Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| REVIEW ARTICLE

Continual Federated Learning for Network Intrusion Detection: An Open-Index Systematic Map

Daryoosh Mansoory[1], Ahmad Wali Noori[2]✉, and Mohammad Peeroz Qaderi[3]

[1]MSc – Master of Science in Computer Science, Department of Information Technology, Faculty of Computer Science, Herat University, Herat, Afghanistan

[2]MSc – Master of Science in Computer Engineering - Software, Faculty of Education, Farah University, Farah, Afghanistan

[3]MSc – Master of Science in Computer Science - Computer Networks, Faculty of Education, Farah University, Farah, Afghanistan

Corresponding Author: Ahmad Wali Noori, **E-mail:** ahmadwali.noori1@gmail.com

| ABSTRACT

Purpose: Continual federated learning (FCL) combines decentralized model updating with sequential adaptation, but network-intrusion studies use incompatible definitions of drift, retention, and evaluation. This open-index systematic map examines how such systems operationalize non-stationarity, preserve earlier knowledge, and report empirical and reproducibility evidence. Method: Eight queries were executed in OpenAlex, Crossref, and DBLP on 26 July 2026. The interfaces returned 8,819 records within documented result limits; deduplication left 3,075 unique records. The initial concept-block rule excluded 2,865 records. A rule-based sensitivity audit reclassified two false negatives, yielding 2,863 automatic exclusions and 212 records for title/abstract screening. Thirty-one database-derived and nine supplementary reports were sought; 24 were not retrieved. Sixteen full texts were assessed, two were excluded, and 14 primary studies were mapped. Results: The corpus comprised three studies from 2023, six from 2024, one from 2025, and four from the partial 2026 search year. Nine studies tested retention after later tasks or classes; five reported adaptation without a prior-task retention outcome. Within the nine retention-tested studies, replay-based designs most often reported favorable within-study stability results, but heterogeneous scenarios and comparators do not establish cross-study superiority. Only two studies reported standardized forgetting or backward transfer, two studies supplied study-specific code, and one tested malicious clients. Conclusion: The accessible open-index evidence supports FCL as a prototype strategy under constructed benchmark sequences. It does not establish cross-network robustness, formal privacy, or deployment-scale reliability. A targeted supplementary update on 27 July 2026 found no additional eligible, accessible full text. Interpretation is constrained by capped interfaces, single-reviewer screening, and 24 unretrieved reports.

| KEYWORDS

continual federated learning; network intrusion detection; concept drift; catastrophic forgetting; systematic mapping

ARTICLE INFORMATION

ACCEPTED: 15 June 2026 **PUBLISHED:** 15 August 2026 **DOI:** 10.32996/jcsts.2026.8.8.27

1. Introduction

Models used in network intrusion detection systems (NIDS) are often evaluated under a closed-world assumption: the classes available during training are the classes encountered at deployment, samples are drawn from a fixed distribution, and a model can be retrained whenever its accuracy declines. Operational networks rarely satisfy that assumption. Attack families appear unevenly across sites, device populations change, benign traffic shifts with software and policy, and labels arrive late or not at all. A detector that adapts to the newest observations may consequently lose sensitivity to earlier attacks; a detector that does not adapt becomes stale (De Lange et al., 2022; Lavaur et al., 2022).

Federated learning (FL) addresses a different constraint. It coordinates model training across clients while keeping raw data decentralized, but it does not by itself solve temporal non-stationarity or forgetting (Kairouz et al., 2021). Continual learning (CL) addresses sequential acquisition and the stability–plasticity trade-off, but it is usually studied with data available at a single learner (De Lange et al., 2022). Continual federated learning (FCL) combines these settings: local streams evolve over time, clients possess different experiences, and the system must fuse new knowledge without erasing old knowledge (Yang et al.,

2024). In intrusion detection this combination is attractive because traffic is privacy-sensitive and new threats are distributed; it is also difficult because client drift and temporal drift interact.

Broad surveys of federated intrusion detection describe architectures, aggregation algorithms, security threats, and privacy mechanisms (Khraisat et al., 2024; Lavaur et al., 2022; Zhang et al., 2025). General FCL surveys classify knowledge-fusion mechanisms and non-stationary learning scenarios (Hamedi et al., 2025; Yang et al., 2024). Neither line of review, however, establishes how primary network-intrusion studies operationalize sequential change, whether they measure retention rather than only current-task accuracy, or whether their experiments can be reproduced. The terminology also causes false positives: some papers call repeated federated rounds "continual" or "incremental" even though the underlying data distribution is static.

This study maps primary experiments at the intersection of federated collaboration, sequential network data, and adaptive intrusion detection. Its contribution is not a pooled estimate of classifier performance. Instead, it separates retention-tested FCL from adaptation-only online FL, maps the experimental design space, audits methodological quality and reproducibility, and derives an evidence agenda for evaluation. That distinction matters: successful updating on a new batch is not evidence that a model retained earlier attack knowledge.

2. Literature Review and Review Gap

A conventional horizontal FL round distributes a global model, performs local optimization on client data, aggregates model updates, and repeats the process (McMahan et al., 2017). In a stationary experiment, the composition of each client's data may be non-independent or non-identically distributed (non-IID) yet fixed. FCL adds a temporal axis: each client receives a sequence of tasks, classes, domains, or unconstrained observations. Spatial heterogeneity determines what different clients know at a given time; temporal heterogeneity determines how each client's knowledge changes. Their interaction produces local forgetting, cross-client interference, and asynchronous knowledge availability (Yang et al., 2024).

For this map, a study qualified as direct FCL evidence only when it combined three observable elements: (i) collaborative federated or decentralized model learning; (ii) network-traffic intrusion or anomaly detection; and (iii) an explicit sequential, evolving, incremental, streaming, or non-stationary evaluation with adaptation or retention. Papers that merely trained for multiple communication rounds were excluded. Included studies were then divided into retention-tested evidence, where performance on earlier knowledge was evaluated after learning later data, and adaptation-only evidence, where online updating was evaluated without a direct forgetting outcome.

This operational rule is narrower than the labels used by authors. It recognizes that online learning may be valuable even when forgetting is not measured, but it prevents those experiments from being interpreted as proof of continual retention. The same rule permits diverse FCL scenarios—class increments, task sequences, explicit concept drift, and unlabeled streams—while preserving an evidence boundary between them.

Table 1. Positioning of this map against adjacent reviews

Review	Primary scope	Treatment of sequential FCL-NIDS	Gap retained
Yang et al. (2024)	General FCL knowledge fusion	General taxonomy; not NIDS-specific	No mapping of NIDS protocols or reproducibility
Hamedi et al. (2025)	FCL concepts, heterogeneity, privacy	Application-level discussion	No direct FCL-NIDS evidence audit
Khraisat et al. (2024)	Federated IDS architectures and aggregation	Sequential learning not separated from static FL	No operational FCL definition
Zhang et al. (2025)	Federated learning in intrusion detection	Broad FL-IDS synthesis	No retention-tested versus adaptation-only distinction
This study	Primary quantitative FCL-NIDS experiments	Strict sequential criterion and two evidence layers	Maps protocols, outcomes, quality, and reproducibility

3. Methodology

The review followed the mapping process described by Petersen et al. (2015) and used PRISMA 2020 for transparent reporting of identification, screening, eligibility, and inclusion (Page et al., 2021). A systematic map was selected because the evidence was heterogeneous in learning scenario, topology, dataset, model, and outcome definition. No prospectively

registered protocol was available; the operational criteria, coding rules, sensitivity correction, and departures from duplicate review are therefore reported within the manuscript rather than attributed to a registration record.

3.1 Research Questions

Nine research questions structured extraction and synthesis. RQ1 asked where and when FCL-NIDS research has been published. RQ2 mapped application domains and federation topologies. RQ3 classified the form of non-stationarity. RQ4 identified mechanisms used for adaptation and retention. RQ5 examined datasets, attack coverage, client partitions, and drift construction. RQ6 mapped models, baselines, and metrics. RQ7 synthesized quantitative evidence without pooling incompatible outcomes. RQ8 assessed methodological quality and reproducibility. RQ9 identified evidence gaps and evaluation priorities.

Table 2. Research questions and corresponding extraction fields

RQ	Question focus	Core extraction fields
RQ1	Publication profile	Year, venue, article type
RQ2	Domains and topologies	Application domain, federation topology, clients
RQ3	Non-stationarity	Class/task/domain increment, stream, drift timing
RQ4	Adaptation and retention	Replay, regularization, distillation, expansion, selective aggregation
RQ5	Data and partitions	Dataset, attacks, chronology, client split, IID/non-IID
RQ6	Evaluation design	Model, aggregation, baselines, metrics, repeats
RQ7	Quantitative evidence	Direction, unit, comparator, boundary conditions
RQ8	Quality and reproducibility	Seven quality domains, code/data/configuration
RQ9	Evidence gaps	Missing protocols, metrics, systems and security tests

3.2 Information Sources and Search Strategy

Searches were executed on 26 July 2026 in OpenAlex, Crossref, and DBLP. These open scholarly indices were chosen because subscription databases and publisher full-text services were not available in the review environment. Eight term-combination queries paired federated learning with continual, incremental, continuous, lifelong, concept-drift, online, or adaptive intrusion-detection terms. The complete strings are reproduced in Appendix A. OpenAlex searched its general work index, Crossref used the bibliographic query field, and DBLP used its publication search endpoint. Search exports and deduplication used DOI as the primary key and normalized title plus publication year as the fallback key.

The application programming interfaces (APIs) returned 8,819 query records: 794 from OpenAlex, 8,000 from Crossref, and 25 from DBLP. OpenAlex was capped at 200 results for each of the three broadest queries, and Crossref was capped at 1,000 results per query. The counts therefore describe retained API records, not every matching record reported by an index. Because publications recurred across queries and sources, 5,744 duplicate query records were removed, leaving 3,075 unique records. This open-index protocol is reproducible from the recorded endpoints and strings, but it is not equivalent to searches of IEEE Xplore, Scopus, Web of Science, or the ACM Digital Library and may under-represent poorly indexed conference reports.

A supplementary search checked reference lists, publisher pages, and targeted title/DOI web results for records that were not captured as retrievable reports in the main exports. Nine additional reports were logged in a separate PRISMA stream. None of those nine full texts could be retrieved, so they were not used in data extraction or synthesis.

The supplementary search was updated on 27 July 2026 with two focused queries in a peer-reviewed-paper discovery index and exact-title searches of publisher domains. Seven high-salience records were checked. Six matched reports already present in the database or supplementary retrieval registers: FedContinualIDS, EIDS, FedIDFR, TSK-FedIDS, TA-FCL, and Falcon. The remaining record evaluated FedAvg, FedProx, and FedDGC under increasing client heterogeneity but did not report a sequential task order, later-task adaptation, or prior-task retention; it was excluded at title/abstract screening. The update therefore changed neither the number of reports sought nor the included-study count. Appendix A records the queries, identifiers, and decisions.

3.3 Eligibility and Screening

Eligible reports were peer-reviewed primary studies published from 2018 to the search date, written in English, and reporting quantitative experiments. Full-text availability was treated as a retrieval outcome, not as an eligibility criterion. The lower bound was set at 2018, the first complete calendar year after the foundational FedAvg report (McMahan et al., 2017), to

delimit the post-FedAvg period. The intervention had to satisfy the three-part operational definition of FCL described above. Studies were excluded if they used static federated rounds without evolving data, addressed endpoint malware rather than network traffic, proposed a conceptual framework without quantitative evaluation, reported only an attack against an FL system, or duplicated an included study.

Screening was performed in a single-reviewer workflow with a second consistency pass over all candidate records. A deterministic concept-block rule retained records only when title or abstract contained a federated-learning term, a continual/adaptive term, and a network-security term. This rule excluded 2,865 records and forwarded 210 for title/abstract review. A post-screening, rule-based sensitivity audit of the complete automatic-exclusion log identified two direct false negatives: FLAME-IoT and FIDWATCH (Alatawi, 2026; Alrashdi et al., 2024). FLAME-IoT was retrievable and eligible; FIDWATCH remained unretrieved behind publisher access. Revised counts are therefore 2,863 automatic exclusions and 212 records in the title/abstract queue. The 181 manual exclusions were unchanged, leaving 31 database-derived reports for retrieval. Nine further reports came from supplementary searching, producing 40 reports sought in total. Sixteen full texts were retrieved and assessed; 24 reports were not retrieved and contributed no extracted evidence. Two of the 16 full texts were excluded: one conceptual framework lacked a primary quantitative experiment, and one arXiv preprint duplicated an included conference paper. Fourteen primary studies remained. The sensitivity audit corrected observed rule errors but was not an independent validation sample. No inter-rater agreement statistic is reported because independent duplicate screening did not occur.

Table 3. Eligibility criteria

Dimension	Included	Excluded
Population	Network-traffic intrusion or anomaly detection	Endpoint-only malware or non-network anomaly tasks
Collaboration	Federated or decentralized collaborative learning	Single-node learning without shared model knowledge
Temporal condition	Sequential/evolving/incremental/streaming data with adaptation or retention	Repeated FL rounds over a fixed data distribution
Evidence	Primary quantitative experiment	Conceptual proposal, editorial, review, protocol only
Language	English-language report	Non-English report
Retrieval handling	Eligibility judged after full-text retrieval	Unretrieved reports logged separately; not treated as ineligible
Date	2018–26 July 2026 (post-FedAvg period)	Outside the window
Publication identity	Peer-reviewed unique report	Duplicate preprint when a conference version was included

3.4 Data Extraction and Classification

A structured form captured bibliographic data, application domain, federation topology, number and type of clients, task or drift sequence, data-access assumptions, adaptation mechanism, privacy or robustness mechanism, model family, dataset, attack classes, partition strategy, baselines, metrics, repeated runs, resource measurements, headline results, limitations, and artifact availability. Publisher PDFs, publisher full-text pages, and complete public-access copies were used for study-level extraction. Extraction preserved the units and comparison direction reported by each study. When a result could not be verified in the retrieved report, it was recorded as not reported rather than inferred from an abstract, a related paper, or common practice.

Evidence was classified in two layers. Retention-tested studies evaluated old-task or old-class performance after later learning, used a forgetting or transfer measure, or compared knowledge before and after adaptation. Adaptation-only studies processed sequential batches or windows but did not measure retention. Mechanisms were coded as replay/retained samples, regularization or distillation, model expansion/isolation, selective aggregation, or uncontrolled online updating. Topologies were coded as central-server, peer-to-peer, or blockchain-committee federation.

3.5 Descriptive Methodological and Reproducibility Appraisal

Seven reporting and design domains were rated 0 (absent), 1 (partial), or 2 (adequate): clarity of the sequential protocol; federation and partition detail; comparator adequacy; outcome and retention measurement; reproducibility artifacts; explicit

limitations or validity threats; and operational realism. Domain ratings were not summed or used to rank studies. This review-specific appraisal is not a validated risk-of-bias instrument, and no study was excluded by a rating. Reproducibility was audited separately for dataset access, study-specific code, environment description, hyperparameters, repeated runs or uncertainty, and an executable end-to-end package.

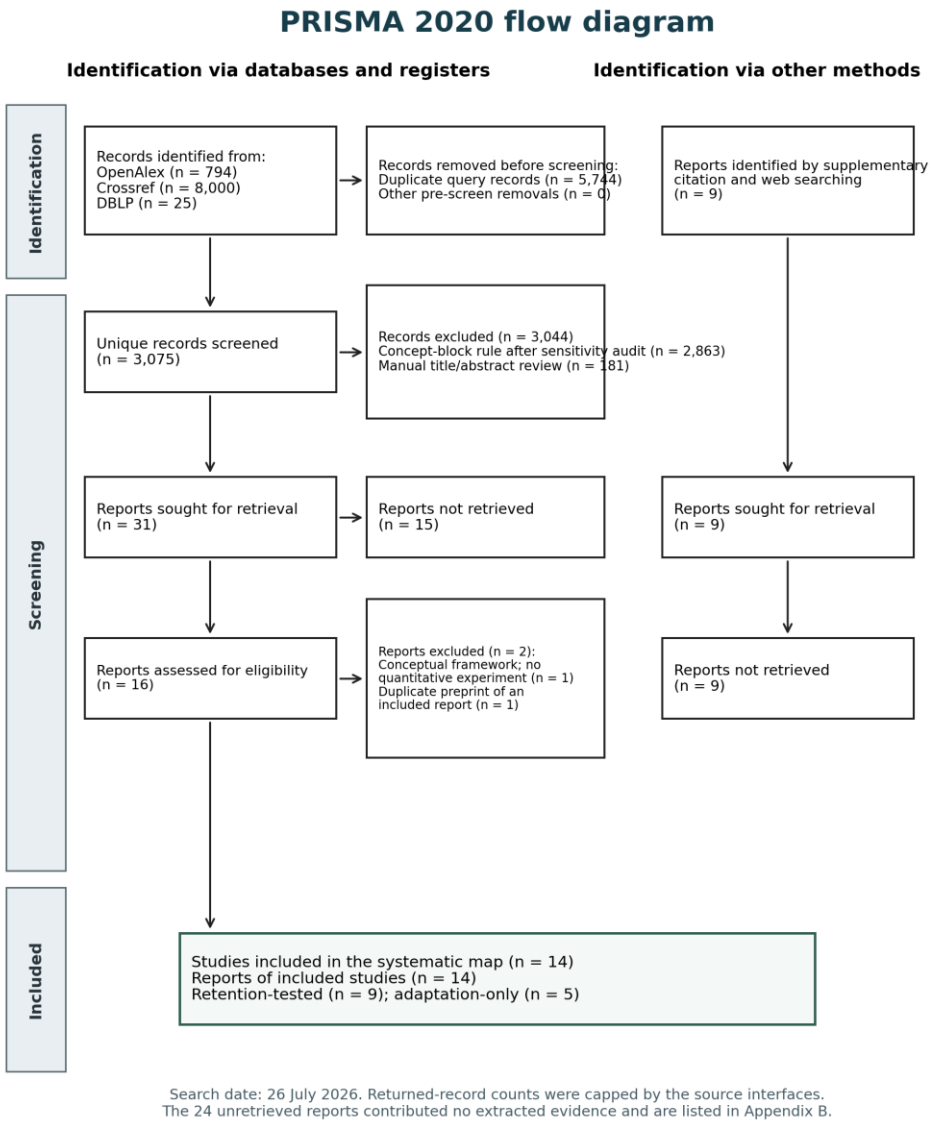

Figure 1. PRISMA 2020 study-selection flow.

3.6 Synthesis

The map uses counts, cross-tabulation, and structured narrative synthesis. A meta-analysis was not attempted. Accuracy on NSL-KDD, Macro-F1 under a four-phase Dirichlet partition, partial area under the curve (pAUC) at a 1% false-positive rate, and backward transfer over nine 5G tasks do not estimate a common effect. Pooling them would hide differences in label space, chronology, client heterogeneity, target access, and baseline strength.

4. Mapping Results

The included evidence comprises 14 studies published between 2023 and 26 July 2026. No eligible accessible full-text study published before 2023 was identified. Six studies appeared in 2024, the highest annual count; three appeared in 2023, one in

2025, and four in the incomplete 2026 observation period. The small and recent corpus supports mapping of design choices but not a claim that the field has converged on a stable evaluation protocol.

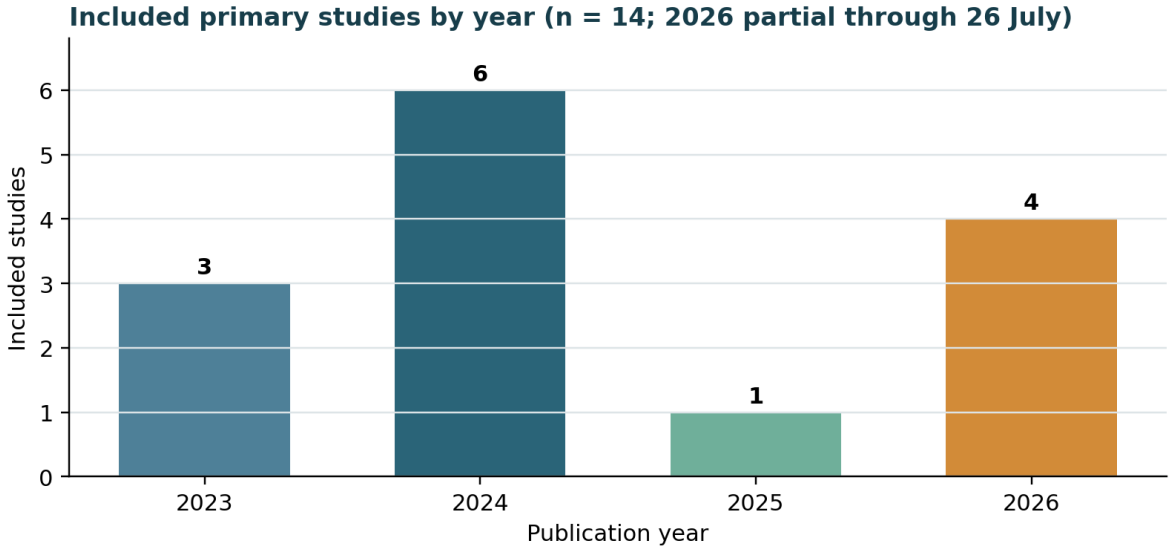

Figure 2. Publication timeline of included studies.

Table 4. Included studies and scope

ID	Study	Domain	Topology	Evidence layer	DOI
S06	Gelenbe et al. (2024)	Health IoT and Internet of Vehicles	Decentralized, asynchronous P2P	Adaptation-only	10.1016/j.iot.2024.101340
S07	Quyen et al. (2024)	IoT class-incremental NIDS	Central-server FL	Retention-tested	10.1109/ATC63255.2024.10908328
S08	Arvidsson et al. (2024)	General streaming NIDS	Central-server FL	Adaptation-only	10.1109/FMEC62297.2024.10710218
S14	He et al. (2023)	UAV edge-network NIDS	Asynchronous central-server FL with digital-twin assistance	Retention-tested	10.1109/JIOT.2023.3282648
S15	Mao et al. (2024)	IIoT class-incremental NIDS	Hierarchical cloud-edge-client FL	Retention-tested	10.1109/JIOT.2024.3408634
S16	Nakıp et al. (2023)	IoT networks	Decentralized P2P	Adaptation-only	10.1109/MASCOTS59514.2023.10387644
S19	Truong and Le (2023)	Metaverse collaborative IDS	Blockchain committee with online FL	Adaptation-only	10.1109/OJCS.2023.3312299
S22	Carillo et al. (2026)	IoT emerging-threat NIDS	Cross-silo central-server FL	Retention-tested	10.1109/TNSM.2026.3675031
S23	Benzaïd et al. (2024)	O-RAN/5G DDoS detection	Central server with two O-RAN trainers	Retention-tested	10.1109/WCNC57260.2024.10570951
S24	Soltani et al. (2024)	Multi-agent online NIDS	Federated multi-agent coordination	Retention-tested	10.1186/s42400-023-00199-0
S26	Wu et al. (2026)	Edge IoT	Central-server FCIL	Retention-tested	10.3390/s26144630
S28	Rehman et al. (2026)	IoMT	Flower central-server FL	Retention-tested	10.5220/0014475900004061

ID	Study	Domain	Topology	Evidence layer	DOI
S29	Su (2025)	Distributed cyber-attack detection	Central-server FCIL	Retention-tested	10.65286/icic.v21i3.80913
S30	Alatawi (2026)	Smart-city IIoT anomaly detection	Central-server adaptive FL	Adaptation-only	10.1007/s44196-026-01252-9

4.1 Domains and Federation Topologies (RQ1–RQ2)

Internet of Things (IoT)-related settings dominated. DISFIDA covered health IoT and vehicles; DOF-ID used IoT nodes; FI-IDS modeled a growing IoT federation; FCL-SBLS examined UAV edge networks; HFIN addressed industrial IoT (IIoT); FLAME-IoT examined adaptive anomaly detection in smart-city IIoT; three studies targeted emerging threats or class increments at the IoT edge; and one focused on IoMT. The remaining studies addressed O-RAN/5G DDoS detection, a metaverse collaboration architecture, general streaming NIDS, or distributed cyber-attack detection. Most experiments nevertheless replayed public benchmark traces rather than collecting simultaneous traffic at independent operational sites.

Eleven studies used central or hierarchical aggregation, two used peer-to-peer decentralized fusion, and one used a rotating blockchain committee. Client scale ranged from two O-RAN trainers to a simulated 100-node blockchain network. FI-IDS expanded from 30 to 50 clients, whereas HFIN modeled three edge servers serving 20 client devices each. Nominal node count did not always indicate training scale: MetaCIDS assigned only 70 of its 100 nodes to supervised or unsupervised training, while other nodes aggregated or detected.

Table 5. Architecture and application map

ID	Learning organization	Client scale	Operational setting
S06	Decentralized, asynchronous P2P	3 IoT nodes plus 6 vehicle nodes	Health IoT and Internet of Vehicles
S07	Central-server FL	30, then 40, then 50	IoT class-incremental NIDS
S08	Central-server FL	20, 30, and 40	General streaming NIDS
S14	Asynchronous central-server FL with digital-twin assistance	50 simulated UAVs	UAV edge-network NIDS
S15	Hierarchical cloud-edge-client FL	3 edge servers; 20 clients per edge	IIoT class-incremental NIDS
S16	Decentralized P2P	3 nodes	IoT networks
S19	Blockchain committee with online FL	100 nodes: 20 aggregators, 60 unsupervised trainers, 10 supervised trainers, 10 detectors	Metaverse collaborative IDS
S22	Cross-silo central-server FL	2 and 5	IoT emerging-threat NIDS
S23	Central server with two O-RAN trainers	2 base-station trainers	O-RAN/5G DDoS detection
S24	Federated multi-agent coordination	Multiple detector agents; scenario-specific	Multi-agent online NIDS
S26	Central-server FCIL	20; 60% participation	Edge IoT
S28	Flower central-server FL	5 IID clients	IoMT
S29	Central-server FCIL	3 IID clients	Distributed cyber-attack detection

ID	Learning organization	Client scale	Operational setting
S30	Central-server adaptive FL	More than 20 per dataset; scale tests at 20, 50, and 100; 18-device testbed	Smart-city IIoT anomaly detection

4.2 Non-Stationarity and Retention Mechanisms (RQ3–RQ4)

Six studies constructed class-incremental phases, one organised nine DDoS tasks, and two used an explicit timeline or injected feature drift. Five studies used online windows, new batches, or event-triggered adaptation. Nine studies evaluated earlier knowledge after later learning, but only two reported standardized forgetting or backward-transfer outcomes. Carillo et al. (2026) reported old/new/all Macro-F1 and gaps from joint retraining, which exposes retention and intransigence but is not identical to backward transfer. Rehman et al. (2026) compared performance along a fixed timeline and retained-sample strategies, again without a standardized forgetting statistic.

Replay or explicit sample retention appeared in seven retention-tested studies. TenaxDoS used reservoir replay; EdgeFedCIL combined exemplars with distillation; the IoMT study compared 100, 500, and 1,000 retained samples; the evolutionary method selected exemplars through clustering and genetic search; the FCIL IoT study used rehearsal with distillation and bias correction; FI-IDS retained 30% or 60% of earlier data while applying EWC; and HFIN allocated 500 exemplars per edge server. FCL-SBLS instead expanded a stacked broad model. The five adaptation-only studies updated models recursively, injected drift, or fused peer parameters without a controlled retention outcome.

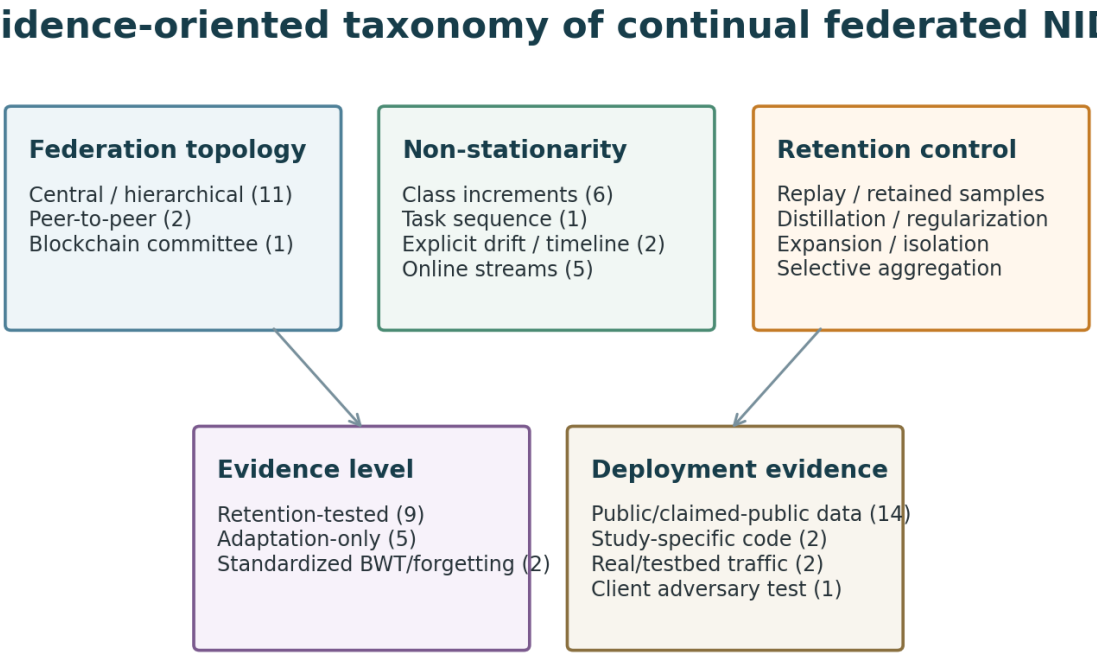

Figure 3. Evidence-oriented taxonomy of the included studies.

Table 6. Sequential scenarios and adaptation mechanisms

ID	Sequential scenario	Principal mechanism	Evidence layer
S06	Online windows; heterogeneous attack streams	Self-supervised DRNN; selective peer parameter fusion	Adaptation-only
S07	Three class increments; 30→40→50 clients; 60%/30% old-data budgets	CNN-GRU with elastic weight consolidation and partial old-data retention	Retention-tested

ID	Sequential scenario	Principal mechanism	Evidence layer
S08	Prequential stream; single-pass chunks	Online Gaussian Naive Bayes with SFLEDS	Adaptation-only
S14	Incremental class ratios with dynamic UAV participation	Stacked broad learning expansion and DDPG-based UAV selection	Retention-tested
S15	Two, five, or ten class-incremental tasks under bandwidth constraints	500-exemplar memory per edge; Replay/iCaRL/WA/DER with weighted transmission optimization	Retention-tested
S16	Time-windowed online updates	DRNN/SWBC with decentralized federated update	Adaptation-only
S19	New local batch each blockchain round	Attention classifier, autoencoder, Multi-Krum, differential privacy	Adaptation-only
S22	Disjoint client-specific attack-class increments	Rehearsal, distillation, bias correction; FedAvg/FedProx/FedDyn	Retention-tested
S23	Nine sequential DDoS tasks	Reservoir replay with FedAvgCL/FedProxCL	Retention-tested
S24	New attack exposure and online model expansion	Federated distillation, Fisher regularization, expansion and frozen units	Retention-tested
S26	Four class-incremental phases	Replay, distillation, class weighting, low-rank compression and quantization	Retention-tested
S28	Timeline t0–t6 with evolving attack families	Cumulative, representative, retained-sample and averaging strategies	Retention-tested
S29	Three tasks, two new classes per task	Evolutionary exemplar replay, clustering and homomorphic encryption	Retention-tested
S30	Synthetic feature drift; non-IID partitions; 20/50/100-client scale tests	Utility-based client selection, momentum aggregation, and hybrid VAE-LSTM scoring	Adaptation-only

4.3 Datasets, Client Partitions, and Drift Construction (RQ5)

The corpus reused a narrow set of public intrusion benchmarks. NSL-KDD appeared twice, UNSW-NB15 twice, and the CIC family in several studies. More recent domain-specific datasets included 5G-NIDD, CICIoMT2024, ToN-IoT, X-IIoTID, IoT-NID, Edge-IIoTset, NF-UQ-NIDS-v2, NF-ToN-IoT-v2, SWaT, and a custom LoRaWAN IIoT collection. FI-IDS, DISFIDA, and DOF-ID used Kitsune, with the latter two also using Bot-IoT. Public availability improves data access, but the widespread use of pre-collected traces means that chronology was usually imposed after collection rather than observed prospectively.

Client partitioning varied substantially. Some studies assigned different attack traces to nodes; Carillo et al. (2026) gave each client a disjoint new attack; Wu et al. (2026) evaluated Dirichlet heterogeneity at α=.1, .5, and 1.0; Rehman et al. (2026) used five balanced IID clients; and Su (2025) assumed three IID clients. Alatawi (2026) used α=.1, .3, and .5 and injected feature shifts, but did not revisit prior tasks after adaptation. Several reports did not distinguish temporal drift from client non-IIDness. Consequently, a gain attributed to continual learning may partly reflect a favorable partition, while a loss may reflect federated optimization rather than forgetting.

Only Carillo et al. (2026) included a dedicated cross-dataset deployment test, and that evaluation showed substantial degradation, with binary pAUC reaching at most 75% in the reported cross-dataset setting. MetaCIDS trained and tested separately on four datasets; this is multi-dataset evaluation, not transfer between networks. The evidence therefore remains primarily within-dataset and does not establish that continual federation compensates for network-specific feature and label shifts.

Table 7. Datasets and client/drift construction

ID	Dataset(s)	Clients	Sequence or partition
S06	Kitsune; Bot-IoT; CAN-bus vehicle traces	3 IoT nodes plus 6 vehicle nodes	Online windows; heterogeneous attack streams
S07	Kitsune	30, then 40, then 50	Three class increments; 30→40→50 clients; 60%/30% old-data budgets
S08	NSL-KDD	20, 30, and 40	Prequential stream; single-pass chunks
S14	CIC-IDS2017	50 simulated UAVs	Incremental class ratios with dynamic UAV participation
S15	NF-UQ-NIDS-v2; NF-ToN-IoT-v2	3 edge servers; 20 clients per edge	Two, five, or ten class-incremental tasks under bandwidth constraints
S16	Kitsune; Bot-IoT	3 nodes	Time-windowed online updates
S19	CSE-CIC-IDS2018; CIC-IDS2017; NSL-KDD; UNSW-NB15	100 nodes: 20 aggregators, 60 unsupervised trainers, 10 supervised trainers, 10 detectors	New local batch each blockchain round
S22	TON_IoT; IoT-NID; Edge-IIoTset	2 and 5	Disjoint client-specific attack-class increments
S23	5G-NIDD	2 base-station trainers	Nine sequential DDoS tasks
S24	CIC-IDS2017; CSE-CIC-IDS2018	Multiple detector agents; scenario-specific	New attack exposure and online model expansion
S26	ToN-IoT; X-IIoTID	20; 60% participation	Four class-incremental phases
S28	CICIoMT2024	5 IID clients	Timeline t0–t6 with evolving attack families
S29	UNSW-NB15; NSL-KDD	3 IID clients	Three tasks, two new classes per task
S30	SWaT; TON_IoT; custom LoRaWAN IIoT	More than 20 per dataset; scale tests at 20, 50, and 100; 18-device testbed	Synthetic feature drift; non-IID partitions; 20/50/100-client scale tests

4.4 Models, Baselines, and Outcomes (RQ6–RQ7)

Model families included deep random neural networks, Gaussian Naive Bayes, convolutional and recurrent networks, stacked broad learning, multilayer perceptrons, autoencoders, attention modules, and Transformer-style classifiers. FedAvg was the most frequent aggregation baseline; FedProx and FedDyn appeared in recent FCIL studies. Comparator quality varied. Stronger designs separated local online learning, static federation, centralized continual learning, and joint retraining. Weaker designs compared only a proposed method with simple averaging or static training.

Accuracy was ubiquitous, but the treatment of class imbalance was uneven. Recent studies reported Macro-F1, pAUC at low false-positive rates, or old/new class scores. Only TenaxDoS and EdgeFedCIL reported standardized backward-transfer or forgetting measures. Systems outcomes were also selective: DISFIDA and DOF-ID reported communication or execution time; MetaCIDS measured blockchain throughput and latency; EdgeFedCIL quantified upload volume; and the IoMT study measured training and inference latency. Energy, peak memory, client dropout, and update failure were almost absent.

Across studies, retention controls generally outperformed static or simple-incremental baselines. TenaxDoS reported at least 13.1 percentage points of F1 improvement over independent continual learning and past-task accuracy above 98.8% (Benzaïd et al., 2024). On ToN-IoT with α=.1, EdgeFedCIL improved Macro-F1 by 9.01 points over GLFC while approximately halving the reported forgetting value (Wu et al., 2026). In IoMT binary classification, representative retention reached 95.73% average accuracy, whereas simple incremental and averaging variants deteriorated later in the stream (Rehman et al., 2026). These results support retention under their constructed sequences; they do not estimate a transferable effect across networks.

Three additional studies reinforced that pattern but also exposed its limits. In FI-IDS, the 60%-memory configuration reached 75.08% accuracy and 79.66% F1 at Task 3, compared with 71.21% and 77.22% for GLFC-60; the full-data FedIDS upper bound still achieved 78.34% accuracy (Quyen et al., 2024). FCL-SBLS ranked above FCL-BLS, FL-Replay, and FL-Regularization across the paper's four incremental scenarios, but the report did not provide a standardized forgetting statistic (He et al., 2023). HFIN showed that smoothed, class-aware transmission weights improved accuracy and Macro-F1 across two IIoT datasets; those experiments used a fixed memory of 500 exemplars per edge server and simulated three edge servers with 20 clients each (Mao et al., 2024).

Joint training remained a demanding upper bound. In Carillo et al. (2026), the best FCIL configuration outperformed centralized continual learning on several dataset/scenario combinations, yet F1all remained 5.21 to 17.30 points below Scratch in the tabulated experiments. This is informative rather than a failure: it quantifies the cost of decentralized incremental acquisition. It also cautions against conclusions based only on comparisons with a static model that has never seen the new classes.

The adaptation-only evidence is useful but narrower. DISFIDA reported 100% TPR and 99% TNR with communication overhead below 0.03% in its unencrypted experiment (Gelenbe et al., 2024). DOF-ID reported node-level metrics above .86 and a total per-window operation time of 48.91 ms (Nakıp et al., 2023). The online SFLEDS experiment reached about 90% accuracy on NSL-KDD and explicitly acknowledged that catastrophic forgetting was not addressed (Arvidsson et al., 2024). MetaCIDS achieved 96.01%, 96.56%, 82.25%, and 98.77% accuracy on four datasets and tested 33% poisoning trainers, but it did not test retention across earlier tasks (Truong & Le, 2023). FLAME-IoT reported F1/AUC values of .854/.92 for the full model, .741/.85 without momentum, and .783/.88 without hybrid scoring under synthetic drift (Alatawi, 2026). Its abstract, main text, and Table 5 give conflicting directions or magnitudes for communication change, so no comparative communication estimate was synthesized.

Table 8. Evaluation outcomes and bounded quantitative findings

ID	Metrics	Verified finding	Interpretive boundary
S06	Accuracy, TPR, TNR; time; communication	TPR 100%, TNR 99%; update traffic <0.03% without encryption; no formal forgetting metric	Online adaptation; prior-task retention not established
S07	Accuracy, precision, recall, F1	At Task 3, FI-IDS-60 reached 75.08% accuracy and 79.66% F1 versus 71.21% and 77.22% for GLFC-60; no standardized forgetting score	Retention directly tested
S08	Accuracy	Federated online setting reached approximately 90% accuracy; authors state forgetting was not addressed	Online adaptation; prior-task retention not established
S14	Accuracy, precision, recall, F1; added training time; modeled cost	FCL-SBLS had the highest reported accuracy across four incremental scenarios relative to FCL-BLS, FL-Replay, and FL-Regularization; no standardized forgetting score	Retention directly tested
S15	Accuracy, Macro-F1; communication and computation analysis	WTO with smoothing (β=.5) consistently exceeded the unsmoothed and quantity-only settings; the abstract reports approximately 10% gains over baselines, but no standardized forgetting score	Retention directly tested
S16	Accuracy, TPR, TNR; latency	All node-level metrics exceeded 0.86; ≥15-point accuracy advantage over reported comparators; 48.91 ms total operation time	Online adaptation; prior-task retention not established
S19	Accuracy, precision, recall; zero-day detection; convergence; blockchain throughput	96.01%/96.56%/82.25%/98.77% accuracy by dataset; 33% poisoners tested; retention not measured	Online adaptation; prior-task retention not established
S22	Macro-F1 old/new/all; pAUC at FPR ≤1%; drop from upper bounds	Best FCIL exceeded centralized CIL on several datasets but remained 5.21–17.30 F1 points below joint-training Scratch across reported scenarios	Retention directly tested

ID	Metrics	Verified finding	Interpretive boundary
S23	Accuracy, precision, recall, F1, BWT, FWT	FCL improved independent CL by ≥14.4 accuracy, 10.8 precision, 13.1 F1, and 19.2 recall points; past-task accuracy >98.8%	Retention directly tested
S24	Detection rate before/after adaptation; adaptation samples; time	CNN average detection rate >95% after 128 new flows; LSTM adapted after the first 15 packets; known-attack performance was retained in reported cases	Retention directly tested
S26	Accuracy, Macro-F1 old/new, forgetting, BWT, upload volume	At α=.1 on ToN-IoT, +9.01 Macro-F1 points and forgetting .1187→.0616 versus GLFC; cumulative upload reduced by up to 10.54×	Retention directly tested
S28	Accuracy, precision, recall, F1, FAR; training and inference latency	Representative strategy averaged 95.73% binary accuracy; cumulative strategy 66.7% six-class accuracy; retained samples reduced training cost	Retention directly tested
S29	Accuracy, precision, recall, F1	Task-3 accuracy exceeded random, GAN, and VAE replay by 4.66, 6.91, and 22.32 points; no standardized forgetting score	Retention directly tested
S30	F1, AUC-ROC, latency, communication, convergence, and energy	Full-model F1/AUC were .854/.92 versus .741/.85 without momentum and .783/.88 without hybrid scoring; communication claims conflict across the abstract, main text, and Table 5; prior-task retention was not evaluated	Online adaptation; prior-task retention not established

4.5 Methodological Reporting and Reproducibility (RQ8)

The domain profile in Table 9 is descriptive rather than a composite quality score. Sequential protocols and client partitions were usually described, whereas prior-task retention, study artifacts, uncertainty, and explicit validity analysis were uneven. These domain ratings do not establish low risk of bias: most experiments used a small number of constructed sequences, and baseline tuning was rarely symmetric or independently verified.

Thirteen studies used data that are publicly obtainable or described as public; FLAME-IoT mixed two named benchmarks with a custom LoRaWAN collection whose access path was not fully traceable. Study-specific executable code was located for Soltani et al. (2024) and Alatawi (2026). Carillo et al. (2026) cited the generic FACIL codebase they extended, but a study-specific implementation was not reported. Wu et al. (2026) stated that code would be made available; it was not available at the search cutoff. Five-run evaluation appeared in TenaxDoS, and Carillo et al. evaluated ten data splits, but confidence intervals and statistical tests were uncommon.

Reproducibility is more than source-code presence. A usable FCL benchmark needs the chronological task manifest, client assignment, preprocessing transforms, random seeds, model initialization, aggregation schedule, buffer policy, and metric implementation. None of the inspected papers supplied a complete machine-executable package covering all of those elements. As a result, headline accuracy can often be recomputed only after reconstructing consequential design choices.

Table 9. Descriptive reporting and design profile (0 = absent, 1 = partial, 2 = adequate)

ID	Sequential	Federation	Baselines	Retention outcome	Artifacts	Limitations	Realism
S06	1	2	2	1	1	1	2
S07	2	2	2	1	0	0	1
S08	2	2	2	1	0	2	1
S14	2	2	2	1	0	1	1
S15	2	2	2	1	0	1	1
S16	2	2	2	1	1	1	1

ID	Sequential	Federation	Baselines	Retention outcome	Artifacts	Limitations	Realism
S19	1	2	2	1	0	1	2
S22	2	2	2	2	1	2	2
S23	2	2	2	2	1	1	2
S24	2	2	2	2	2	2	1
S26	2	2	2	2	1	2	1
S28	2	2	2	2	0	2	1
S29	2	2	2	1	0	2	1
S30	2	2	2	1	2	1	2

Scoring domains: sequential protocol; federation and partition detail; comparator adequacy; retention-aware outcomes; reproducibility artifacts; explicit limitations; and operational realism. Ratings are displayed by domain and are not summed or interpreted as risk-of-bias scores.

Table 10. Reproducibility audit

ID	Public data	Study code	Environment	Hyperparameters	Repeats/uncertainty	End-to-end package
S06	Yes	No	Yes	Yes	No	No
S07	Yes	No	Yes	Yes	Not reported	No
S08	Yes	No	Partial	Partial	No	No
S14	Yes	No	Yes	Yes	Not reported	No
S15	Yes	No	Yes	Yes	Not reported	No
S16	Yes	No	Yes	Yes	No	No
S19	Yes	No	Yes	Yes	No	No
S22	Yes	No; base FACIL framework cited	Yes	Yes	10 splits; mean reported	Partial
S23	Yes	No	Yes	Yes	Five runs	No
S24	Yes	Yes	Partial	Yes	Not reported	Yes
S26	Yes	Promised, not available at cutoff	Yes	Yes	Not reported	No
S28	Yes	No	Yes	Yes	Not reported	No
S29	Yes	No	Partial	Yes	Not reported	No
S30	Partial	Yes	Yes	Yes	Not reported	Partial

4.6 Robustness of the Synthesis to Reporting Completeness

A post hoc robustness check removed four studies with partial reporting in at least five of the seven descriptive appraisal domains (S07, S14, S15, and S19). This left ten studies: six retention-tested and four adaptation-only. Five of the six retention-tested studies still used replay or retained samples, and the two standardized forgetting or backward-transfer reports remained in the subset. The principal interpretation therefore did not depend on the four least completely reported studies: retention controls were favorable within constructed sequences, but the evidence remained too heterogeneous for a pooled or cross-network superiority claim.

The check also sharpened a weakness. The only included malicious-client experiment was S19, one of the four less completely reported studies. After its removal, the higher-information subset contained no direct client-adversary test. Claims about poisoning resilience must therefore remain outside the central effectiveness conclusion.

5. Discussion

Four findings define the present evidence boundary. FCL-NIDS is an emerging rather than mature research area: the earliest accessible eligible studies appeared in 2023, and much of the corpus used terminology that did not map cleanly to a shared continual-learning protocol. Second, within the nine retention-tested studies, replay or retained samples most often accompanied favorable within-study stability results, although cross-study superiority cannot be inferred and local storage conflicts with strong privacy interpretations of federated learning. Third, gains are conditional on constructed within-dataset streams and frequently disappear relative to joint retraining. Fourth, reproducibility and operational evaluation lag behind algorithmic design.

5.1 What Counts as Continual Federated Detection?

The boundary between online FL and FCL should be empirical, not terminological. A model that receives a new batch each round is adaptive, but continual retention is demonstrated only when the evaluation revisits earlier knowledge or computes a retention measure. Five included studies satisfy the broader adaptation criterion without satisfying that stronger retention criterion. Reporting them separately avoids discarding useful streaming evidence while preventing a category error.

Future papers should state the learning scenario before presenting the algorithm. At minimum, authors should specify whether labels expand, domains shift, task identity is known, old data can be stored, clients observe synchronized or asynchronous changes, and the global model is evaluated on local, pooled, or external traffic. Without those statements, the same term—incremental federated learning—can describe incompatible problems.

5.2 Stability, Plasticity, and the Cost of Memory

Replay was effective across several experiments because it anchors the decision boundary with earlier samples. Yet in network security, keeping raw exemplars may retain payload-derived features, rare device signatures, or quasi-identifiers. None of the replay studies jointly measured forgetting, membership inference, and storage exposure. Distillation and regularization reduce raw-memory dependence but can preserve biased teacher errors or constrain adaptation to genuinely novel attacks. Model expansion protects old parameters at the cost of memory growth and more complex synchronization.

The evidence does not identify a universally best mechanism. EdgeFedCIL's gains were strongest under severe heterogeneity on ToN-IoT, while a comparator remained marginally better in some milder X-IIoTID settings (Wu et al., 2026). Carillo et al. (2026) found that additional synchronization helped some new-class outcomes but showed diminishing returns and dataset-specific reversals. These interactions argue for factorial experiments crossing buffer size, heterogeneity, task order, and communication budget.

5.3 Privacy, Robustness, and Operational Claims

Keeping raw traffic local is a data-movement property, not a complete privacy guarantee. Model updates can leak information, and replay changes the local attack surface. MetaCIDS tested poisoning convergence and used differential privacy and a blockchain committee; Su (2025) incorporated homomorphic encryption. Most other studies assumed honest clients and protected channels. No included study reported a joint privacy–retention frontier, and only one tested a malicious-client fraction.

Operational claims were similarly underdetermined. Millisecond update time on one workstation does not predict performance over a congested wide-area federation. Model-upload reductions do not establish energy savings unless computation and retransmission are measured. O-RAN integration and the 5G-NIDD testbed are valuable steps toward realism, but the training experiment still involved only two trainer virtual machines. Claims of scalability should therefore be bounded by client count, network model, hardware, availability pattern, and measured resource.

5.4 Implications for Practitioners

The current literature supports prototypes, not unattended deployment. A practitioner evaluating FCL-NIDS should begin with a chronological local baseline and a periodic joint-retraining upper bound. The FCL candidate should then be evaluated under client dropout, delayed labels, attack imbalance, and an external network shift. Retention should be measured separately for old and new attacks, and the alert operating point should be fixed before comparison. A method that raises Macro-F1 while increasing false alarms beyond the operational budget is not an improvement for a security operations center.

Memory policy deserves explicit governance. If exemplars are retained, their feature content, retention duration, encryption, access control, and deletion path should be documented. If only distilled outputs or prototypes are stored, authors should test

whether they leak client membership or attack-specific signatures. Federation should be treated as one layer in a security design, not as evidence that privacy has been solved.

6. Evidence Gaps and Research Agenda (RQ9)

The immediate need is a benchmark protocol rather than another isolated classifier. A useful benchmark would preserve acquisition time, split traffic by independently collected sites, introduce attacks and benign shifts at different clients, and publish a task manifest. It should support both class-incremental and domain-incremental evaluation, with an external network held out until the end. Such a design would separate forgetting from client drift and cross-network shift.

Minimum baselines should include local static training, local online learning, static FedAvg, the same continual method without federation, centralized continual learning, cumulative retraining, and joint training as an upper bound. The primary outcome set should include Macro-F1 or area under the precision–recall curve (AUCPR), false-positive rate at a fixed detection rate, old/new/all performance, average forgetting, backward and forward transfer, adaptation delay, cumulative communication, client computation, memory, and energy. Results should be reported over several task orders and random seeds with uncertainty.

Security evaluation must include poisoning, backdoors, inference, and compromised aggregators. Defenses should be evaluated for their effect on both plasticity and retention; robust aggregation that suppresses unusual updates may also suppress legitimate early evidence of a new attack. Finally, study packages should include code, containers, immutable dataset identifiers, preprocessing, client manifests, task sequences, configuration files, seeds, and scripts that regenerate every table.

Table 11. Prioritized evidence agenda

Priority	Current weakness	Minimum design response	Primary value
1	Constructed single-dataset streams	Chronological multi-site benchmark plus external-network holdout	Separates forgetting, client drift, and domain shift
2	Inconsistent sequential definitions	Publish task manifests and declare class/task/domain/online scenario	Makes studies comparable
3	Accuracy-dominant evaluation	Macro-F1/AUCPR, fixed operating point, old/new/all, forgetting, BWT/FWT	Measures stability and security utility
4	Weak or incomplete baselines	Local/static FL, local CL, centralized CL, cumulative and joint-training upper bounds	Attributes gains to the correct component
5	Limited uncertainty	Multiple task orders and seeds with intervals and paired tests	Tests robustness of conclusions
6	Sparse systems evidence	Bytes, latency, energy, memory, dropout and failure recovery	Supports deployment claims
7	Privacy assumed from locality	Inference, poisoning, backdoor and compromised-aggregator tests	Evaluates the actual threat model
8	Few reusable artifacts	Container, split/task registry, configuration, seeds and table-generation scripts	Enables independent verification

7. Threats to Validity and Limitations

The largest limitation is retrieval coverage. The review searched three open scholarly indices rather than subscription databases, and 24 of 40 reports sought after screening or supplementary discovery could not be retrieved in full text. Those reports were excluded from evidence extraction, even when their abstracts appeared relevant. The resulting map is therefore an open-index, accessible-full-text evidence layer, not a census of every FCL-NIDS paper. Publisher access, indexing delay, and language restriction may have produced selection bias.

Screening and extraction were not independently duplicated. A repeated consistency pass reduced clerical error but cannot replace two reviewers, and no agreement statistic is available. The rule-based sensitivity audit corrected two observed false negatives but does not estimate the overall false-negative rate. The search vocabulary may also miss studies that implement

continual behavior without using continual, incremental, online, adaptive, or drift terminology. Conversely, broad adaptive queries produced many false positives.

Construct validity remains a concern because authors use "continual," "online," and "incremental" inconsistently. The review addressed this by applying a three-part operational definition and by separating retention-tested from adaptation-only evidence. Even so, classification required judgment when a study processed sequential batches but did not expose performance on earlier tasks.

Quantitative conclusions are limited by study design. Datasets, attack labels, preprocessing, task order, client partitions, metrics, baselines, and hardware were incompatible, so no pooled effect was calculated. The appraisal domains are review-specific and descriptive. They summarize reporting and design features; they are not estimates of causal bias, certainty ratings, or guarantees of external validity.

The main index search includes studies published through 26 July 2026, with a targeted supplementary update on 27 July 2026. The 2026 count therefore represents a partial year and must not be compared directly with complete annual counts. One included article was available as an unedited accepted manuscript close to the cutoff. Fast-moving publication and later artifact release mean that the reproducibility assessment can change. An update should repeat searches in IEEE Xplore, Scopus, Web of Science, ACM Digital Library, and publisher platforms with institutional access.

The 2018 lower bound was chosen to cover the post-FedAvg period, but this operational boundary may exclude earlier decentralized online-learning studies that later authors would describe as federated or continual. Restricting inclusion to English-language reports may also under-represent work published in other languages.

8. Conclusion

Fourteen accessible primary studies show that continual federated learning is being used to update intrusion detectors across evolving and decentralized traffic. Nine studies tested retention; five demonstrated only online adaptation. Within the retention-tested subset, replay or retained samples most often accompanied favorable stability results, but the evidence does not establish their superiority across studies. Recent methods combined replay with distillation, bias correction, compression, or model expansion. The evidence is promising within constructed benchmark streams, but it does not yet establish robust transfer across networks, end-to-end privacy, or large-scale deployment. These conclusions apply to the accessible open-index corpus and are constrained by the unretrieved-report rate and single-reviewer process.

The decisive research question is no longer whether a federated detector can be updated. It is whether the detector can retain earlier attacks, absorb genuinely new behavior at heterogeneous clients, respect an alert budget, survive adversarial participation, and do so within measurable communication, memory, and energy constraints. Common chronological benchmarks and executable study packages would address the most immediate comparability and reproducibility gaps.

Declarations

Funding: This research received no external funding. The APC will be funded by the authors.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: The bibliographic metadata, search logs, deduplicated records, screening decisions, sensitivity audit, retrieval register, and processing scripts are provided in the supplementary reproducibility package accompanying this submission. The package contains no copyrighted full texts.

Ethics Approval and Consent to Participate: Not applicable. This study analyzed published bibliographic records and reports and did not involve human participants, personal data, or animal subjects.

Author Contributions: All authors reviewed and approved the submitted manuscript and accept responsibility for the integrity of the work.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1]. Alatawi, M. N. (2026). FLAME-IoT: Federated learning with adaptive model evolution for smart city industrial IoT anomaly detection. *International Journal of Computational Intelligence Systems*. https://doi.org/10.1007/s44196-026-01252-9

[2]. Alrashdi, I., Sallam, K. M., Alrowaily, M. A., Alruwaili, O., & Arain, B. (2024). FIDWATCH: Federated incremental distillation for continuous monitoring of IoT security threats. *Ad Hoc Networks, 165*, 103637. https://doi.org/10.1016/j.adhoc.2024.103637

[3]. Arvidsson, V., Alawadi, S., Boldt, M., Angelsmark, O., & Söderlund, F. (2024). A novel approach for intrusion detection using online federated learning on streaming data. In *2024 9th International Conference on Fog and Mobile Edge Computing (FMEC)* (pp. 114–121). IEEE. https://doi.org/10.1109/FMEC62297.2024.10710218

[4]. Benzaïd, C., Hossain, F. M., Taleb, T., Merino Gómez, P., & Dieudonne, M. (2024). A federated continual learning framework for sustainable network anomaly detection in O-RAN. In *2024 IEEE Wireless Communications and Networking Conference (WCNC)* (pp. 1–6). IEEE. https://doi.org/10.1109/WCNC57260.2024.10570951

[5]. Carillo, R., Cerasuolo, F., Bovenzi, G., Ciuonzo, D., & Pescapé, A. (2026). A federated and incremental network intrusion detection system for IoT emerging threats. *IEEE Transactions on Network and Service Management*, 23, 3865–3880. https://doi.org/10.1109/TNSM.2026.3675031

[6]. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2022). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(7), 3366–3385. https://doi.org/10.1109/TPAMI.2021.3057446

[7]. Gelenbe, E., Gül, B. C., & Nakıp, M. (2024). DISFIDA: Distributed self-supervised federated intrusion detection algorithm with online learning for health Internet of Things and Internet of Vehicles. *Internet of Things, 28*, 101340. https://doi.org/10.1016/j.iot.2024.101340

[8]. Hamedi, P., Razavi-Far, R., & Hallaji, E. (2025). Federated continual learning: Concepts, challenges, and solutions. *Neurocomputing, 651*, 130844. https://doi.org/10.1016/j.neucom.2025.130844

[9]. He, X., Chen, Q., Tang, L., Wang, W., Liu, T., Li, L., Liu, Q., & Luo, J. (2023). Federated continuous learning based on stacked broad learning system assisted by digital twin networks: An incremental learning approach for intrusion detection in UAV networks. *IEEE Internet of Things Journal, 10*(22), 19825–19838. https://doi.org/10.1109/JIOT.2023.3282648

[10]. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., … Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning, 14*(1–2), 1–210. https://doi.org/10.1561/2200000083

[11]. Khraisat, A., Alazab, A., Singh, S., Jan, T., & Gomez, A. (2024). Survey on federated learning for intrusion detection system: Concept, architectures, aggregation strategies, challenges, and future directions. *ACM Computing Surveys, 57*(1), Article 7. https://doi.org/10.1145/3687124

[12]. Lavaur, L., Pahl, M.-O., Busnel, Y., & Autrel, F. (2022). The evolution of federated learning-based intrusion detection and mitigation: A survey. *IEEE Transactions on Network and Service Management, 19*(3), 2309–2332. https://doi.org/10.1109/TNSM.2022.3177512

[13]. Mao, J., Wei, Z., Li, B., Zhang, R., & Song, L. (2024). Toward ever-evolution network threats: A hierarchical federated class-incremental learning approach for network intrusion detection in IIoT. *IEEE Internet of Things Journal, 11*(18), 29864–29877. https://doi.org/10.1109/JIOT.2024.3408634

[14]. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273–1282). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html

[15]. Nakıp, M., Gül, B. C., & Gelenbe, E. (2023). Decentralized online federated G-network learning for lightweight intrusion detection. In *2023 31st International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)* (pp. 1–8). IEEE. https://doi.org/10.1109/MASCOTS59514.2023.10387644

[16]. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

[17]. Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology, 64*, 1–18. https://doi.org/10.1016/j.infsof.2015.03.007

[18]. Quyen, N. H., Hoang, N. V., Duy, P. T., & Pham, V.-H. (2024). FI-IDS: A federated incremental learning approach for intrusion detection system. In *2024 International Conference on Advanced Technologies for Communications (ATC)* (pp. 432–437). IEEE. https://doi.org/10.1109/ATC63255.2024.10908328

[19]. Rehman, M., Bahsi, H., & Kalakoti, R. (2026). Incremental federated learning for intrusion detection in IoT networks under evolving threat landscape. In *Proceedings of the 12th International Conference on Information Systems Security and Privacy* (pp. 97–107). SCITEPRESS. https://doi.org/10.5220/0014475900004061

[20]. Soltani, M., Khajavi, K., Jafari Siavoshani, M., & Jahangir, A. H. (2024). A multi-agent adaptive deep learning framework for online intrusion detection. *Cybersecurity, 7*, Article 5. https://doi.org/10.1186/s42400-023-00199-0

[21]. Su, J. (2025). Evolutionary replay-driven federated class-incremental learning for cyber-attack detection. In *Poster Volume III: The 2025 Twenty-First International Conference on Intelligent Computing* (pp. 2517–2534). https://doi.org/10.65286/icic.v21i3.80913

[22]. Truong, V. T., & Le, L. B. (2023). MetaCIDS: Privacy-preserving collaborative intrusion detection for metaverse based on blockchain and online federated learning. *IEEE Open Journal of the Computer Society, 4*, 253–266. https://doi.org/10.1109/OJCS.2023.3312299

[23]. Wu, Z., He, B., Si, Z., Qiu, C., Liao, X., & Su, C. (2026). Communication-efficient federated class-incremental intrusion detection for edge IoT networks. *Sensors, 26*(14), 4630. https://doi.org/10.3390/s26144630

[24]. Yang, X., Yu, H., Gao, X., Wang, H., Zhang, J., & Li, T. (2024). Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering, 36*(8), 3832–3850. https://doi.org/10.1109/TKDE.2024.3363240

[25]. Zhang, H., Ye, J., Huang, W., Liu, X., & Gu, J. (2025). Survey of federated learning in intrusion detection. *Journal of Parallel and Distributed Computing, 195*, 104976. https://doi.org/10.1016/j.jpdc.2024.104976

Appendix A. Complete Search Strings

Each string was submitted separately to OpenAlex, Crossref, and DBLP on 26 July 2026. Source limits and APIs can return different rankings over time; the raw query-record counts reported in the PRISMA diagram correspond to that date.

Q1. "continual federated learning" "intrusion detection"

Q2. "federated continual learning" "intrusion detection"

Q3. "federated incremental learning" "intrusion detection"

Q4. "federated continuous learning" "intrusion detection"

Q5. "federated lifelong learning" "intrusion detection"

Q6. "federated learning" "concept drift" "intrusion detection"

Q7. "federated learning" "online learning" "intrusion detection"

Q8. "federated learning" "adaptive" "network anomaly detection"

Table A1. Search interfaces, limits, and retained query records

Source	Interface and searched field	Date restriction	Result limit	Returned records
OpenAlex	Works API; general search field	1 Jan 2018–26 Jul 2026	200 per query	794
Crossref	Works API; query.bibliographic	1 Jan 2018–26 Jul 2026	1,000 per query	8,000
DBLP	Publication search API	Year checked after retrieval	1,000 per query	25

Automated screening matched the concatenated title and abstract without regard to case. A record advanced only when at least one expression from every block in Table A2 was present. Hyphens and spaces were both accepted in compound expressions such as concept drift and zero-day.

Table A2. Exact concept blocks used for automated screening

Required block	Case-insensitive expressions
Federated learning	federated; federation; federating (including longer word forms beginning with these stems)

Required block	Case-insensitive expressions
Continual or adaptive learning	continual; continuous; incremental; lifelong; online; stream; streaming; concept drift; drift-aware; drift-adapt; evolving; evolvable; evolutionary; class-incremental; task-incremental; domain-incremental; catastrophic forgetting; adaptive
Network security	intrusion; network anomaly/anomalies; cyber-attack; cyber-threat; cybersecurity; DDoS; DoS attack; zero-day; malicious network; malicious traffic; attack detection

Supplementary update queries executed on 27 July 2026 were: "federated continual learning" intrusion detection year:2017–2026 and "federated incremental learning" intrusion detection year:2017–2026. The queries were run in the Consensus academic discovery index, followed by exact-title and DOI checks on publisher domains. This update supplemented discovery only; it did not replace the three reproducible main index searches.

Table A3. Targeted supplementary search update (27 July 2026)

Record	Report	DOI	Decision
S03	FedContinualIDS for fog computing with HIKARI	10.1007/s13748-026-00443-x	Duplicate of existing database retrieval record; no new full text
S17	Federated incremental EIDS for zero-day attacks	10.1109/MNET.018.2200349	Duplicate of existing database retrieval record; no new full text
S20	TA-FCL for blockchain intrusion detection	10.1109/RMKMATE69073.2026.11518984	Duplicate of existing database retrieval record; no new full text
S21	Falcon for heterogeneous consumer IoT	10.1109/TCE.2026.3690484	Duplicate of existing database retrieval record; no new full text
X33	FedIDFR in Kubernetes	10.1109/CAC63892.2024.10865025	Duplicate of supplementary retrieval record; no new full text
X34	TSK-FedIDS for IIoT	10.1109/FASTA70174.2026.11548795	Duplicate of supplementary retrieval record; no new full text
U01	Distributed collaborative intrusion detection with FedDGC	10.1109/ICSP69961.2026.11540768	Excluded at title/abstract: heterogeneity experiment without sequential-task or retention evaluation

Appendix B. Full-Text Retrieval and Exclusion Register

Twenty-four reports could not be retrieved in full: 15 arose from the database/register stream and nine from supplementary searching. They were not counted as eligibility exclusions and contributed no extracted evidence. The register makes the coverage boundary auditable.

Table B1. Reports sought but not retrieved (n = 24)

ID	Identification stream	Title	DOI
S01	Database/register search	New Continual Federated Learning System for Intrusion Detection in SDN-Based Edge Computing	10.1002/cpe.8332
S02	Database/register search	Continual Federated Learning Approach for Distributed Network Intrusion Detection: A Case Study on Software Defined Networking	10.1007/978-3-032-06777-7_2
S03	Database/register search	Intrusion detection model for secure fog computing: harnessing with federated continual learning using the HIKARI dataset to combat emerging attacks	10.1007/s13748-026-00443-x

ID	Identification stream	Title	DOI
S04	Database/register search	Federated continual representation learning for evolutionary distributed intrusion detection in Industrial Internet of Things	10.1016/j.engappai.2024.108826
S05	Database/register search	FL-IIDS: A novel federated learning-based incremental intrusion detection system	10.1016/j.future.2023.09.019
S09	Database/register search	Stacked Broad Learning System Empowered FCL Assisted by DTN for Intrusion Detection in UAV Networks	10.1109/globecom54140.2023.10436983
S10	Database/register search	A Robust and Scalable Federated Continual Learning Framework for Adaptive DDoS Detection in Heterogeneous IoT Environments	10.1109/icc52391.2025.11161279
S11	Database/register search	FedGuard: Federated Network Intrusion Detection with Differential Privacy, Byzantine Fault Tolerance,Concept Drift Awareness, and Zero-Day Threat Response	10.1109/icpcsn68523.2026.11543526
S12	Database/register search	Intrusion Detection for Unmanned Aerial Vehicles (UAVs) via Lightweight Federated Continuous Learning	10.1109/icsadl67539.2026.11451982
S13	Database/register search	Fed-CLIDS: Network intrusion detection system based on federated meta-continuous learning	10.1109/ijcnn64981.2025.11228189
S17	Database/register search	Federated Incremental Learning based Evolvable Intrusion Detection System for Zero-Day Attacks	10.1109/mnet.018.2200349
S18	Database/register search	FNEL:An Evolving Intrusion Detection System Based on Federated Never-Ending Learning	10.1109/msn53354.2021.00047
S20	Database/register search	Adaptive Trust-Aware Federated Continual Learning for Blockchain Intrusion Detection	10.1109/rmkmate69073.2026.11518984
S21	Database/register search	Falcon: Federated Incremental Learning Framework for Intrusion Detection in Heterogeneous Consumer-Centric Internet of Things	10.1109/tce.2026.3690484
X39	Database/register sensitivity audit	FIDWATCH: Federated Incremental Distillation for Continuous Monitoring of IoT Security Threats	10.1016/j.adhoc.2024.103637
X30	Supplementary search	Federated Hoeffding Tree-Based Intrusion Detection for Data Streams	10.1002/spy2.70192
X31	Supplementary search	Online Personalized Federated Learning for UAV Intrusion Detection	10.1007/978-981-96-8725-1_32
X32	Supplementary search	User-Interactive Federated Learning for Incremental Threat Adaptation	10.1016/j.procs.2026.06.553
X33	Supplementary search	FedIDFR: Federated Incremental Detection for Kubernetes Environments	10.1109/CAC63892.2024.10865025
X34	Supplementary search	TSK-FedIDS: Interpretable Federated Incremental Intrusion Detection	10.1109/FASTA70174.2026.11548795
X35	Supplementary search	ST-FedxIDS: Drift-Adaptive Federated Intrusion Detection	10.29284/zcrk2p65
X36	Supplementary search	A Self-Healing Federated Learning Intrusion Detection System	10.34190/iccws.21.1.4467
X37	Supplementary search	Adaptive Federated Collaborative Learning for Intrusion Detection	10.65204/djes.v3i1.408

ID	Identification stream	Title	DOI
X38	Supplementary search	Adaptive Federated Learning Defense Against Byzantine Attacks and Concept Drift in IIoT	No DOI located

Table B2. Full-text reports excluded after assessment (n = 2)

ID	Title	DOI	Reason
S25	A Framework for Real-Time Continual Learning Federated Intrusion Detection Systems	10.22555/pjets.v13i2.1434	Conceptual framework; no primary quantitative experiment
S27	Incremental Federated Learning for Intrusion Detection in IoT Networks under Evolving Threat Landscape	10.48550/arXiv.2603.10776	Duplicate preprint of included conference report S28

Appendix C. Extraction Form and Coding Rules

Bibliographic fields: title, authors, year, venue, DOI, publication type. Context fields: application domain, network layer, federation topology, client roles and count. Sequential fields: task identity, class/domain/task increment, stream order, drift trigger, synchronization and target access. Algorithm fields: model, aggregation, replay, regularization, distillation, model expansion, compression, privacy and robustness. Evaluation fields: dataset, attack classes, preprocessing, client partition, IID/non-IID parameter, baselines, metrics, repeats, statistical uncertainty, hardware, communication, latency, energy and memory. Reproducibility fields: data URL, study code, environment, hyperparameters, seeds, split manifest and executable package.

Retention-tested coding required an explicit evaluation of earlier tasks/classes after later learning, a forgetting or transfer measure, or a before/after comparison of retained knowledge. Studies with sequential updates but no such outcome were coded adaptation-only. Missing information was coded ‘not reported’; it was not reconstructed from common practice.

Appendix D. Automatic-Screen Sensitivity Audit

The initial three-block rule excluded 2,865 records. A secondary rule was applied to the complete exclusion log to search for temporal or adaptive terminology combined with network, IoT, industrial-control, or anomaly language that the original security block could miss. This rule-based audit identified two direct false negatives. FLAME-IoT was retrieved and included as adaptation-only evidence. FIDWATCH was moved to the database retrieval register but remained unavailable in full text. The audit revised the automatic exclusions to 2,863 and the title/abstract queue to 212; it was not an independent human validation sample.

Table D1. Records reclassified by the sensitivity audit

Record	Report	Original decision	Revised decision	Reason
R00239 / S30	FLAME-IoT (Alatawi, 2026)	Automatic exclusion	Included; adaptation-only	IIoT anomaly language was not captured by the original security block; full text reported synthetic drift and quantitative FL experiments
R00266 / X39	FIDWATCH (Alrashdi et al., 2024)	Automatic exclusion	Report sought; not retrieved	IoT security-threat language was not captured by the original block; publisher page required institutional full-text access

Appendix E. PRISMA 2020 Reporting Map

The table maps each PRISMA 2020 item to a manuscript location. Items designed for meta-analysis are marked not applicable because the review used a systematic mapping synthesis and did not estimate pooled effects.

Table E1. PRISMA 2020 checklist-to-manuscript crosswalk

Item	PRISMA topic	Manuscript location	Status or qualification
1	Title	Title	Reported
2	Abstract	Structured abstract	Reported
3	Rationale	Sections 1–2	Reported
4	Objectives	Section 1 and RQ table	Reported
5	Eligibility criteria	Section 3.3; Table 3	Reported
6	Information sources	Section 3.2; Tables A1 and A3	Reported with access limitation
7	Search strategy	Section 3.2; Appendix A	Main and supplementary update queries reported
8	Selection process	Section 3.3; Figure 1	Single-reviewer process disclosed
9	Data collection process	Section 3.4	Reported
10a	Outcome data items	Section 3.4; Appendix C	Reported
10b	Other data items	Section 3.4; Appendix C	Reported
11	Study risk-of-bias assessment	Section 3.5; Table 9	Not performed; descriptive methodological appraisal only
12	Effect measures	Section 3.6	No common effect measure
13a	Eligibility for each synthesis	Sections 3.4–3.6	Reported
13b	Data preparation	Sections 3.4 and 3.6	No statistical conversion
13c	Display methods	Figures 2–3; Tables 4–10	Reported
13d	Synthesis methods	Section 3.6	Narrative mapping
13e	Heterogeneity exploration	Sections 4.2–4.4 and 5.2	Qualitative
13f	Sensitivity analyses	Sections 3.3 and 4.6; Appendix D	Screening and reporting-completeness checks reported
14	Reporting-bias assessment	Section 7	Not formally assessed
15	Certainty assessment	Section 7	Not formally assessed
16a	Study selection	Section 3.3; Figure 1	Reported
16b	Excluded studies	Table B2	Reported with reasons
17	Study characteristics	Tables 4–7	Reported
18	Risk of bias in studies	Table 9	No validated risk-of-bias assessment; descriptive domains reported
19	Results of individual studies	Table 8	Reported in original units
20a–d	Results of syntheses	Sections 4.4–5.3	Narrative synthesis
21	Reporting biases	Section 7	Discussed as limitation
22	Certainty of evidence	Section 7	No formal grading
23a–d	Discussion	Sections 5–8	Reported

Item	PRISMA topic	Manuscript location	Status or qualification
24a	Registration	Section 3	Not registered
24b	Protocol access	Section 3	No public protocol
24c	Amendments	Section 3	No registered protocol amendments
25	Support	Declarations	Reported; no external funding. The APC will be funded by the authors.
26	Competing interests	Declarations	Reported; the authors declare no conflict of interest.
27	Availability of data/code/materials	Data Availability; Appendices A–D	Supplementary reproducibility package supplied with the submission; no public repository deposit claimed.