-9.

Truong Thu Huong, Ta Phuong Bac, Kieu Ngan Ha, Nguyen Viet Hoang, Nguyen Xuan Hoang, Nguyen Tai Hung, and Kim Phuc Tran. Federated learning-based explainable anomaly detection for industrial control systems. *IEEE Access*, 10:53854–53873, 2022. doi:10.1109/ACCESS.2022.3173288.

Razi Iqbal, Shereen Ismail, and Geetanjali Rathee. Hardware aware federated learning with sparse models for adaptive anomaly detection in IoT. *Discover Artificial Intelligence*, 2026. doi:10.1007/s44163-026-01171-w.

Md Morshedul Islam, Wali Mohammad Abdullah, and Baidya Nath Saha. Privacy-preserving hierarchical fog federated learning (PP-HFFL) for IoT intrusion detection. *Sensors*, 25(23):7296, 2025. doi:10.3390/s25237296.

Mohannad Abu Issa, Abdelrahman Eldosouky, Ashraf Matrawy, and Mohamed Ibnkahla. Multi-temporal device clustering for federated learning–Internet of Things intrusion detection systems. *IEEE Transactions on Machine Learning in Communications and Networking*, 2026. doi:10.1109/TMLCN.2026.3706262.

Xiaojun Jin, Chao Ma, Song Luo, Pengyi Zeng, and Yifei Wei. Distributed IIoT anomaly detection scheme based on blockchain and federated learning. *Journal of Communications and Networks*, 26(2):252–262, 2024. doi:10.23919/JCN.2024.000016.

J. Jithish, Bithin Alangot, Nagarajan Mahalingam, and Kiat Seng Yeo. Distributed anomaly detection in smart grids: a federated learning-based approach. *IEEE Access*, 11:7157–7179, 2023. doi:10.1109/ACCESS.2023.3237554.

Linda Joseph, B. Prabha, and M. Sambath. TrustFed-RHIO: an optimization-driven differential privacy federated learning framework for secure and explainable IIoT attack detection. *Scientific Reports*, 2026. doi:10.1038/s41598-026-45277-8.

Sajjad Khan and Davor Svetinovic. Adaptive privacy-preserving federated learning for robust IoT systems: a defence against data poisoning attacks. *IoT and Cyber-Physical Systems*, 2026. doi:10.1016/j.iotcps.2026.03.006.

Ansam Khraisat, Ammar Alazab, Moutaz Alazab, Areej Obeidat, Sarabjot Singh, and Tony Jan. Federated learning for intrusion detection in IoT environments: a privacy-preserving strategy. *Discover Internet of Things*, 5:72, 2025a. doi:10.1007/s43926-025-00169-7.

Ansam Khraisat, Ammar Alazab, Dev Vasan, Andrew Marshall, and Samar M. Alqhtani. RF-FedAvg: federated learning-based random forest model for intrusion detection in wireless sensor networks. *Cluster Computing*, 28, 2025b. doi:10.1007/s10586-025-05591-8.

Nickolaos Koroniotis, Nour Moustafa, Francesco Schiliro, Praveen Gauravaram, and Helge Janicke. Edge-federated learning-based

intelligent intrusion detection system for heterogeneous internet of things. *IEEE Access*, 12:83375–83393, 2024. doi:10.1109/ACCESS.2024.3410046.

Li Ma, Jicheng He, Kai Lu, Dan Wang, Long Yin, and Zhaokun Li. A contrastive learning and knowledge distillation-based framework for efficient federated intrusion detection in IoT. *Systems Science & Control Engineering*, 13(1):2518963, 2025. doi:10.1080/21642583.2025.2518963.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *Proceedings of Machine Learning Research*, 54:1273–1282, 2017.

Mohammad Y. Mhawish, Mahmoud Aljamal, Ayoub Alsarhan, Bashar S. Khassawneh, Adnan Akhunzada, and Ahmad Sami Al-Shamayleh. A novel adaptive generalization-guided federated learning for multimodal industrial IoT intrusion detection under statistical heterogeneity. *IEEE Open Journal of Computer Society*, 2026. doi:10.1109/OJCS.2026.3703977.

Viraaji Mothukuri, Prachi Khare, Reza M. Parizi, Seyedamin Pouriyeh, Ali Dehghantanha, and Gautam Srivastava. Federated-learning-based anomaly detection for IoT security attacks. *IEEE Internet of Things Journal*, 9(4):2545–2554, 2022. doi:10.1109/JIOT.2021.3077803.

Hope Leticia Nakayiza, Love Allen Chijioke Ahakonye, Dong-Seong Kim, and Jae Min Lee. PureFL: trust-weighted federated learning with noise-resilient homomorphic encryption for blockchain-based IoV networks. *IEEE Internet of Things Journal*, 2026. doi:10.1109/JIOT.2026.3691349.

Dinh C Nguyen, Ming Ding, Pubudu N Pathirana, Aruna Seneviratne, Jun Li, and H Vincent Poor. Federated learning for Internet of Things: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 23(3):1622–1658, 2021. doi:10.1109/COMST.2021.3075439.

Dinesh Kumar Nishad, Rashmi Singh, and Saifullah Khalid. A lightweight blockchain-enabled federated intrusion prevention framework for resource-constrained industrial IoT devices to detect and mitigate emerging cyberattacks. *Journal of Cloud Computing*, 2026. doi:10.1186/s13677-026-00843-3.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi:10.1136/bmj.n71.

Haonan Peng, Chunming Wu, and Yanfeng Xiao. FD-IDS: federated learning with knowledge distillation for intrusion detection in non-IID IoT environments. *Sensors*, 25(14):4309, 2025. doi:10.3390/s25144309.

S. Phani Praveen, Kanhaiya Sharma, Deepak Parashar, V. S. N. Murthy, Uddagiri Sirisha, and Deshinta Arrova Dewi. Design of an iterative method for adaptive federated intrusion detection for energy-constrained edge-centric 6G IoT cyber-physical systems. *Scientific Reports*, 15, 2025. doi:10.1038/s41598-025-25293-w.

A. Puviarasu and V.K. Sudha. Enhanced IoT security: privacy-preserving federated learning model for accurate, real-time intrusion detection across devices. *Ain Shams Engineering Journal*, 17:103866, 2026. doi:10.1016/j.asej.2025.103866.

Mahmoud Ragab, Ehab Bahaudien Ashary, Bandar M. Alghamdi, Rania Aboalela, Naif Alsaadi, Louai A. Maghrabi, and Khalid H. Allehaibi. Advanced artificial intelligence with federated learning framework for privacy-preserving cyberthreat detection in IoT-assisted sustainable smart cities. *Scientific Reports*, 15, 2025. doi:10.1038/s41598-025-88843-2.

Ripal Ranpara, Shobhit K. Patel, Om Prakash Kumar, and Fahad Ahmed Al-Zahrani. Scalable architecture for autonomous malware detection and defense in software-defined networks using

federated learning approaches. *Scientific Reports*, 15, 2025. doi:10.1038/s41598-025-14512-z.

Pedro Ruzafa-Alcázar, Pablo Fernández-Saura, Enrique Mármol-Campos, Aurora González-Vidal, José L. Hernández-Ramos, Jorge Bernal-Bernabe, and Antonio F. Skarmeta. Intrusion detection based on privacy-preserving federated learning for the industrial IoT. *IEEE Transactions on Industrial Informatics*, 19(2):1145–1154, 2023. doi:10.1109/TII.2021.3126728.

Xabier Sáez-de Cámara, Jose Luis Flores, Cristóbal Arellano, Aitor Urbieta, and Urko Zurutuza. Clustered federated learning architecture for network anomaly detection in large scale heterogeneous IoT networks. *Computers & Security*, 131:103299, 2023. doi:10.1016/j.cose.2023.103299.

Yousef Sanjalawe, Salam Al-E'mari, Tahani Alqurashi, Zahyah H. Alharbi, Sharif Naser Makhadmeh, and Mohammad Alsharaiah. Adaptive graph attention-based federated learning for IoT intrusion detection: mitigating poisoning attacks. *PeerJ Computer Science*, 11:e3281, 2025. doi:10.7717/peerj-cs.3281.

Sanaa A. Sharaf and Sameer Nooh. Identifying significant features in adversarial attack detection framework using federated learning empowered medical IoT network security. *Scientific Reports*, 15, 2025. doi:10.1038/s41598-025-14913-0.

Imtiaz Ali Soomro, Hamood ur Rehman Khan, Syed Jawad Hussain, Zeeshan Ashraf, Mrim M. Alnfiai, and Nouf Nawar Alotaibi. Lightweight privacy-preserving federated deep intrusion detection for industrial cyber-physical system. *Journal of Communications and Networks*, 26(6):632–649, 2024. doi:10.23919/JCN.2024.000054.

Shaymaa E. Sorour, Mohammed Aljaafari, Amany M. Shaker, and Ahmed E. Amin. LSTM-JSO framework for privacy preserving adaptive intrusion detection in federated IoT networks. *Scientific Reports*, 2025. doi:10.1038/s41598-025-95966-z.

T. Swarnalatha, S.L. Aruna Rao, Rahul Suryodai, Desidi Narsimha Reddy, A. Vanathi, and K. Sudheer Kumar. CyberFedEdgeAI framework for real time cyberattack detection in heterogeneous IoT systems using federated edge intelligence. *Discover Computing*, 2026. doi:10.1007/s10791-026-10129-6.

Hong-Yen Tran, Jiankun Hu, Xuefei Yin, and Hemanshu R. Pota. An efficient privacy-enhancing cross-silo federated learning and applications for false data injection attack detection in smart grids. *IEEE Transactions on Information Forensics and Security*, 18:2538–2552, 2023. doi:10.1109/TIFS.2023.3267892.

Saeed Ullah, Junsheng Wu, Mian Muhammad Kamal, Mohammed K. Alzaylaee, and Mohammad Alibakhshikenari. Spectre-Fed: evolving federated edge intelligence from FedEdge-ID to robust-private IoT intrusion detection via hybrid adversarial training. *IEEE Open Journal of Communications Society*, 2026. doi:10.1109/OJCOMS.2026.3665325.

Priyanka Verma, Nitesh Bharot, John G. Breslin, Donna O'Shea, Ankit Vidyarthi, and Deepak Gupta. Zero-Day Guardian: a dual model enabled federated learning framework for handling zero-day attacks in 5G-enabled IIoT. *IEEE Transactions on Consumer Electronics*, 70(1):3856–3866, 2024. doi:10.1109/TCE.2023.3335385.

APPENDIX A. CHARACTERISTICS OF INCLUDED STUDIES

Table 8. Characteristics of the 61 included studies. Quality (Q): S = Strong; A = Adequate; W = Weak.

ID	Author (Year)	Journal	Domain	FL Aggregation	Primary Dataset(s)	Q
S0	Chen et al.	IEEE Access	Wireless edge	FedAGRU	CICIDS2017	A

ID	Author (Year)	Journal	Domain	FL Aggregation	Primary Dataset(s)	Q
1	(2020)				, WSN-DS	
S02	Attota et al. (2021)	IEEE Access	General IoT	FedAvg	MQTT dataset	W
S03	Huong et al. (2022)	IEEE Access	ICS/SCADA	FedAvg (uniform)	SWaT, SCADA	S
S04	Driss et al. (2022)	Complex Intell. Syst.	Vehicular	FedAvg	Car Hacking 2020	A
S05	Ruzafa-Alcázar et al. (2023)	IEEE T-ITS	IIoT	FedAvg / Fed+	CIC-ToN-IoT	A
S06	Jithish et al. (2023)	IEEE Access	Smart grid	FedAvg	Ausgrid, CIDDS-001	S
S07	Tran et al. (2023)	IEEE T-IFS	Smart grid	DCR+Shamir	SimBench	S
S08	Hajj et al. (2023)	Sensors	General IoT	Statistic-sharing	NSL-KDD	A
S09	Sáez-de Cámara et al. (2023)	Comput. Secur.	General IoT	Clustered FedAvg	Gotham testbed	S
S10	Jin et al. (2024)	JCN	IIoT	FedAcc	SKAB	A
S11	Bhavsar et al. (2024)	IEEE Access	Vehicular	FedAvg	Car Hacking 2020	A
S12	Soomro et al. (2024)	JCN	IIoT	FedAvg+SMPC	X-IIoTID	S
S13	Verma et al. (2024)	IEEE TCE	IIoT/5G	FedAvg	X-IIoTID	S
S14	Agrawal et al. (2024)	IEEE Access	General IoT	Hierarchical FedAvg	CICIoT2023	S
S15	Koroniotis et al. (2024)	IEEE Access	Heterogeneous IoT	FedAvg+AdaBoost	BoT-IoT, UNSW-NB15	S
S16	Ahmad et al. (2025)	Sensors	Fog IoT	FedAvgM	ToN-IoT	A
S17	Anjum et al. (2025)	Sci. Rep.	General IoT	FedAvg	CICIoT2023	S

ID	Author (Year)	Journal	Domain	FL Aggregation	Primary Dataset(s)	Q
S18	Bukhari et al. (2024)	Internet of Things	Edge IIoT	Async FedAvg	Edge-IIoTset	A
S19	Chaurasia et al. (2024)	J. Supercomput.	IIoT	FedAvg	X-IIoTID	S
S20	Mothukuri et al. (2022)	IEEE IoTJ	General IoT	Async FedAvg	Modbus dataset	A
S21	Bouzeraib et al. (2025)	IEEE Access	General IoT	FedAvg+WGAN-GP	CSE-CIC-IDS2018	A
S22	Abd Elaziz et al. (2025)	Front. Big Data	General IoT	FedAvg+EEFO	N-BaIoT, CICIoT2023	S
S23	Almansour et al. (2025)	Sci. Rep.	Vehicular IoT	APFed	UNSW-NB15, CICIDS2017	S
S24	Peng et al. (2025)	Sensors	General IoT	FedProx+KD	UNSW-NB15, BoT-IoT	S
S25	Al Mazroa (2025)	Sci. Rep.	IIoT	FedAvg+GNN	UNSW-NB15	S
S26	Alsaleh et al. (2025)	Sensors	General IoT	Semi-decentralised FedAvg	CICIoT2023, WUSTL-IIoT	S
S27	Beshah et al. (2025)	Sci. Rep.	General IoT	FedDWC	NSL-KDD, CICIDS2017	S
S28	Praveen et al. (2025)	Sci. Rep.	Edge/6G IoT	Adaptive FedAvg	NSL-KDD	A
S29	Sharaf and Nooh (2025)	Sci. Rep.	General IoT	FedAvg	IoT healthcare dataset	A
S30	Devi et al. (2025)	ISA	WSN/smart city	FedAvg	NSL-KDD	A
S31	Alabdulatif (2025)	CMC	IIoT/smart city	Trust-weighted FedAvg	WUSTL-IIoTCC-2021	S
S32	Farooq et al. (2025)	CMC	Smart city	FedAvg	NSL-KDD (schema)	A
S33	Khan and Svetinovic	IoTCPS	General IoT	FedAvg	CICIoT2023	A

ID	Author (Year)	Journal	Domain	FL Aggregation	Primary Dataset(s)	Q
	(2026)					
S34	Danquah et al. (2025)	ISA	General IoT	FedAvg	N-BaIoT	S
S35	Alqazzaz (2025a)	Sci. Rep.	Smart industrial	FedAvg+HE+DP	X-IIoTID	S
S36	Khraisat et al. (2025b)	Cluster Comput.	WSN	Weighted FedAvg	WSN-DS	A
S37	Al Tfaily et al. (2025)	Sci. Rep.	General IoT	Graph FedAvg	ToN-IoT, UNSW-NB15	S
S38	Ragab et al. (2025)	Sci. Rep.	Smart city	FedAvg	Unnamed benchmark	W
S39	Hossain et al. (2025)	Complex Intell. Syst.	General IoT	FedAvg+KD	N-BaIoT	S
S40	Alqazzaz (2025b)	Int. J. Comput. Intell. Syst.	Smart city	FedAvg+HE	Unnamed smart city	A
S41	Ranpara et al. (2025)	Sci. Rep.	SDN/IoT	FedAvg	NSF-UNR SDN	A
S42	Sanjalawe et al. (2025)	PeerJ CS	General IoT	FedAvg+GAT	CIC-ToN-IoT	S
S43	Ma et al. (2025)	SSCE	General IoT	FedAvg+KD	N-BaIoT	S
S44	Babar et al. (2026)	IEEE OJCOMS	6G healthcare IoT	Trust-weighted FedAvg	IoT-Health, ToN-IoT	A
S45	Nakayiza et al. (2026)	IEEE IoTJ	Vehicular (IoV)	PureFL	CICIoV2024, 5G-NIDD	S
S46	Mhawish et al. (2026)	IEEE OJCS	IIoT (multimodal)	GE2F	TON-IoT	S
S47	Issa et al. (2026)	IEEE TMLCN	MQTT IoT	FedAvg	MQTT-IoT-IDS	S
S48	Ullah et al. (2026)	IEEE OJCOMS	Edge IoT	FedAvg+adversarial	CICIoT2023	A
S49	Joseph et al. (2026)	Sci. Rep.	IIoT	FedAvg	CICIoT2024-DIAD	W

ID	Author (Year)	Journal	Domain	FL Aggregation	Primary Dataset(s)	Q
S50	Sorour et al. (2025)	Sci. Rep.	General IoT	FedAvg	TON-IoT telemetry	A
S51	AlHayan and Al-Muhtadi (2026)	Sci. Rep.	General IoT	FedAvg+LLM	NSL-KDD	A
S52	Islam et al. (2025)	Sensors	Fog IoT	Hierarchical FedAvg	CIC-IoT2023	S
S53	Bilal et al. (2026)	Sci. Rep.	General IoT	FedAvg/FedProx/FedNova	CIC-IoT2023, TII-SSRC-23	A
S54	García-Sáez et al. (2026)	Comput. Netw.	Heterogeneous IoT	SCAFFOLD+FedAdX	X-IIoTID, Edge-IIoTset	S
S55	Puviarasu and Sudha (2026)	Ain Shams Eng. J.	General IoT	FedAvg+DP+HE	Kaggle IoT IDS	A
S56	Iqbal et al. (2026)	Discov. AI	IoT edge	FedAvg+SPADIS	Synthetic temp. data	W
S57	Swarnalatha et al. (2026)	Discov. Comput.	Heterogeneous IoT	FedAvg (sync/async)	TON-IoT, N-BaIoT	S
S58	Nishad et al. (2026)	J. Cloud Comput.	IIoT	Trust-weighted FedAvg	ToN-IoT, UNSW-NB15	S
S59	Khraisat et al. (2025a)	Discov. IoT	Smart home IoT	FedAvg/FedAvgM	N-BaIoT	A
S60	Alharthi et al. (2025)	Appl. Sci.	IIoT	Hierarchical FedAvg	Edge-IIoTset	S
S61	Albanbay et al. (2025)	JSAN	General IoT			