Pinto *et al. Cybersecurity* (2026) 9:135
https://doi.org/10.1186/s42400-026-00567-6

Cybersecurity

SURVEY

Open Access

Balancing the trilemma: a survey of federated anomaly detection for secure cyber-physical systems

Andrea Pinto[1*], Yezid Donoso[1] and Jairo A. Gutierrez[2]

Abstract

The proliferation of Cyber-Physical Systems (CPS) across critical infrastructure has created an unprecedented attack surface where digital threats may precipitate catastrophic physical consequences. As conventional centralized security paradigms fail to address the scale and complexity of these environments, Federated Learning (FL) has emerged as a transformative approach, enabling collaborative, edge-native anomaly detection without centralizing sensitive data. This paper presents a comprehensive survey and critical analysis of the state-of-the-art in securing CPS through advanced FL. We introduce a novel multi-axis taxonomy that systematically categorizes the field by architecture, detection methodology, application domain, and privacy-preservation scheme. Building on this analysis, we synthesize these findings into a prescriptive framework to guide the selection of appropriate security archetypes for different CPS domains. Through this lens, we deconstruct the—the trade-off between accuracy, communication, and privacy- that governs every FL design. Our analysis synthesizes the dominant trends, including the convergence of deep learning with edge computing and the increasing sophistication of privacy-enhancing technologies. We further identify critical research gaps, including the scarcity of physical testbeds, limited resilience against advanced adversarial attacks, and underdeveloped explainability. The paper concludes by defining the critical frontiers for future research, emphasizing the need to resolve the inherent tension between FL's privacy goals and the transparency requirements of Explainable AI (XAI) to build truly trustworthy systems.

Keywords Federated learning, Cybersecurity, Critical infrastructures, Cyber-physical systems, Anomaly detection

Introduction

Cyber-Physical Systems (CPS) are a fundamental manifestation of the convergence of cyber and physical worlds, representing sophisticated integrations of computational entities (such as sensors and actuators) with physical objects and infrastructure (Ali et al. 2024; Singh et al. 2022). These systems are designed to monitor and control physical processes by leveraging interconnected networks and advanced digital technologies. Unlike traditional isolated industrial systems, CPS operates in open, digitalized environments, enabling real-time data exchange, autonomous decision-making, and intelligent control. This transformation from closed to open structures has revolutionized various industries by integrating Information Technology (IT) with Operational Technology (OT) (Friha et al. 2023). The proliferation of CPS is particularly evident across critical infrastructures (CIs) and vital societal sectors due to their ability to enhance efficiency, reliability, and smartness; sectors such as: Energy, Transportation, Healthcare, and Manufacturing (Sharma and Shambharkar 2025; Yang et al. 2023a).

*Correspondence:
Andrea Pinto
ya.pinto10@uniandes.edu.co
[1] Systems and Computing Engineering Department, School of Engineering, Universidad de los Andes, 111711 Bogotá, Colombia
[2] Department of Computer and Information Sciences, School of Engineering, Computer, and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand

The increasing frequency and sophistication of cyberattacks against these interconnected systems highlight the urgent need for robust and adaptable security measures; in 2024 alone, ransomware attacks targeting the industrial sector surged by 87% (Dragos 2024). Conventional security methods are often insufficient due to the inherent characteristics of CPS and Industrial Internet of Things (IIoT) devices, such as resource constraints, heterogeneity, and the scale of deployments (Sharma and Shambharkar 2025; Nandanwar and Katarya 2025a). The attack surface is no longer confined to the digital realm; it extends directly into the physical world, where cyber intrusions can precipitate tangible, often destructive, consequences (Ahanger et al. 2025). This risk is not theoretical; a recent U.S. intelligence report confirmed that foreign adversaries have successfully manipulated water and energy systems, demonstrating the potential to cause direct physical damage and deny critical services. The vulnerabilities inherent in these systems can be exploited by malicious actors to compromise not just data confidentiality or integrity, but the very safety and operational stability of critical infrastructure (Singh 2025). This necessitates the development of advanced cybersecurity systems specifically designed to address the intricacies of these environments.

Conventional security paradigms, such as signature-based Intrusion Detection Systems (IDS), are inherently reactive, designed for known threats (Manivannan 2024), while proving ineffective against novel, zero-day attacks (Friha et al. 2023). This deficiency is amplified as adversaries increasingly leverage artificial intelligence (AI) to launch more sophisticated and evasive campaigns, with some analysts predicting that 17% of all cyberattacks will soon involve generative AI (Bhogal et al. 2025). Furthermore, the centralized architectures of traditional security solutions introduce performance bottlenecks and single points of failure (Belenguer et al. 2025). The protracted remediation timeline—averaging 272 days for an industrial sector breach—underscores the inadequacy of currents models and results in severe financial and operational costs (Security 2024). These limitations are compounded by the inherent processing, memory, and energy constraints of most CPS devices, which make embedding complex, resource-intensive security features impractical (Huong et al. 2021). Legacy systems also lack the adaptability for real-time response, and their anomaly-based counterparts are often plagued by high false-positive rates (Manivannan 2024), leading to critical alert fatigue. Therefore, the cyber-physical age necessitates a paradigm shift towards robust, intelligent, and adaptive security frameworks.

To address these challenges, a paradigm shift towards decentralized architecture is essential. Edge computing emerges as a foundational solution, moving computation and data storage closer to the source to overcome the latency and bandwidth limitations of centralized models. This architectural evolution is critical for the real-time responsiveness demanded by CPS (Makris et al. 2025; Javeed et al. 2024). However, architecture alone is insufficient; it requires an equally advanced intelligence paradigm. This is where Federated Learning (FL) provides a transformative approach. Specifically, as a decentralized Machine Learning (ML) technique, FL enables multiple distributed edge devices to collaboratively train a shared AI model without ever exposing their raw, sensitive data (Agrawal et al. 2022). This method not only preserves data privacy -a critical concern in CPS- but also mitigates the single points of failure and communication overheads associated with traditional, centralized ML (Friha et al. 2022).

However, the transition to decentralized security paradigms reveals a complex landscape of technical trade-offs necessitated by the inherent heterogeneity and extreme resource constraints of CPS architectures (Belenguer et al. 2025; Elkhodr 2025). These environments impose a critical performance trilemma—the inescapable tension between detection accuracy, communication efficiency, and data privacy (Wehbi et al. 2023)—where the optimization of any single dimension typically introduces technical friction into the remaining two. For instance, while perturbation-based methods such as Differential Privacy (DP) provide formal guarantees against inference attacks, the resulting statistical noise can obscure subtle anomalies in high-velocity sensor data, potentially degrading the precision required for safety–critical actuation (You et al. 2024). Similarly, the application of robust cryptographic techniques to ensure update confidentiality significantly escalates computational overhead and network latency (Namakshenas et al. 2024), which often conflicts with the real-time operational requirements of critical infrastructure. Furthermore, prioritizing communication efficiency through aggressive model compression or reduced update frequencies can limit the global model's ability to capture the sophisticated temporal dynamics characteristic of modern industrial processes, thereby diminishing its effectiveness against zero-day threats. This trilemma underscores that implementing federated security in CPS is not merely a matter of linear enhancement but a strategic balancing of competing requirements.

The core contributions of this paper are threefold. First, we develop a novel, dual-layered taxonomic framework. We begin by establishing a comprehensive multi-axis taxonomy that provides a structured overview of the state-of-the-art. We then synthesize this analysis into a prescriptive synthetic taxonomy that guides the selection of security archetypes for different CPS domains. Second,

we provide a critical analysis of the field's foundational challenges, deconstructing the performance trilemma with a novel visual model and systematically identifying research gaps in a structured format. Finally, based on this comprehensive analysis, we identify critical research gaps and "blind spots" in the current literature and propose a concrete agenda of open challenges and future research directions to guide the advancement of the field.

The remainder of this paper is structured to build upon this foundation. Sect. "Survey Methodology" details our Survey Methodology, utilizing the PRISMA framework to ensure a transparent and reproducible literature selection process. Sect. "Foundational Concepts" establishes the Foundational Concepts of CPS and FL, culminating in a synthesis that maps these core principles to our classification criteria. The core of our contribution is presented in Sect. "A Multi-Axis Taxonomy of FL-Based Anomaly Detection Systems", where we introduce a comprehensive multi-axis taxonomy—organized by architecture, methodology, threat model, and privacy—and synthesize these dimensions into a prescriptive framework for selecting security archetypes. Sect. "Deconstructing the Field: Trends, Trade-offs, and Foundational Challenges" transitions to a synthesis and critical analysis of the field, deconstructing the performance trilemma and highlighting systemic research gaps. Finally, Sect. "Open Challenges and Future Research Directions" outlines future research directions and Sect. "Conclusion" offers concluding remarks.

Survey methodology

To ensure a comprehensive, transparent, and reproducible analysis of the state-of-the-art, this survey was conducted following the systematic principles of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. This structured approach provides a clear methodological foundation for the identification, screening, and selection of relevant literature, ensuring that the resulting analysis is both rigorous and unbiased. The methodology encompassed a precise literature search strategy, the application of strict inclusion and exclusion criteria, and a multi-stage screening process to distill the most relevant and high-impact research from a broad initial pool of articles.

Search strategy and data sources

The literature search was executed across five leading academic databases: IEEE Xplore, Scopus, Web of Science (WoS), SpringerLink, and ScienceDirect (Elsevier). To focus on the most contemporary research, the search was constrained to peer-reviewed journal and review articles published between January 2023 and the present. A highly specific search query was constructed to capture the exact intersection of the core technologies relevant to this survey. The query, uniformly applied across all databases, was: ("Federated Learning" OR "Distributed Machine Learning") AND "Privacy Preserving" AND ("Cyber Physical Systems" OR "Industrial Control Systems") AND ("Anomaly Detection" OR "Intrusion Detection" OR "Attack Detection"AND ("Edge Computing" OR "Fog Computing" OR "Edge AI" OR "On-device AI").

Inclusion and exclusion criteria

A stringent set of criteria was established to govern the selection process.

Inclusion Criteria: An article was considered for inclusion only if it was a peer-reviewed journal or review article, written in English, and published within the specified 2023–2025 timeframe.

Exclusion Criteria: An article was excluded if it met one or more of the following conditions:

- Irrelevant Application Domain: The research did not have a mandatory and explicit application to CPS, Industrial Internet of Things (IIoT), or other CIs). This was the most critical exclusion factor.
- Lack of Anomaly Detection Focus: The paper did not primarily focus on anomaly detection, intrusion detection, or attack detection, even if it discussed FL in a security context.
- Absence of Core Technologies: The work did not involve FL or a comparable Distributed Machine Learning (DML) approach, or it failed to implement these within an edge, fog, or on-device computing paradigm.
- Non-Applicable Document Type: The publication was a conference proceeding, book chapter, patent, thesis, or non-peer-reviewed article.

Multi-stage screening and selection results

The screening and selection process was conducted in multiple stages, with each decision validated by at least two authors to ensure consensus and minimize subjective bias. The initial search yielded a total of 744 articles across all databases before deduplication: Web of Science (294), Scopus (162), ScienceDirect (64), IEEE Xplore (42), and SpringerLink (182).

First, all retrieved results were aggregated, and duplicates were systematically removed. The remaining articles then underwent title and abstract screening, where publications clearly outside the research scope were discarded. Subsequently, the full text of each remaining article was thoroughly evaluated against the exclusion criteria defined in Sect. "Inclusion and Exclusion Criteria". This critical stage ensured that only papers satisfying

the precise intersection of all required domains were retained.

The rigorous application of this filtering protocol significantly refined the initial results. The final number of articles deemed relevant from each database were: 21 from Web of Science, 32 from Scopus, 11 from ScienceDirect, 29 from IEEE Xplore, and 30 from SpringerLink. The final, deduplicated corpus of these selected articles forms the foundational body of literature for the taxonomic analysis presented in this survey.

Foundational concepts

This section lays the theoretical groundwork essential for the analyses presented in this paper. We begin by delineating the core principles of CPS, detailing their fundamental architecture, operational feedback loops, and the inherent characteristics that define their unique security challenges. Subsequently, we introduce FL, explaining the mechanics of its decentralized training process and the key attributes that position it as a transformative solution for securing these complex environments. A clear grasp of these domains is crucial for contextualizing the advanced taxonomies and vulnerabilities discussed in the subsequent sections.

Cyber-physical systems (CPS)

CPS are systems that integrate computational entities with the physical world and its associated processes, defined by the tight coupling of cyber systems with physical ones (Singh et al. 2022) (Alturki, et al. 2025). This integration combines physical components, such as energy generation sources and storage devices, with digital technologies like sensors, communication networks, and control systems (Ahmad et al. 2025). Through this combination, a CPS can be described as a physical system controlled by embedded software, where computational entities like sensors and actuators collaborate (Kheddar et al. 2023). This structure facilitates real-time monitoring, control, and optimization of physical processes, leveraging data-accessing and data-processing services available on the Internet to execute common tasks (Singh et al. 2022). The implementation of CPS is instrumental across numerous critical sectors, found in industrial environments like manufacturing plants, power systems, and transportation networks (Macas et al. 2022). Given the essential services they provide, CPS are classified as CIs, making their robustness crucial. These systems require an Internet of Things (IoT) Click or tap here to enter text.infrastructure and include applications such as smart grids, Intelligent Transportation Systems (ITS) (Makris et al. 2025), and healthcare systems (Internet of Medical Things/IoMT) (Jayanthiladevi et al. 2025).

The key difference between IT security and CPS security lies in their primary goals and the nature of consequences when those goals are not met. IT security primarily focuses on safeguarding the Confidentiality, Integrity, and Availability (CIA) of data, where a failure typically results in data breaches, financial loss, or service disruptions (Belenguer et al. 2025). In contrast, CPS security extends beyond data protection to directly secure physical processes and ensure safety (Nandanwar and Katarya 2025b). Because CPS integrates computational entities with the physical world, data security is a crucial means to an end, not the end itself. Consequently, failure in CPS security can have catastrophic real-world impacts far beyond data loss, including physical equipment damage, widespread power outages, environmental disasters, and direct threats to human life, as seen in attacks on industrial control systems, medical devices, or smart grids (Rizvi and Demeri 2025).

The core operational principle of a CPS is the Cyber Physical Feedback Loop, a continuous, interactive cycle between the system's computational components and its physical processes. This tight coupling, illustrated in Fig. 1, is essential for enabling real-time monitoring, control, and adaptation. In a security context, this loop introduces a critical vulnerability: an attacker who compromises the 'processing and decision making' phase can inject malicious commands, leading to physical damage in the 'actuation' phase. For example, a compromised industrial controller could receive false sensor data, causing it to command an actuator to dangerously

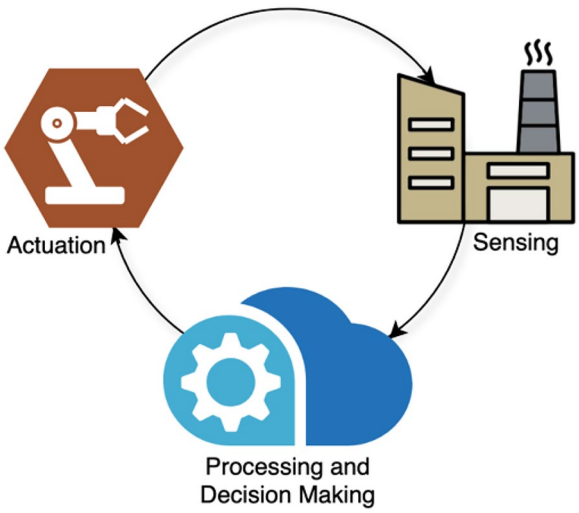

Fig. 1 The cyber-physical feedback loop and its role in the threat model axis. This figure illustrates the bidirectional dependency between the cyber and physical layers, establishing the foundational rationale for analyzing physical-layer consequences within the adversarial landscape

over-pressurize a pipeline. This is a defining characteristic of a cyber-physical attack. Additionally, this loop function has three distinct phases. It initiates with sensing, where physical components like sensors capture real world data and transmit it to the cyber layer (Ahmad et al. 2025). In the subsequent processing and decision-making phase, computational platforms analyze this data to derive actionable insights and make autonomous decisions (Singh et al. 2022). The cycle concludes with actuation, where commands based on these decisions are sent back through control systems to influence or adjust the physical processes (Huong et al. 2021). This dynamic and continuous process allows CPS to monitor and control physical operations in real-time and adapt to changing conditions. In cybersecurity contexts, this feedback loop is particularly crucial. It enables security systems to continuously refine their models based on new data, allowing for rapid, adaptive responses to evolving cyber threats near the data sources (Duy et al. 2024; Jayanthiladevi et al. 2025; Kheddar 2025).

The architecture of a CPS is defined by the tight coupling of computational elements with physical processes, facilitated by robust communication networks. While specific implementations vary, CPS architectures are commonly conceptualized through layered models that delineate functional responsibilities. Several layered architectural models exist, tailored to specific domains. A prevalent three-layer model, particularly in the context of Industry 5.0, consists of Nandanwar and Katarya 2025b:

- The Physical Layer: Comprising tangible assets such as sensors and actuators that directly interact with the physical environment to capture real-world data (Senthil et al. 2025).
- The Cyber Layer: This middleware layer includes computational resources like edge computing nodes and cloud servers, along with communication networks, to aggregate and analyze data from the physical layer.
- The Cognitive Layer: This layer integrates advanced AI and ML algorithms to enable autonomous decision-making, predictive maintenance, and intelligent process control.

Similarly, modern Industrial Control Systems (ICS) often feature a three-tiered structure composed of a physical field layer, a supervisory layer for monitoring and control such as SCADA systems, and an enterprise application layer for management (Kheddar et al. 2023). IoT-enabled systems also frequently adopt a three-layer design, consisting of a perception layer for information sharing among devices, a network layer for data transmission, and an application layer for user interaction (Ahanger et al. 2025). Beyond these layered models, specific CPS applications deploy distinct architectures to meet specialized demands. For instance, a decentralized IIoT architecture utilizes an edge-cloud structure to perform processing intelligence close to the data sources, thereby reducing bandwidth requirements and attack response times. In contrast, the Internet of Robotic Things (IoRT) often employs a hierarchical framework with an IoRT device layer, more powerful edge nodes, and a central cloud server (Nkoom et al. 2025a). Other specialized architectures include those for Smart Grid Cyber-Physical Power Systems (SG-CPPS) (Abdelkader et al. 2024), which integrate traditional power systems with advanced communication and control technologies, and IoT-enabled healthcare systems, which may incorporate blockchain for security and deep learning for data analysis.

The increasing deployment and interconnectedness of CPS, driven by the rise of IoT and 5G/6G mobile communications, have increased their susceptibility to cyber threats and vulnerabilities (Nandanwar and Katarya 2025b). This creates significant challenges related to security and reliability, with the potential for cyber-attacks to cause cascading effects (Abdelkader et al. 2024) and physical damages that extend beyond the digital realm to include property losses and threats to human lives (Makris et al. 2025). Securing these systems requires overcoming a range of inherent challenges, beginning at the device level with significant resource constraints (Yousefnezhad et al. 2020) and issues of data scarcity and quality (Friha et al. 2023) (Ahanger et al. 2025) (Abdelkader et al. 2024) that hinder the training of effective models. The inherent data heterogeneity from diverse sources further complicates security analytics (Ahsan et al. 2025). Beyond these data-centric issues, CPS face critical architectural hurdles, including the need for real-time processing without introducing unacceptable latency (Kheddar et al. 2023) and ensuring scalability as device numbers grow (Ali et al. 2025a). Furthermore, many environments incorporate legacy systems with outdated protocols that are difficult to secure (Kheddar et al. 2023). Compounding these issues is the dynamic nature of threats themselves, such as zero-day exploits and Advanced Persistent Threats (APTs) (Nandanwar and Katarya 2025b), and the fact that many advanced detection models are "black boxes" that lack interpretability (Kanyama et al. 2024). This forces designers to navigate complex trade-offs between security and operational efficiency (Yousefnezhad et al. 2020). Finally, overarching concerns of data privacy and confidentiality and the unpredictable risk of human factors (Abdelkader et al. 2024) underscore the critical need for effective security

solutions like Intrusion Detection Systems (IDS) (Nandanwar and Katarya 2025b).

Federated learning (FL)

The deployment of traditional machine learning-based IDS has historically relied on data-centric centralized architecture, wherein raw training data from all network nodes are aggregated and processed on a single server or cloud platform (Zhang et al. 2024). This paradigm, however, introduces significant impediments when applied to the security of modern CPS and the IIoT. A fundamental challenge posed by this paradigm is the inherent risk to data privacy. The aggregation of potentially sensitive operational data from disparate industrial organizations or smart devices creates a trove of confidential information susceptible to leakage (Yang et al. 2023b) as even anonymized data can be re-identified, and a single breach of the central server compromises the entire dataset. Sharing such raw data with third-party entities is often untenable and may conflict with data management regulations. Furthermore, the massive volume of data generated by IIoT devices imposes considerable communication overhead, leading to increased network latency and power consumption that is prohibitive for real-time threat detection (Huong et al. 2021).

Architecturally, the centralized model introduces a single point of failure and attack (Ali et al. 2024). A compromise or malfunction of the central server can neutralize the entire security apparatus of the system (Belenguer et al. 2025). This model also faces scalability challenges, as conventional algorithms struggle to scale with the massive data volumes characteristic of CPS, and may fail to generalize effectively across the diverse network contexts and heterogeneous devices present in these ecosystems (Belenguer et al. 2025). Consequently, a paradigm shift toward a decentralized approach is not advantageous but necessary for robust CPS security.

FL has emerged as the leading decentralized machine learning technique to address the limitations of centralized models. Its core principle is the collaborative training of a shared model by multiple clients without the explicit exchange of their local, raw data (Ali et al. 2025a; Friha et al. 2022). This preserves privacy while minimizing training-associated costs. The FL process, illustrated in Fig. 2, is fundamentally iterative and can be delineated into a distinct cycle:

- *Local Training on Private Data*: The process begins with the clients with edge devices or sensors, each of which holds a private, local dataset. Each client independently trains a local model version using its own data, ensuring that sensitive information never leaves the device perimeter (Carvalho Bertoli et al. 2023).
- *Model Parameter Communication:* Instead of transmitting raw data, clients send their updated model parameters (learned weights or gradients) to a central aggregator (Ma and Su 2025).
- *Global Model Aggregation:* The aggregation server collects these parameter updates from multiple clients and combines them to produce an improved, more robust global model. The canonical algorithm for this step is Federated Averaging (FedAvg), where client model updates are averaged, often weighted by the size of their respective local datasets (Kheddar 2025).
- *Iterative Refinement:* The server distributes the newly consolidated global model parameters back to the clients, who use them as the baseline for the next round of local training. This iterative cycle allows the global model to progressively incorporate collective knowledge from all participating clients, enhancing its accuracy and generalization capabilities over time (Belenguer et al. 2025).

By design, this process significantly reduces data transfer overhead (Huong et al. 2021) and system communication costs while offering a scalable (Makris et al. 2025) and resilient framework that mitigates the single-point-of-failure risk.

FL topologies for CPS

The implementation of the FL process can be instantiated through distinct network topologies, the selection of which is contingent upon the specific resilience, control, and trust requirements of the target CPS. The prevailing topology is the Centralized, Server-Orchestrated FL Model (Tabassum et al. 2022), wherein a dedicated central server assumes an authoritative role, managing the end-to-end training lifecycle. This includes client selection, model distribution, update aggregation, and redistribution (Zhang et al. 2025). Such a hierarchical structure demonstrates high suitability for structured CPS environments (Javeed et al. 2024), like smart manufacturing plants, that possess a logical central control plane.

Conversely, to more effectively mitigate the risks associated with a central point of failure, a Decentralized, Peer-to-Peer (P2P) Model may be implemented (Friha et al. 2023). This server-less architecture empowers clients to coordinate directly for the purpose of exchanging and aggregating model updates (Belenguer et al. 2025). The P2P model affords superior fault tolerance and is therefore particularly advantageous for dynamic, ad-hoc CPS networks, such as vehicular ad-hoc networks (VANETs) or autonomous drone swarms, where a central server is either impractical or

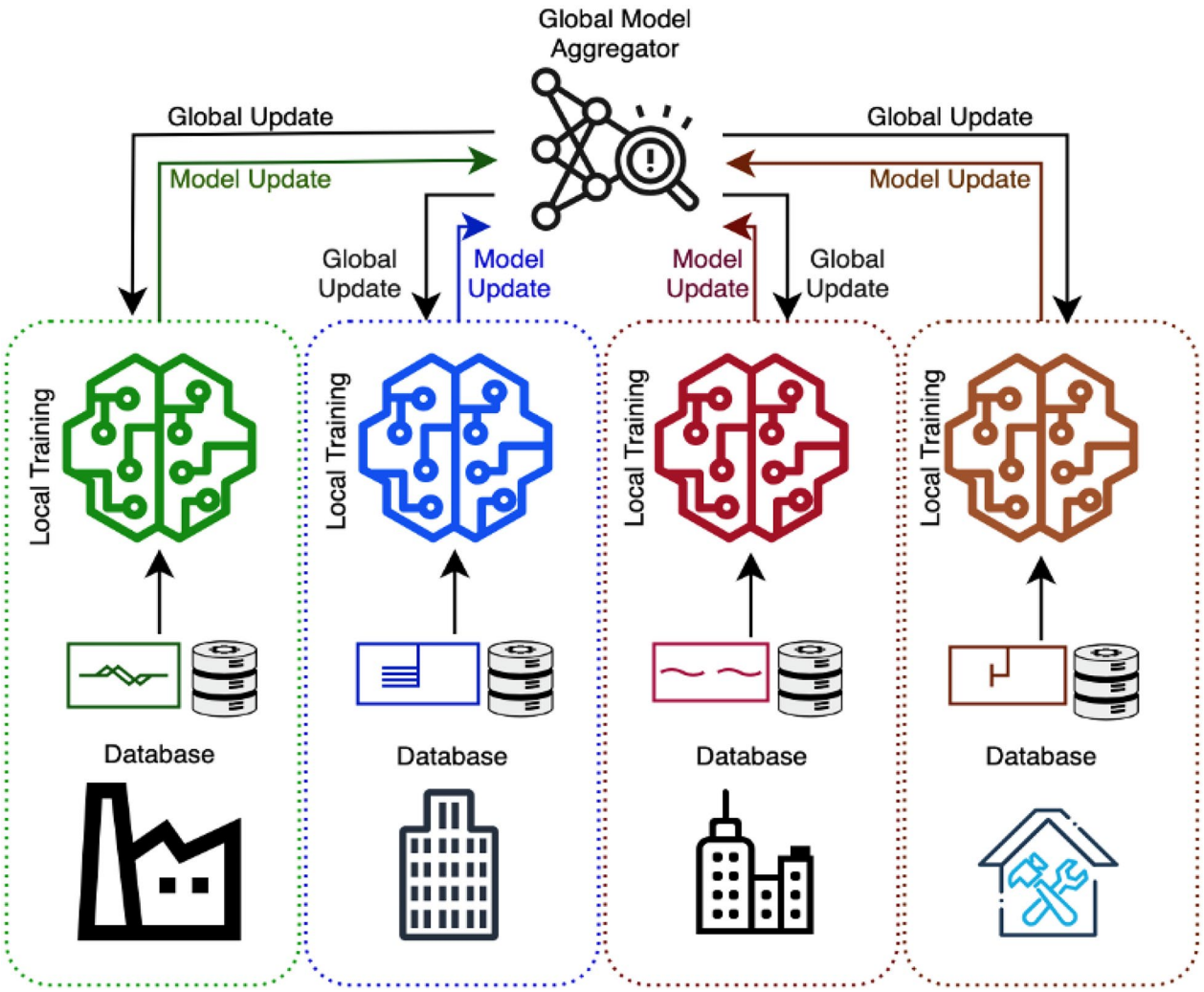

Fig. 2 The federated learning process as the basis for the anomaly detection methodology axis. This diagram deconstructs the iterative cycle of local training and global aggregation, providing the technical context for the methodological taxonomy of model updates and robust aggregation

introduces an unacceptable critical vulnerability (Shenoy et al. 2025). This topology can be further hardened through the integration of distributed ledger technologies, such as blockchain, to ensure the integrity and auditable transparency of model exchanges.

While decentralized and peer-to-peer topologies offer superior fault tolerance and effectively mitigate the single-point-of-failure risks associated with centralized models, they simultaneously introduce a more complex and fragile trust model (Friha et al. 2023). In the absence of a central authority to curate or validate participating clients, these server-less architectures become inherently more susceptible to the sophisticated Byzantine and Collusive attacks that emerge during edge implementation (Singh et al. 2022; Praharaj et al. 2025).

Implementation challenges at the CPS edge

A primary impediment is statistical heterogeneity. The data distributions across edge clients in a CPS are typically Non-Identically and Independently Distributed (Non-IID), a direct consequence of their unique operational environments (Taslimasa et al. 2023). This statistical divergence can impede the convergence of the global model, thereby degrading its predictive accuracy (Su and Zhang 2025; Makris et al. 2025). Furthermore, the FL process itself introduces new attack surfaces. Because the central aggregator must, by design, trust the updates submitted by participating clients, the system becomes vulnerable to adversarial attacks. These include model poisoning, where malicious clients submit corrupted updates to sabotage the global model, and inference attacks, which aim to reverse-engineer private training

data from the communicated model parameters (Ali et al. 2024).

This navigation of the performance trilemma—balancing detection accuracy, communication efficiency, and data privacy—represents the core technical friction in CPS security (You et al. 2024; Gupta, et al. 2023). The trade-off is rooted in the conflicting mathematical and architectural requirements of each objective. For instance, the implementation of robust Privacy-Enhancing Technologies (PETs) like DP often requires the injection of statistical noise into model updates to thwart inference attacks; however, this perturbation can obscure the subtle, high-velocity signal patterns necessary for high-accuracy anomaly detection (Yang et al. 2023c). Similarly, while cryptographic methods like Homomorphic Encryption (HE) offer strong privacy, they significantly escalate communication overhead and computational latency, which conflicts with the real-time responsiveness required by critical infrastructure (Poorazad et al. 2024; Hossain et al. 2021). Furthermore, addressing data heterogeneity (Non-IID data) often necessitates more frequent communication rounds or more complex global models, which directly compromises the communication efficiency essential for resource-constrained edge devices. Consequently, an optimization in one dimension inevitably introduces performance degradation in the remaining two, necessitating a context-aware selection of security archetypes.

Beyond these foundational hurdles, the complexity of the CPS edge introduces multi-faceted dimensions of heterogeneity and an expanding adversarial landscape (Su and Zhang 2025). While statistical heterogeneity remains a recognized barrier, it is compounded by client heterogeneity, where disparate hardware constraints, energy profiles, and communication reliability among IIoT devices create "stragglers" that destabilize the iterative training process (Xia et al. 2025). Furthermore, model heterogeneity emerges in complex ecosystems where different network tiers utilize varying neural architectures or hyperparameter configurations to meet local operational requirements, complicating the seamless aggregation of learned weights.

This structural complexity is mirrored by an evolving threat surface that extends significantly beyond standard model poisoning and inference attempts. CPS environments are uniquely vulnerable to Byzantine threats and Collusive attacks, where multiple compromised clients coordinate their updates to subtly manipulate the global model while remaining below the threshold of statistical outlier detection (Belenguer et al. 2025). Additionally, sophisticated vectors such as backdoor attacks, gradient leakage, and Membership Inference Attacks (MIA) allow adversaries to potentially reverse-engineer sensitive industrial states or operational patterns from high-resolution updates (Singh et al. 2022; Friha et al. 2023). These advanced adversarial risks necessitate a shift from simple trust-based aggregation toward more robust, reputation-based defense mechanisms (Ahmad et al. 2025; Alabdulatif 2025).

Mapping fundamentals to the taxonomy

The foundational characteristics of CPS and FL established in the preceding sections provide the technical rationale for the multi-axis taxonomy introduced in Sect. "A Multi-Axis Taxonomy of FL-Based Anomaly Detection Systems". This synthesis ensures that the classification of existing research is grounded in the operational and security realities of the cyber-physical domain.

The tight coupling of the cyber-physical feedback loop (detailed in Sect. "Cyber-Physical Systems (CPS)") and the catastrophic potential of compromised actuation directly shapes the Application Domain and Threat Model Axis (Sect. 4.3). This link ensures that the taxonomy categorizes defense mechanisms based on their ability to protect physical outcomes—such as grid stability or patient safety—rather than just computational data integrity.

Furthermore, the pervasive issues of client and model heterogeneity (discussed in Sect. "Federated Learning (FL)") at the edge necessitate the Architectural Paradigm Axis (Sect. "Architectural Paradigms for CPS Resilience"). The selection of centralized, decentralized, or hierarchical models is presented not merely as a structural choice, but as a strategic response to managing diverse hardware profiles and statistical Non-IID data distributions inherent in CPS environments.

Finally, the Performance Trilemma—the inescapable tension between detection accuracy, communication efficiency, and data privacy (introduced in Sect. "Federated Learning (FL)")—serves as the primary driver for the Anomaly Detection Methodology Axis (Sect. "Anomaly Detection Methodologies for the CPS Edge") and the Privacy-Preservation Axis (Sect. "Domain-Specific Threat Models and Use Cases"). By establishing these links, the taxonomy functions as a structured mapping of how the foundational constraints of CPS are translated into specific federated security solutions.

A multi-axis taxonomy of FL-based anomaly detection systems

To systematically analyze the field of federated anomaly detection for CPS, this section introduces a novel multi-axis taxonomy. This framework moves beyond a simple linear review by organizing the existing research according to four distinct, yet related, dimensions that directly respond to the foundational CPS characteristics and

the performance trilemma established in Sect. "Survey Methodology". Such a methodology allows for detailed analysis, revealing key relationships between different approaches and identifying significant research gaps.

Our analysis is structured along the following four axes: (1) the architectural paradigm, (2) the core anomaly detection methodology, (3) the target CPS application domain and threat model, and (4) the implemented privacy-preservation scheme. Each of the following subsections is dedicated to examining literature from one of these perspectives, thereby providing a comprehensive and structured view of the state-of-the-art and demonstrating how specific design choices mitigate the technical friction inherent in mission-critical environments.

Architectural paradigms for CPS resilience

The architecture of a FL system fundamentally dictates how learning is coordinated, how model updates are aggregated, and how the system scales and defends against structural vulnerabilities. Beyond mere connectivity, the selection of a paradigm serves as a strategic response to the foundational risks identified in Sect. "Foundational Concepts", such as single points of failure and the presence of "straggler" nodes in heterogeneous environments. The extant literature reveals three primary architectural paradigms, which we categorize as Centralized, Decentralized (Peer-to-Peer), and Hierarchical FL, as depicted in Fig. 3**.** Each model presents distinct trade-offs regarding efficiency, robustness, and communication overhead, directly influencing the system's position within the performance trilemma between accuracy, privacy, and latency. Table 1 provides a comparative analysis of these paradigms, highlighting their respective impact on CPS security and operational performance.

Server-orchestrated models and centralized trust

In the centralized paradigm, a single, monolithic server orchestrates the entire training process, often utilizing frameworks like Keras and TensorFlow for implementation (Khacha et al. 2024). This server is responsible for initializing and distributing a global model to a cohort of participating clients. Each client then performs local training on its private dataset and transmits only the resulting model parameters or gradients back to the server. The central server aggregates these updates, most commonly using the Federated Averaging (FedAvg) algorithm (Belenguer et al. 2025), to generate a refined global model for the subsequent training round. This iterative process continues until model convergence is achieved. Specific implementations, such as FED-IDS, may utilize a central discriminator network as the core aggregator component (Tabassum et al. 2022).

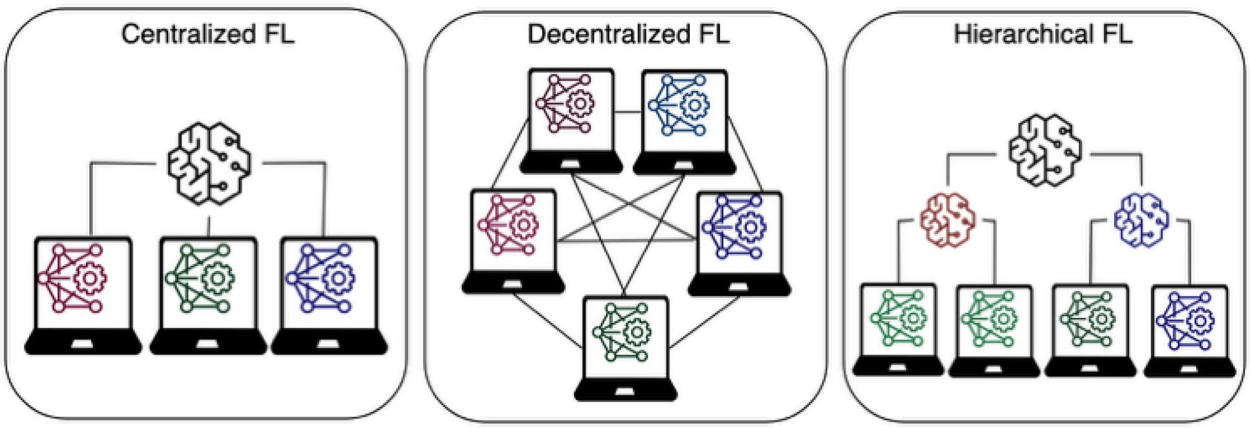

Fig. 3 Architectural paradigms within the system architecture axis. This visual comparison of centralized, decentralized, and hierarchical models highlights the structural trade-offs regarding scalability and resilience

Table 1 Comparative impact of FL architectures on CPS security

Architecture	Defense against single-point-of-failure	Resilience to Byzantine/collusive attacks	Communication efficiency
Centralized	Low (Server is a bottleneck/target)	Moderate (Requires server-side filtering)	Lower (High long-range traffic)
Decentralized	High (No central target)	Low (Vulnerable without consensus)	Moderate (High synchronization)
Hierarchical	Moderate (Isolated clusters)	High (Multi-tier validation)	High (Reduced long-range traffic)

The principal advantage of CFL is its inherent privacy preservation by keeping raw data localized. This privacy-preserving design makes it an effective foundational benchmark against which novel systems are evaluated (Friha et al. 2023, 2022; Ma and Su 2025; Javeed et al. 2024; Xia et al. 2025; Makris et al. 2025). Further surveys have reinforced this comparative analysis, statistically demonstrating FL's superior data utilization and performance against purely centralized models in various configurations (Agrawal et al. 2022; Makris et al. 2025). Even within this paradigm, variations in optimization algorithms like FedAvg, FedAvgM, and FedAdam are actively benchmarked to find the most effective approach (Praharaj et al. 2025). This model also accommodates hybrid implementations; for instance, local models like autoencoders can be trained on edge data while final classifiers are trained on public data in the cloud (Su and Zhang 2025). While foundational, the reliance on a single orchestrator introduces a potential single point of failure and can create a communication bottleneck (Agrawal et al. 2022), prompting the development of alternative architectures.

Decentralized and blockchain-enabled P2P frameworks

The decentralized, or Peer-to-Peer, architecture also referred to as a 'distributed' deployment (Khacha et al. 2024) obviates the need for a central aggregation server. In this model, clients engage in direct P2P communication to collaboratively exchange and aggregate model parameters, with consensus typically facilitated through communication with neighboring devices (Hamouda et al. 2023). This structure inherently mitigates the single point of failure risk present in CFL and can enhance system resilience (Ali et al. 2024).

Several novel IDSs have been proposed based on this paradigm. For instance, the authors of Friha et al. (2023) designed a fully decentralized FL system where clients collaboratively exchange models without central coordination. Similarly, models leverage a completely decentralized P2P mode for anomaly detection (Belenguer et al. 2025). This approach is not limited to IDS; in Ma and Su (2025), authors implemented a decentralized FL scheme for DDoS defense where autonomous systems communicate via East–West interface protocols. The PPSS framework also implements and evaluates a Decentralized FL (DFL) mode, demonstrating its effectiveness in Non-IID settings (Hamouda et al. 2023). This distributed configuration allows individual nodes to monitor their local context and neighbors while sharing insights globally, a concept explored in various works on distributed intrusion detection (Carvalho Bertoli et al. 2023; Taslimasa et al. 2023; Agrawal et al. 2022).

The resilience of the decentralized paradigm is significantly augmented by the integration of blockchain technology, which transforms a standard peer-to-peer network into a cryptographically verifiable and auditable framework (Yang et al. 2023c; Saha et al. 2024). In these blockchain-enabled decentralized FL systems, the traditional central aggregator is replaced by a distributed ledger that records model update transactions and facilitates consensus among clients. This integration is critical for addressing the trust deficit inherent in peer-to-peer environments; by using the blockchain as an immutable record, the system can enforce non-repudiation and ensure the integrity of the global model throughout its lifecycle. For instance, the Privacy-Preserving and Secure Framework (PPSF) leverages blockchain-enabled federated deep learning to secure Industrial IoT environments, demonstrating that a decentralized trust mechanism can effectively manage the coordination of updates while preventing unauthorized tampering (Mohamed 2025; Li et al. 2024). Furthermore, such architectures allow for the implementation of novel consensus protocols, such as Proof of Federated Deep Learning (PoFDL), which incentivize honest participation and provide a robust defense against the Byzantine threats and Collusive attacks prevalent in server-less CPS networks (Bhardwaj and Sumangali 2025).

Multi-tier and hierarchical edge architectures

Hierarchical Federated Learning presents a hybrid, multi-layered architecture that combines elements of both centralized and decentralized models to enhance scalability and efficiency (Mughal et al. 2024). In a typical HFL structure, model aggregation occurs at multiple tiers. Local devices on the edge perform an initial round of training and send their updates to an intermediate aggregation layer, such as an edge server or a regional data center. These intermediate servers perform a regional aggregation before forwarding the consolidated parameters to a top-level cloud server for final global model synthesis (Nguyen, et al. 2025).

This tiered structure is particularly effective for large, complex networks and offers several advantages. It can significantly improve communication efficiency and is robust to the challenges posed by Non-IID data, as statistical variations can be partially absorbed at lower aggregation tiers (Hamouda et al. 2023). Several advanced frameworks utilize this model, as discussed in broad surveys of the field (Zhang et al. 2025). For example, FedDEL is a chained, semi-asynchronous HFL framework with clients, edge servers, and a cloud server (Xia et al. 2025). Similarly, in Zhang et al. (2025) named a hierarchical framework for the Internet of Medical Things (IoMT) using "dew servers" as an edge aggregation layer.

In (Huong et al. 2021), a Federated Deep Reinforcement Learning Empowered Anomaly Detection (FLAD) was analyzed, a three-level hierarchy with Global, Regional, and Local detection centers. Another example, HFed-IDS, integrates this architecture with 5G technology in smart grids (Kheddar 2025). This hierarchical concept is also discussed as a method for managing Non-IID data by grouping clients into clusters, each with its own local aggregator that then communicates with a global server (Agrawal et al. 2022).

Anomaly detection methodologies for the CPS edge

This axis of our taxonomy provides a deep analysis of the specific anomaly detection algorithms employed *within* the FL frameworks surveyed. The choice of a particular algorithmic engine is not arbitrary; it is fundamentally dictated by the target environment's specific challenges, such as the temporal nature of CPS data, edge computational constraints, and the critical need to defend the learning process itself. For this reason, our analysis reveals a distinct evolution from applying standard detection models as local clients in FL architecture to engineering specialized techniques that secure the integrity of the federated aggregation process. The surveyed methods can be categorized into three primary groups: core deep learning engines, robust and secure aggregation techniques, and integrated privacy-enhancing frameworks.

Core deep learning detection engines

The foundational approach to FL-based anomaly detection involves deploying sophisticated deep learning models on individual clients. A dominant trend within this category is the application of recurrent and hybrid architectures to model the temporal dynamics inherent in CPS data. This preference stems from the need to extract both spatial and sequential features from network traffic and sensor readings. This is exemplified in works that combine Convolutional Neural Networks (CNNs) for feature extraction with Gated Recurrent Units (GRUs) or Bidirectional LSTMs (BiLSTMs) to capture time-dependent patterns, as seen in Li et al. (2021). This hybrid strategy is further refined in works that propose a unified Convolutional Recurrent Neural Network (CRNN), as in Selvam et al. (2025); (Husnoo, et al. 2023; Hossain et al. 2021), or combine GRUs with classical ensembles like Random Forest, as demonstrated in systems for vehicular sensor networks.

In unsupervised FL contexts where labeled attack data is scarce -a common scenario in CPS- reconstruction-based models offer a powerful paradigm. Frameworks like FeDiSa leverages Deep Auto-Encoders (DAEs) on client devices. By training the model to accurately reconstruct benign operational data, any input that result in a high reconstruction error is flagged as a potential anomaly (Husnoo, et al. 2023). This approach is extended in other systems that employ outlier-aware Autoencoder-Multi-Layer Perceptrons (AE-MLP) for Distributed Denial of Service (DDoS) defense or utilize Stacked Sparse Autoencoders (SSAE) for dimensionality reduction at the edge, demonstrating the versatility of reconstruction techniques.

To address the prevalent issues of data imbalance and privacy, generative models, particularly Generative Adversarial Networks (GANs), are increasingly employed. Rather than relying solely on real data, frameworks like (Selvaraj et al. 2024) and other privacy-preserving IDS use GANs at the client level to synthesize high-fidelity data. This augmented data balances the training set, improving the classifier's robustness against rare attack classes while simultaneously enhancing privacy.

Techniques for robust and secure aggregation

While effective, the core detection engines are vulnerable to attacks targeting the FL process. This has spurred the development of techniques focused on robust and secure aggregation, where the "anomaly" is a malicious client or a poisoned model update.

A key strategy is adversarial mitigation through model scrutiny at the aggregation server. Instead of blindly averaging model updates, these systems inspect incoming contributions. The framework proposed in Chen et al. (2021) implements an unsupervised clustering method to group model weights, allowing the server to identify and isolate malicious updates. Similarly, the work in Shen et al. (2024) uses knowledge distillation as a sophisticated fusion mechanism to refine the global model and mitigate the impact of heterogeneous client data.

The architecture of the FL system itself is also leveraged as a defensive technique, particularly through hierarchical and clustered topologies. HFL frameworks, such as those in Althunayyan et al. (2024) and (Xia et al. 2025), introduce intermediate aggregation layers. This structure improves scalability and resilience by containing threats at a regional level before they can impact the global model.

A further sign of the field's maturation is the integration of Explainable AI (XAI) for trustworthy aggregation. Recognizing that "black box" models are insufficient for critical systems, frameworks integrate XAI for trustworthy aggregation. For instance, the system in Aflaki et al. (2024) integrates LIME and Blockchain, while the framework in Namakshenas et al. (2024) uses Shapley Values (SV). These techniques provide insight into why a model makes a certain decision, which can be used to validate

the contributions of clients and build trust in the aggregated global model.

Integrated privacy and detection frameworks

The most advanced systems treat privacy not as an add-on, but as a fundamental component of the detection architecture itself. In these frameworks, the lines between the privacy mechanism and the security technique are blurred.

Cryptography-empowered learning is a cornerstone of this approach. Systems in Poorazad et al. (2024) and (Alabdulatif 2025) utilize Homomorphic Encryption (HE) to allow the server to perform aggregation directly on encrypted model parameters. This ensures that the server, or any eavesdropper, can learn nothing about a client's update. To provide formal mathematical guarantees of privacy, Differential Privacy (DP) is widely integrated. Frameworks designed for decentralized environments, such as the one in Alabdulatif (2025), achieve DP by injecting carefully calibrated noise into gradients before transmission. While providing strong privacy, this introduces a critical trade-off between privacy and model accuracy. This occurs because the statistical noise added to protect privacy can obscure genuine patterns within the model updates, potentially reducing the global model's ability to learn effectively. Balancing this is a central challenge these papers seek to resolve.

To overcome the limitations of single-party encryption, the integration of Secure Multi-Party Computation (MPC) has emerged as a vital mechanism for ensuring privacy-preserving model aggregation (Li et al. 2024; He et al. 2025). MPC enables a cohort of clients to collaboratively compute an aggregation function, such as the weighted average of model gradients, over their private inputs without any party—including the central aggregator—revealing their individual updates. While MPC provides strong confidentiality guarantees by distributing the computation across multiple nodes, its deployment in real-time CPS environments is often constrained by high communication and synchronization overheads (Li et al. 2024).

Addressing the need for integrity alongside confidentiality, researchers are increasingly exploring Zero-Knowledge Proofs (ZKP) and Functional Encryption (FE) as specialized primitives. ZKPs allow clients to provide a "proof-of-training," enabling the aggregator to verify that a model update was generated following the prescribed protocol and on legitimate data without exposing the actual weights or sensitive local samples (Nguyen, et al. 2025). This is particularly effective for defending against Byzantine and poisoning attacks, as it provides a verifiable trail of computation. Simultaneously, Functional Encryption (specifically Inner Product Functional Encryption) offers a more computationally efficient alternative to general-purpose HE for the specific mathematical tasks of federated aggregation. By allowing the server to derive only the required inner product of weight vectors without fully decrypting individual updates, FE achieves a significant reduction in the computational and energy burdens typical of cryptographic methods in edge-native CPS devices (Shanmugarasa et al. 2023).

Finally, to ensure the integrity and auditable transparency of the entire federated process, several advanced frameworks employ decentralized trust mechanisms using blockchain. The technical mechanism involves using the blockchain as a decentralized, immutable ledger where all federated transactions -such as client registrations, model update submissions, and the final aggregated global model- are recorded as cryptographically signed transactions. This provides a robust defense against repudiation attacks and enhances the integrity and auditability of the FL process by making it computationally infeasible to alter the historical record. Authors of Aflaki et al. (2024) use a distributed ledger to create an immutable record of all model updates, representing a significant step towards fully trustworthy, decentralized anomaly detection.

Despite the transparency benefits of XAI-driven aggregation, significant technical bottlenecks and adversarial failure modes persist. A primary limitation is the computational overhead associated with calculating feature attributions (e.g., SHAP or LIME) for high-dimensional model updates in real-time; such latency can be prohibitive for high-velocity CPS actuation loops. A critical failure mode involves 'Explanation Manipulation' attacks, wherein sophisticated adversaries craft malicious updates that mimic benign explanation patterns, effectively bypassing interpretability-based filters. This effectively "explains away" the poisoning attempt, allowing malicious gradients to bypass interpretability-based filters while remaining below the threshold of statistical outlier detection. Consequently, ensuring the integrity of the explanation itself remains an open technical challenge in adversarial federated environments.

To provide a consolidated overview of the approaches discussed, Table 2 synthesizes the core anomaly detection techniques, highlighting their principal advantages, inherent trade-offs, and representative examples from the literature. This synthesis underscores the critical balance that researchers must strike between detection accuracy, computational overhead, privacy guarantees, and resilience against adversarial attacks. The trend toward hybrid systems that combine multiple techniques -such as using a robust deep learning detector at the client and a secure aggregation mechanism at the server- demonstrates the

Table 2 Summary of anomaly detection techniques in federated learning for CPS

Taxonomic category	Core technique	Key advantages	Potential disadvantages/trade-offs	Representative papers
Core Detection Engines	Recurrent/Hybrid (RNN, LSTM, GRU, CNN)	Excellent for modeling temporal CPS data; captures complex sequential patterns	Can be computationally intensive for resource-constrained edge devices	Li et al. (2021), Selvam et al. (2025), Alazab et al. (2025)
	Reconstruction-Based (AE, DAE)	Unsupervised (no labeled attack data needed); effective for defining normalcy	May fail to detect novel anomalies that do not produce high reconstruction errors	Duy et al. (2024), Husnoo, et al. (2023)
	Generative Models (GANs)	Addresses data imbalance by synthesizing minority class data enhances privacy	Training is complex and can be unstable (e.g., mode collapse)	Duy et al. (2024),Selvaraj et al. (2024), Wen et al. (2025)
Robust Aggregation	Adversarial Mitigation (Clustering, Distillation)	Directly secures the FL process against poisoning and backdoor attacks	Adds computational and communication overhead to the aggregation step	Chen et al. (2021), Shen et al. (2024), Hamad et al. (2025)
	Hierarchical/Clustered Topologies	Improves scalability and resilience; contains threats at a regional level	Increase architectural complexity and management overhead	Xia et al. (2025), Althunayyan et al. (2024), Orabi et al. (2025)
	Automated Model Design	Flexible and adaptable to diverse scenarios; reduces manual design effort	Can be computationally expensive; may lead to overly complex models	Hayawi et al. (2025), Zhang and Su (2025),
	Adaptive/Personalized Learning	Adapts to data heterogeneity (Non-IID) and evolving threats	May lead to model divergence if personalization is too strong	Su and Zhang (2025), Kumar and Khari (2025)
	Explainable AI (XAI)	Provides transparency and trust in model decisions; aids in validating client updates	High computational cost for generating explanations, especially for complex models	Namakshenas et al. (2024)
	Reinforcement Learning (RL)	Enables dynamic, adaptive defense strategies against evolving threats	Requires careful design of reward functions and simulation environments	Hossain et al. (2021), Gupta, et al. (2023)
Integrated Frameworks	Cryptography (Homomorphic Encryption)	Provides strong confidentiality guarantees for model updates during aggregation	Significant computational overhead, potentially slowing down the learning process	Namakshenas et al. (2024), Poorazad et al. (2024), Alabdulatif (2025)
	Differential Privacy (DP)	Offers formal, mathematical guarantees of privacy against inference attacks	Involves a direct trade-off between the level of privacy and model accuracy	Hamouda et al. 2023) Hamouda et al. (2023), Hossain et al. (2021)
	Blockchain	Ensures integrity, non-repudiation, and an auditable trail of the FL process	Potential scalability limitations and high energy consumption	Hamouda et al. (2023), Aflaki et al. (2024), Hayawi et al. (2025)

field's movement toward holistic, multi-layered defense strategies.

Domain-specific threat models and use cases

The efficacy of federated anomaly detection is inextricably linked to the specific operational constraints and adversarial landscape of the target environment. While the preceding sections examined the architectural and methodological foundations of FL, this section shifts the focus toward the Domain-Specific Threat Model, analyzing how unique industrial requirements dictate the selection of security archetypes. To provide a rigorous foundation for this analysis, it is essential to first deconstruct the multifaceted threat surface that characterizes mission-critical CPS. By mapping the lifecycle of federated training against the specific vulnerabilities of the physical layer, we can better understand how generic adversarial vectors manifest as catastrophic operational risks.

Taxonomic deconstruction of the CPS adversarial landscape

The adversarial landscape in federated CPS is defined by a sophisticated spectrum of threats that target the integrity, confidentiality, and availability of both the digital model and the physical process (Blika et al. 2025; Khan et al. 2023). At the inception of the federated lifecycle, Integrity-Based Threats, such as data and model poisoning, represent a primary concern. In these scenarios, compromised edge devices inject malicious gradients or manipulated datasets to subtly influence the global model's parameters. Unlike traditional IT poisoning, these attacks in a CPS context are frequently designed as "stealthy" injections that remain dormant until specific physical triggers are met, such as reaching a critical pressure or voltage threshold. This latent instability poses a severe risk to safety–critical operations, where a false negative during a genuine anomaly event can lead to irreparable physical damage (Chiriac et al. 2025).

Parallel to integrity concerns, Privacy-Based Threats leverage the shared nature of model updates to facilitate industrial espionage. Through techniques such as Membership Inference Attacks (MIA) or Gradient Leakage, an adversary can reverse-engineer sensitive operational patterns or proprietary industrial logic from the very updates intended to secure the system (Mia and Hadi Amini 2024). This risk is compounded by Operational Availability Threats, which manifest as either Byzantine behaviors or coordinated collusive attacks. Byzantine faults—whether stemming from hardware failure or malicious intent—aim to prevent the global model from converging, effectively blinding the system's monitoring capabilities. Furthermore, during the post-deployment inference phase, Adversarial Evasion techniques allow attackers to bypass real-time detectors by crafting subtle perturbations in sensor data. This ensures that unauthorized physical actuation remains undetected by the global model, thereby bridging the gap between cyber-manipulation and physical disruption (Nkoom et al. 2025b; Sharma et al. 2025).

By analyzing these multifaceted dimensions, the extant literature reveals that the selection of an FL system fundamentally dictates how learning is coordinated and how the system defends against such structural vulnerabilities (Khan et al. 2023; Shao et al. 2024). Building upon this general threat framework, the subsequent analysis examines how these vulnerabilities manifest across specific industrial domains, ranging from the high-latency requirements of vehicular networks to the extreme data sensitivity of healthcare environments. Each model presents distinct trade-offs regarding efficiency, robustness, and communication overhead, which are explored in detail through the following use cases (Shao et al. 2024).

To synthesize the theoretical framework established above, Table 3 provides a consolidated view of the security landscape, mapping specific adversarial vectors to their operational objectives and physical impacts.

Domain-specific manifestations and analysis

This axis of our taxonomy provides a technical analysis of the surveyed literature based on the specific CPS domain and its corresponding threat model. The design of an FL-based anomaly detection system is not generic; it is intrinsically shaped by the operational context and the unique vulnerabilities of its application environment. Unlike standard IT systems where the primary consequence of a breach is data loss, in CPS, a cyber anomaly

Table 3 Summary of security threats and adversarial vectors in federated CPS

Attack category	Specific threat	Adversarial objective	Physical impact on CPS
Poisoning	Backdoor/Model Tampering	Inject hidden triggers into the detector	Latent instability; false negatives during safety events
Inference	Gradient Leakage/MIA	Reverse-engineer industrial states	Exposure of proprietary operational logic; espionage
Availability	Byzantine and Collusive Attacks	Disrupt model convergence or take control	Loss of monitoring capability; synchronized failures
Evasion	Adversarial Perturbations	Bypass real-time detection logic	Undetected physical intrusion; unauthorized actuation

can precipitate catastrophic physical events. Therefore, our analysis will not only categorize systems by their application domain but will also deconstruct the specific adversarial goals, the attack vectors, and the technical rationale for selecting a particular FL-based defense.

While CI traditionally encompasses sectors like energy and water, this survey adopts a broader definition to include emerging, large-scale interconnected systems whose failure can precipitate significant societal or economic disruption. This analysis therefore includes not only stationary CI like smart grids and ICS, but also dynamic and mobile systems. For instance, vehicular networks (IoV, V2X, UAVs) are included as a critical category. While a single vehicle is not CI, a coordinated attack on a fleet of autonomous vehicles or a city's traffic management system could paralyze transportation, constituting a critical failure. Their high mobility and real-time safety requirements create a unique and highly dynamic attack surface where a cyber-physical anomaly directly translates to physical risk. Similarly, industrial robotics are included because the disruption of large-scale robotic fleets in logistics or manufacturing can cripple key supply chains, representing a significant economic threat. The healthcare sector (IoMT) is also analyzed as a CI domain where a cyberattack can directly impact patient safety. By categorizing the literature across these domains, we can reveal how distinct threat models drive the development of tailored FL security frameworks.

The domain of vehicular networks presents a unique threat model where the adversary's goal extends beyond data theft to include inducing kinetic impact. An attack is not merely a data integrity issue but a direct threat to physical safety. For example, the threat vector of a data poisoning attack against an object detection model, as addressed in Hamad et al. (2025), could cause a vehicle to misinterpret its environment (e.g., mistake a pedestrian for a harmless object), leading to a collision. This direct link between model integrity and physical outcome is a defining characteristic of this domain's threat model.

To counter the threat of inference attacks on sensitive location and trajectory data, the PPFedSL framework combines Split Learning with Federated Learning (Soares et al. 2025). This hybrid architecture is specifically chosen because it provides a stronger defense against an adversary attempting to reconstruct a vehicle's travel history from its shared model updates. In this model, the neural network is literally split into a client-side portion and a server-side portion. The client-side model, which processes raw, sensitive data directly, remains on the vehicle. Only its intermediate output -the "smashed data"- is sent to the server for the completion of the forward pass. By combining this with federated learning for the aggregation step, the central server never has access to either the raw data or the full model gradients, adding a critical layer of obfuscation that makes successful reconstruction attacks significantly more difficult than in a standard FL implementation.

A framework for UAVs in maritime operations uses FL to specifically counter threats of data confidentiality and leakage during the data collection phase. In this threat model, the raw data collected by a UAV -which could be high-resolution video for surveillance or sensor readings for environmental monitoring- is the primary asset an adversary seeks. Transmitting this data to a central server creates two major risks: first, an eavesdropping attack could intercept the data stream in transit, and second, a breach of the central server would create a mass data leakage event. The FL approach mitigates these threats by ensuring the raw, high-dimensional data never leaves the UAV. The UAV, acting as an edge device, uses its onboard computational resources to process the data and train a local machine learning model (e.g., for path optimization). Instead of transmitting the raw data, it only sends the resulting low-dimensional model parameters (gradients or weights) to the central server for aggregation. These parameters represent the abstract "learnings" from the data, not the data itself, making it computationally infeasible for an adversary to reconstruct the original sensitive information from the intercepted updates (Min et al. 2025).

The threat model in this domain extends beyond direct cyberattacks to include safety–critical operational anomalies. For instance, the work in Peng et al. (2023) uses FL not just to protect driver privacy but also to enhance uncertainty-awareness in trajectory predictions. Here, a deviation from a predicted trajectory with high certainty is itself an anomaly that could signal an imminent safety risk, a concern non-existent in traditional IT systems. Other works, such as the one presented in Hamad et al. (2025), focus on safeguarding the FL process itself against data poisoning and model replacement attacks, which are critical threats in a domain where manipulated object detection could have catastrophic physical consequences. The use of hierarchical FL in works is specifically tailored to the distributed nature of in-vehicle networks (Althunayyan et al. 2024). This architecture not only improves scalability but also strategically reduces the system's attack surface. By introducing intermediate aggregators, it prevents a single point of failure at the central server and makes it more difficult for an adversary to influence the global model by compromising a small number of clients.

In industrial and robotic environments, the threat model also expands beyond data privacy to include attacks that could disrupt physical processes, compromise intellectual property, or cause physical harm. The

Internet of Robotic Things (IoRT) is particularly vulnerable to attacks like DDoS, which can paralyze entire fleets of autonomous robots. The framework designed by the authors of Orabi et al. (2025) aims to detect and mitigate DDoS attacks in IoRT environments by using a combination of FL and differential privacy clustering.

A critical concern in this domain is the threat of model-based attacks, where an adversary attempts to reverse-engineer a trained model to infer sensitive information about the system's operations. A study focusing on electric robot charging infrastructure directly confronts the threat of Model Inversion (MI) attacks (Nkoom et al. 2025b). In this context, an MI attack could reveal proprietary charging patterns or robot operational schedules, which could be exploited for industrial espionage or physical disruption. The proposed solution uses Federated Transfer Learning to build a robust detection system that is resilient to such inference attacks.

Unlike the vehicular domain where the primary adversarial goal is often inducing immediate kinetic impact on a single or small group of targets, the threat model for stationary CI frequently involves adversaries seeking to cause widespread, cascading service disruption across a large geographic area. For example, in a smart grid, an adversary's goal could be to trigger a cascading blackout. To achieve this, they might use a False Data Injection (FDI) attack that mimics the signature of a natural grid fault. This makes the defense proposed in the FeDiSa framework particularly relevant, as it is explicitly designed for this purpose (Husnoo, et al. 2023). By using a Deep Auto-Encoder within an FL framework, the system learns a precise model of normal power grid behavior, enabling it to effectively discriminate between natural disturbances and adversarial intrusions. Similarly, systems like DeepFed (Li et al. 2021), BFL (Poorazad et al. 2024), and 2DF-IDS (Friha et al. 2023) are tailored to detect a range of threats in industrial settings.

The threat model for Advanced Metering Infrastructure (AMI) is particularly complex, including not only DoS, Probe, and User to Root (U2R) attacks but also FL-specific vulnerabilities like gradient leakage, which is especially potent in FL scenarios with few participants or when high-resolution model updates are shared (Xia et al. 2025). An adversary exploiting this vector could infer the energy consumption patterns of a specific facility, which could be used to identify CIs for a targeted physical attack or to plan load-altering attacks on the grid. The broader goal of fighting cyber threats in Critical Energy Infrastructure (CEI) is a key motivation for adopting FL, as highlighted in Razzak et al. (2022).

In the healthcare domain, the consequences of a cyberattack can be life-threatening. The threat model extends beyond patient data privacy to include direct physical harm through malicious command injection. For example, an attacker could tamper with the functionality of an insulin pump or alter the rhythm of a pacemaker. This makes the security of Wireless Body Area Networks (WBANs), as discussed in Razzak et al. (2022), a safety–critical concern. To defend against not only privacy breaches but also model poisoning attacks that could corrupt a diagnostic model, some frameworks employ GANs, as detailed in Selvaraj et al. (2024). In this threat model, a poisoned diagnostic model that, for instance, learns to misclassify a malignant tumor as benign, represents a direct threat to patient outcomes. The use of GANs to create a balanced and high-fidelity synthetic dataset is a defense aimed at improving the robustness and reliability of the final diagnostic model against such manipulations.

Other frameworks focus on detecting anomalous physiological signals that could indicate either a medical emergency or device tampering. The review in Selvam et al. (2025), for instance, highlights two distinct approaches. The first system performs adaptive anomaly detection on ECG signals using Transformer-based Autoencoders at the edge. This technique is particularly well-suited for identifying subtle deviations in quasi-periodic signals that could represent either cardiac arrhythmia or malicious manipulation of sensor data. The system's adaptability is further enhanced by using Support Vector Data Description (SVDD) to account for changes in a patient's specific data distribution over time.

The second system discussed, the FIDChain model, addresses the need for integrity and auditability by integrating blockchain to create a secure and transparent ledger for all model updates within the IoMT network. Collectively, these approaches demonstrate a sophisticated understanding of the IoMT threat landscape, where anomaly detection must protect both patient data and patient safety.

Privacy-preserving mechanisms and cryptographic defenses

This axis of our taxonomy provides a deep technical analysis of the spectrum of privacy-preservation techniques employed within the surveyed FL frameworks. While the foundational principle of FL-training on local data without centralizing it- provides an inherent layer of privacy, it is not a complete safeguard against all privacy threats. The exchange of model parameters creates a new attack surface, exposing the system to sophisticated inference attacks that can potentially reconstruct sensitive information from the shared updates. Consequently, a significant body of research, as highlighted in comprehensive surveys like (Makris et al. 2025) and (Dhanushkodi et al. 2024) focuses on augmenting the

Table 4 A comparative analysis of privacy-preservation techniques in federated learning for CPS

Technique category	Specific method	Primary defense against	Computational overhead	Key trade-off/disadvantage	Representative papers
Architectural Baseline	Inherent FL Design	Mass data leakage from a central server; direct exposure of raw data	Low	Vulnerable to sophisticated inference attacks on model updates	Nkoom et al. (2025b), Husnoo, et al. (2023), Razzak et al. (2022) (Dhanushkodi et al. (2024)
Cryptographic Methods	Homomorphic Encryption (HE) and Secure Multi-Party Computation (SMPC)	Eavesdropping attacks; inference attempts by the aggregation server	High	Significant performance impact; can introduce high latency, slowing down the learning process	Taslimasa et al. (2023), Hamouda et al. (2023), Shen et al. (2024), Kamatchi and Uma (2025)
Perturbation Methods	Differential Privacy (DP)	Inference attacks (e.g., membership inference) by making individual contributions statistically undetectable	Medium	The direct "privacy-utility trade-off": stronger privacy (more noise) can reduce model accuracy	Soares et al. (2025), Vemulapalli and Sekhar (2025), Yang et al. (2025), kumar sah et al. (2025)
Architectural Defenses	Split Learning and Data Obfuscation	Targeted inference attacks like Membership Inference Attacks (MIA)	Medium–Low	Defenses are often tailored to specific attack vectors and may lack general applicability	Kumar and Khari (2025), Shao et al. (2024)
Decentralized Trust	Blockchain	Tampering with model updates; repudiation attacks; lack of auditability in the FL process	High	Potential scalability limitations and high latency/energy consumption due to consensus mechanisms	Xia et al. (2025), Chen et al. (2021), Xiao et al. (2025)

basic FL paradigm with advanced Privacy-Enhancing Technologies (PETs). This analysis categorizes the surveyed literature based on the primary privacy mechanism employed, moving from the baseline privacy of the FL architecture itself to more robust cryptographic methods, perturbation-based techniques, and architectural defenses, the key characteristics of which are compared in Table 4.

The most fundamental level of privacy preservation is the adoption of the FL architecture itself. The privacy benefit is achieved by design: raw data remains within the client's administrative domain, and only abstract model parameters are transmitted. This approach, leveraged by a vast number of works such as (Althunayyan et al. 2024; AbuElHassan et al. 2025) and (Hossain et al. 2025) inherently mitigates the risk of mass data leakage from a central server. It is computationally efficient compared to more advanced PETs and is a core feature in frameworks for diverse domains including industrial monitoring (Wen et al. 2025), hypervisor security (Kumar and Khari 2025), and intelligent vehicles. However, this baseline privacy is vulnerable to sophisticated inference attacks. As noted in surveys like (Yang et al. 2023c), an adversary with access to the shared gradients can potentially reconstruct parts of the private training data. Therefore, while suitable for low-risk environments, this inherent privacy is often considered insufficient for highly sensitive CPS applications, necessitating the more advanced techniques discussed below.

To provide stronger, more formal privacy guarantees against specific threats, many frameworks integrate one or more advanced PETs. As systematically reviewed in Ali et al. 2025b and (Shenoy et al. 2025), these technologies are not mutually exclusive and are often combined to create a multi-layered defense. To defend against eavesdropping and aggregation servers that, while following the protocol, may attempt to infer information from model updates, frameworks employ cryptographic techniques that allow for computation on protected data.

HE allows the aggregation server to perform mathematical operations directly on encrypted model parameters, while Secure Multi-Party Computation (SMPC) enables a group of clients to jointly compute a function without revealing their individual inputs.

These methods provide strong, provable guarantees of confidentiality. However, their primary disadvantage is high computational and communication overhead. Consequently, many frameworks, such as (Li et al. 2021) and (Namakshenas et al. Jul. 2024) apply more efficient variants like Additively Homomorphic Encryption (AHE). The use of the Paillier cryptosystem (a form of AHE) is noted by authors in Vemulapalli and Sekhar (2025) and in specific applications for smart grids (Yang et al. 2025). SMPC, as used in Xia et al. (2025), often requires multiple communication rounds, which can introduce latency.

DP offers a formal, mathematical guarantee of privacy by making the participation of any single client statistically undetectable. As mentioned before, this is typically achieved by injecting calibrated statistical noise into the model updates before they are shared. DP provides a strong, quantifiable guarantee against a wide range of inference attacks and is often less computationally intensive than cryptographic methods. However, the core challenge of DP is the inherent privacy-utility trade-off. As highlighted in kumar sah et al. (2025), adding more noise for better privacy inevitably degrades the quality of the model updates, which can harm the accuracy of the final global model. This trade-off is critical in CPS, where a loss of accuracy could have safety implications. Advanced implementations, such as the Contextual Anonymization Aware Differential Privacy (CtxADP) in Shao et al. (2024), attempt to mitigate this by dynamically adjusting the privacy noise. DP is widely used in frameworks like (Nkoom et al. 2025b; Xiao et al. 2025), and (Alhazmi et al. 2025) where strong privacy guarantees are required.

GAN-based privacy enhancement is frequently proposed as a robust mechanism for generating synthetic training data, yet it faces the dual challenges of training instability and mode collapse (Hossain et al. 2025). In Non-IID (heterogeneous) CPS environments, GANs often fail to represent the full diversity of edge data, leading to a biased global model. Moreover, there exists a fundamental Privacy-Fidelity Trade-off: if the generator is overly regularized to satisfy strict privacy guarantees (e.g., via Differential Privacy) (Rahdari et al. 2025), the resulting synthetic samples may lose the high-resolution signal patterns necessary for accurate anomaly detection. This failure to capture the "long tail" of anomaly distributions can lead to dangerous false negatives in mission-critical monitoring.

Beyond general-purpose PETs, some of the most advanced research focuses on designing the FL architecture itself to be resilient against specific, targeted threats, most notably Membership Inference Attacks (MIA). An MIA adversary with access to the trained model, attempts to determine if a specific data point was part of a particular client's training set -a significant privacy breach in domains like healthcare. These architectural defenses modify the training process or model structure to break the statistical links that MIA attacks exploit. This approach can be highly effective against a known threat vector and may have lower computational overhead than general—purpose PETs. Its primary disadvantage, however, is a lack of generality, as a defense tailored for MIA may not protect against

other inference attacks. The PPFL-DCS framework (Mehta et al. 2025) provides a prime example, integrating a Neural Transformer System (NTS) that trains the model on "mixup" data -synthetic examples created by interpolating between real data points. This process inherently obfuscates the properties of individual data records, making it significantly harder for an MIA model to succeed. Other architectural choices, such as the Split Learning used in Mehta et al. (2025), also function as a defense by splitting the model itself and only sharing intermediate activations, thus shielding the raw data from direct inference.

While encryption and DP focus on the confidentiality of model updates, another class of threats targets the integrity and auditability of the FL process. To counter these, several frameworks integrate blockchain technology. The technical mechanism involves using the blockchain as a decentralized, immutable ledger where all federated transactions—such as client registrations, model update submissions, and the final aggregated global models- are recorded as cryptographically signed transactions. This approach fundamentally shifts the trust model from a centralized, potentially fallible server to a distributed, verifiable consensus.

This integration provides a robust defense against several critical threat vectors. First, it directly mitigates repudiation attacks, where a malicious client might submit a poisoned update and later deny having done so; the immutable record on the blockchain serves as undeniable proof of the transaction. Second, it enhances resilience against tampering, as altering any historical record would require compromising the entire distributed ledger, a task considered computationally infeasible. This is a key motivation in frameworks like those in, which recommend this approach for securing smart electric vehicles (Bhardwaj and Sumangali 2025).

Furthermore, blockchain enables novel mechanisms for secure aggregation and validation. In (Aflaki et al. 2024), authors use the blockchain to enhance the transparency and security of the global model update process. Other advanced systems, such as the (Hamouda et al. 2023), introduce novel consensus protocols like Proof of Federated Deep Learning (PoFDL) to validate and add new model-containing blocks to the chain. However, the advantages of blockchain are not without significant trade-offs. The primary disadvantages are potential scalability limitations and high latency introduced by the consensus mechanisms (e.g., Proof-of-Work), which can be prohibitive for real-time, large-scale CPS applications. This makes the choice of blockchain a critical design decision, best suited for environments where auditability and integrity are paramount, and some latency is acceptable.

Having analyzed the literature across the four axes of architecture, detection technique, application domain, and privacy scheme, we can now synthesize these findings into a prescriptive taxonomy. This framework, presented in Table 5, moves beyond description to offer guidance on selecting an appropriate FL Security Archetype based on the primary constraints and threat models of a given CPS domain. It serves as a decision-making tool that connects a domain's specific challenges to a holistic security solution.

Deconstructing the field: trends, trade-offs, and foundational challenges

Having systematically categorized the surveyed literature across multiple technical axes, this section transitions from a taxonomic review to a critical analysis of the overarching trends, trade-offs, and challenges that define the field. The deployment of FL for anomaly detection in CPS is not a straightforward application of a single technology but a complex balancing act between competing technical and operational requirements.

This analysis will first deconstruct the fundamental performance trilemma -the trade-off between accuracy, communication, and privacy- that governs the design of these systems. Subsequently, we identify the prevailing research trends that characterize the state-of-the-art and conclude by highlighting the significant research gaps and "blind spots" that must be addressed to advance the field toward robust, real-world deployment.

The performance trilemma: accuracy, communication, and privacy

The design of an FL system for CPS security is governed by a fundamental performance trilemma: the inherent trade-off between detection accuracy, communication overhead, and data privacy, as shown in Fig. 4. Achieving gains in one of these domains, such as strengthening privacy guarantees through cryptographic or perturbation methods, often comes at the cost of the others -for instance, by increasing communication overhead or potentially reducing model accuracy. This section deconstructs each axis of this trilemma, analyzing how state-of-the-art frameworks strategically balance these competing requirements based on the specific threat model and operational context of their target domain.

Accuracy versus privacy

The most direct trade-off exists between model accuracy and the strength of privacy guarantees. While the inherent privacy of FL provides a baseline, achieving robust protection against inference attacks often requires advanced PETs that can impact model performance. For instance, perturbation-based methods like DP achieve

Table 5 Taxonomy for selecting FL security archetypes

CPS domain	Primary threat vector	Key system constraints	Recommended FL security archetype	Justification
Vehicular Networks (IoV/V2X)	Data poisoning of object detection models; inference attacks on location/trajectory data	High mobility; ultra-low latency requirements; avoidance of single points of failure	Decentralized or Hierarchical FL with lightweight Recurrent/Hybrid detectors and Split Learning/Functional Encryption (FE) for efficient privacy	Decentralized architecture directly mitigates the single point of failure risk in dynamic VANETs (Shenoy et al. 2025). This is paired with Split Learning, which provides a robust defense against inference attacks by creating a critical layer of obfuscation that protects sensitive trajectory data, directly addressing the primary threat model for this domain
Stationary CI (Smart Grids/ICS)	False Data Injection (FDI) attacks that mimic natural faults; cascading service disruption; gradient leakage from high-resolution updates	Need for high scalability; integration with legacy systems; severe impact of failure	Hierarchical FL (HFL) using Reconstruction-based (e.g., Autoencoder) detectors and strengthened with Homomorphic Encryption (HE) and Secure Multi-Party Computation (MPC)	A hierarchical architecture is chosen to address the large-scale nature of smart grids. The use of Auto-Encoders is specifically tailored to the threat of FDI attacks, as they can learn a precise model of normal grid behavior (Husnoo, et al. 2023). Finally, Homomorphic Encryption is included to provide strong confidentiality for sensitive operational data during aggregation, a technique noted for this specific application (Vemulapalli and Sekhar 2025)
Healthcare (IoMT)	Malicious command injection (e.g., altering pacemakers); diagnostic model poisoning; breach of sensitive patient data	Extreme data sensitivity and regulatory compliance; safety–critical real-time monitoring	Centralized or Blockchain-integrated FL with Generative Models (GANs) for data augmentation and strong Differential Privacy (DP) for formal guarantees	Given the extreme data sensitivity, a Blockchain-integrated model is recommended to create a secure and transparent ledger for all model updates, ensuring auditability and integrity. The use of GANs directly counters the critical threat of diagnostic model poisoning by creating robust and balanced training data, while strong DP provides the necessary formal guarantees for protecting patient data (Selvam et al. 2025)
Industrial Robotics (IoRT)	DDoS attacks against robot fleets; Model Inversion (MI) attacks to steal proprietary operational schedules and intellectual property	High data heterogeneity (Non-IID) from diverse environments; need for both physical safety and IP protection	Clustered/Hierarchical FL with Robust Aggregation techniques (e.g., knowledge distillation) and Explainable AI (XAI) to ensure model integrity	The Clustered/Hierarchical approach is selected to specifically manage the high data heterogeneity (Non-IID) common in IoRT environments. This is paired with Robust Aggregation to mitigate the impact of heterogeneous client data during model fusion (Shen et al. 2024). The inclusion of XAI addresses the need for high trust and integrity by enabling the validation of client contributions (Namakshenas et al. 2024), which is crucial when protecting valuable intellectual property

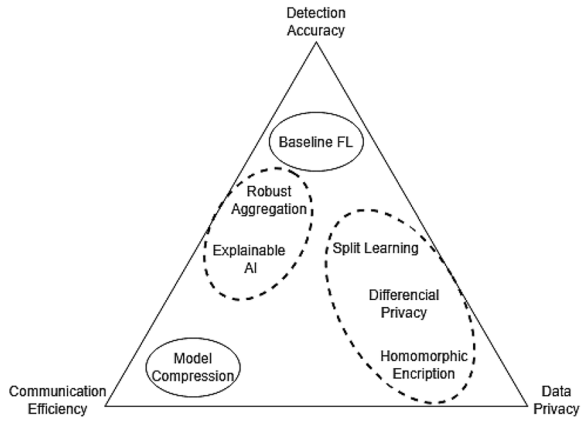

Fig. 4 The performance trilemma in federated anomaly detection

privacy by injecting calibrated noise into model updates. While this provides a formal, mathematical guarantee, the noise can degrade the quality of the aggregated global model, leading to a reduction in anomaly detection accuracy, a trade-off explicitly noted in Zhang et al. (2024) and (Hamad et al. 2025). This "Privacy-performance balance," as it is termed in Orabi et al. (2025) is particularly acute in safety–critical CPS, where even a marginal decrease in accuracy could have catastrophic consequences.

Frameworks navigate this challenge in several ways. Some, like the AnoFed approach for cardiac arrhythmia detection (Latif et al. 2025), demonstrate that high accuracy can be maintained while keeping sensitive health data local. Others, such as the FedMDO framework (You et al. 2024) explicitly aim to improve both, reporting a 10% accuracy improvement by using a redesigned DP mechanism that has less negative impact. Besides, the FedAMLrn framework for IIoT (Shao et al. 2024), attempt to find an optimal balance by developing adaptive DP mechanisms (CtxADP) that dynamically adjust the noise level based on data sensitivity. Others achieve high accuracy alongside strong privacy by accepting a controlled increase in computational cost. For example, a FL CNN Intrusion Detection System (Torre et al. 2025) achieves an accuracy of 97.31% by integrating DP, Diffie-Hellman Key Exchange, and Homomorphic Encryption, incurring only a minimal 10% increase in computation time.

Accuracy versus communication

The demand for high-accuracy anomaly detection, with some systems reporting rates as high as 99.00% (Abbas et al. 2023), often leads to the use of complex, high-parameter deep learning models. However, training these complex models on individual, often resource-constrained, IoT or edge devices can be computationally infeasible. FL addresses this by allowing devices to collaboratively train a shared model, but larger models result in larger model updates, which increases the communication overhead in each federated round. The literature presents several strategies to manage this trade-off. The framework for hypervisor security in Kumar and Khari (2025) achieves a high accuracy of 92.6% while simultaneously reducing communication overhead by 55% and training time by 32% compared to centralized methods. Specialized architectures like the DEAFL-ID for IIoT are specifically designed to be delay and energy-efficient while maintaining high detection accuracy (Latif et al. 2025).

For highly resource-constrained environments, some frameworks prioritize communication efficiency without sacrificing accuracy by modifying the FL architecture itself. The GC-FADA framework for ICS (Zhu et al. 2025) achieves up to 98.29% accuracy while reducing overall costs by avoiding complex cryptographic primitives and relying on lightweight masking techniques. Similarly, a Resource-Efficient Clustered Federated Learning Framework (Takele and Villányi 2025) minimizes communication by forming client clusters and having only resourceful cluster heads transmit model updates. Intelligent client selection, as seen in FedMint (Wehbi et al. Dec. 2023), also improves efficiency by selecting more suitable participants and avoiding stragglers.

Communication versus privacy

Stronger privacy guarantees often come at the cost of increased communication and computational overhead. Cryptographic methods like Homomorphic Encryption (HE) and technologies like blockchain, while providing robust security, are computationally intensive and can increase the size of transmitted data or add latency through consensus mechanisms. The FedAMLrn (Shao et al. 2024) framework serves as an excellent example: its use of blockchain for integrity and DP for privacy results in a low communication overhead of 5.5 MB but introduces a notable latency of 100 ms, which can increase significantly with more nodes. This highlights a clear design choice where increased latency is an accepted trade-off for achieving tamper-proof model updates and strong privacy. In contrast, solutions for dynamic environments like intelligent connected vehicles prioritize low latency. The approach in Kamatchi and Uma (2025) uses a dynamic asynchronous aggregation method to reduce training time, while the PPFedSL strategy for IoV (Soares et al. 2025) employs lightweight cryptography to minimize computational overhead in a mobile setting. Novel approaches, such as using Generative AI to create and transfer a lightweight synthetic data surrogate instead of large data instances (Hussien et al. 2025),

represent a promising direction for achieving both strong privacy and energy efficiency.

Ultimately, this analysis of the performance trilemma demonstrates a sophisticated and context-aware navigation of these competing requirements within the surveyed literature. The absence of a single, universally optimal solution underscores a key finding of this survey: the choice of techniques reflects a strategic balancing of priorities dictated by the specific operational requirements and threat model of the target CPS domain. This leads to a clear divergence in design, with frameworks prioritizing formal privacy guarantees in healthcare, low latency in vehicular networks, and computational efficiency in resource-constrained industrial settings.

Observed trends and prevailing approaches

Beyond the navigation of the performance trilemma, our systematic review of the literature reveals several dominant trends that characterize the current state-of-the-art in federated anomaly detection for CPS. These trends signify a concerted effort to build resilient, private, and efficient cybersecurity solutions for distributed and critical environments.

Pervasive shift towards decentralized learning for privacy and scalability

The most significant trend is the widespread adoption of FL as a decentralized machine learning approach for cybersecurity. This shift is a direct response to the critical challenges of traditional centralized methods, namely data privacy concerns, high communication burdens, and difficulties handling the heterogeneous (Non-IID) data common in CPS. FL is now considered a pivotal enabler for Trustworthy AI (TAI), allowing distributed network nodes to collaboratively train models without exchanging raw sensitive data (Saha et al. 2024). This privacy-preserving security model enhances real-time cyber defense by enabling collaborative threat intelligence without compromising data sovereignty, a key advantage highlighted in Mohamed (2025). This approach is widely applied to ICS and IIoT to improve security, overcome data scarcity for training, and enable privacy-preserving collaborative detection (Abdullahi and Lazarova-Molnar 2025).

Integration with edge and fog computing

A dominant trend is the convergence of FL with Edge and Fog Computing to process data closer to its source. This is vital for the operational efficiency and real-time responsiveness required in CPS. By performing training locally on edge/IoT devices, this paradigm significantly reduces the need to transmit massive datasets to a central cloud, leading to enhanced privacy, improved bandwidth efficiency, and greater scalability (Li et al. 2024). Solutions like Edge-FLGuard propose lightweight deep learning models for on-device inference and real-time anomaly detection in 5G-enabled IoT environments, explicitly addressing data heterogeneity and distributed attack surfaces (Reis 2025). This localized processing is crucial for the resource-constrained edge devices common in Industry 4.0, which face major bottlenecks in communication and computational costs when sharing model updates (Takele and Villányi 2025).

Increasing sophistication in privacy-enhancing technologies (PETs)

Beyond FL's inherent privacy-preserving nature, research actively integrates a portfolio of advanced PETs to bolster data confidentiality against sophisticated attacks. A clear trend is the combination of FL with multiple PETs to create a multi-layered defense.

- Differential Privacy (DP) is frequently combined with FL to add noise to model updates, further protecting sensitive information, as seen in frameworks for IIoT (Shao et al. 2024) and smart grids.
- Homomorphic Encryption (HE) is often integrated to allow computations to be performed directly on encrypted data, a technique used for Industrial IoMT to analyze encrypted medical records (Ghadi et al. 2025).
- Blockchain Technology is increasingly integrated with FL to provide enhanced security, transparency, and decentralization. In FL-based IDS, blockchain helps remove single points of failure and ensures trust among devices without requiring a central authority, as explored in and (Shawkat et al. 2025).

Deepening specialization in critical application domains and threat models

FL is evolving from a general-purpose technology to a set of highly tailored solutions for specific critical infrastructure sectors. Research is heavily concentrated on securing essential services, with a strong focus on:

- ICS and Smart Grids: A significant body of work focuses on developing advanced IDSs for SCADA networks and smart grids, which are highly susceptible to cyber-physical attacks (Akpolat and Kalay 2025).
- Healthcare (IoMT): FL is extensively employed for disease diagnosis and health monitoring, enabling the development of personalized models while preserving patient data privacy, as seen in applications for cardiac disease detection and medical image analysis (Makris et al. 2025).

- Vehicular Systems (ITS/CAVs): In intelligent transportation, FL is used for safety–critical tasks such as safeguarding against malicious attacks on autonomous vehicles (Hamad et al. 2025) and addressing persistent threats in Vehicle–Road Cooperation Systems (Kumar et al. 2024).

This specialization extends to the types of anomalies detected. While early work focused on general intrusion detection, a clear trend is the development of frameworks to counter specific, sophisticated attacks such as DDoS (Sakr et al. 2024), jamming and spoofing (Rehman et al. 2025) and zero-day attacks (Verma et al. 2024).

Emergence of explainable AI (XAI) for trust and accountability

With the growing deployment of complex "black box" AI models in high-stakes CPS domains, a nascent but significant trend is the integration of Explainable AI (XAI). Recognizing that transparency is crucial for trust and accountability, particularly in regulated sectors like finance and healthcare, researchers are beginning to integrate XAI methods like LIME and SHAP into FL frameworks. This aims to provide interpretability for AI-driven security decisions, a critical step for the operational adoption of these advanced systems.

Identification of research gaps and "blind spots"

Despite the significant progress and prevailing trends identified in this survey, a critical analysis of the literature reveals several research gaps and "blind spots" that are not yet adequately addressed. These omissions represent significant barriers to the practical, large-scale deployment of robust and trustworthy FL-based security solutions for CPS and highlight crucial areas for future research.

The following discussion expands on each of the foundational challenges summarized in Table 6, providing a deeper analysis of the specific limitations and blind spots in the current literature.

Data-related challenges: scarcity, quality, and real-world validation

A pervasive blind spot across the surveyed literature is the reliance on outdated or synthetic datasets that may not accurately reflect the complexity and dynamism of modern CPS environments. Many studies, as noted in Mahmoodi et al. (2023) and (Raza et al. 2024), still use datasets generated in controlled, private settings. This significantly limits the generalizability of the proposed models. The challenge is compounded by the scarcity of high-quality, labeled data, as industrial clients are often reluctant to share sensitive samples due to security concerns (Zhu et al. 2025). Furthermore, there is a notable lack of empirical validation on actual physical testbeds or large-scale, long-term deployments, a gap explicitly acknowledged in Sarker et al. (2024). While some works, such as the CaixaBank pilot (Karampasi, et al. 2024), provide valuable examples of real-world deployment, such comprehensive validations remain an exception rather than the norm.

Resilience against advanced and collusive adversarial attacks

While many frameworks are designed to detect external intrusions, a significant research gap lies in the resilience of the FL process itself against sophisticated, internal adversarial attacks. The vulnerability of FL models to data poisoning and Byzantine attacks is a frequently cited area for future research (Blika et al. 2025). As noted in Dhanushkodi et al. (2024) attackers can manipulate input data to trick AI systems, highlighting a critical need for

Table 6 Summary of identified research gaps and future directions

Research gap category	Key limitations and 'Blind spots'	Implied future direction
Data and Validation	•Reliance on outdated or synthetic datasets •Scarcity of high-quality, labeled data •Lack of validation on actual physical testbeds	Development of standardized, realistic CPS datasets and greater focus on empirical validation
Adversarial Resilience	•Limited resilience against sophisticated, internal adversarial attacks •Insufficient defenses against stealthy, multi-party, and collusive attacks	Design of advanced defenses against multi-party and Byzantine threats, moving toward reputation-based aggregation
Scalability & Efficiency	•Performance bottlenecks from central aggregators •Difficulty of learning on resource-constrained edge devices •Management of Non-IID data requires more robust solutions	Development of resource-aware FL algorithms, optimization of on-device training, and new strategies for statistical heterogeneity
Trust & Transparency	•The black box nature of deep learning models hinders trust •A comprehensive privacy evaluation metric is absent •Ethical issues like algorithmic fairness remain largely unexplored	Deeper integration of Explainable AI (XAI), creation of quantitative privacy metrics, and research into the ethical dimensions of FL

enhanced model robustness. Current research predominantly focuses on common attacks like DoS and DDoS and general model poisoning.

However, there is a recognized gap in developing defenses against more sophisticated, stealthy, and multi-party threats. Works like (Husnoo, et al. 2023) highlight the vulnerability of FL to byzantine threats, while (Husnoo, et al. 2023) points out that existing poisoning attacks are not sufficiently stealthy and that targeted model poisoning remains an open challenge. Crucially, (Yang et al. 2023c) states that defenses against single attackers are insufficient, as multiple attackers with different roles may compromise the entire FL system in the form of collusion, a threat not adequately covered by current approaches. For example, one malicious client could submit a model update containing a backdoor, while another colluding client submits updates that subtly amplify the backdoor's effect, an attack that is much harder to detect by methods that look for single statistical outliers. This threat is not adequately covered by current approaches.

Gaps in scalability, resource optimization, and heterogeneity management

While FL is proposed as a solution to the scalability issues of centralized systems, many FL frameworks themselves face practical scalability challenges. Traditional FL architectures that rely on a single central aggregator can create performance bottlenecks and single points of failure, compromising robustness in large-scale deployments (Althunayyan et al. 2024). Furthermore, implementing exhaustive learning on resource-constrained edge devices remains a significant hurdle. As noted in Latif et al. (2025), there is a pressing need for more energy-efficient algorithms and resource optimization strategies. A key blind spot identified is that most research focuses on global optimization, with limited attention paid to the critical issue of executing effective local training and inference at the edge. The management of Non-IID data, while a recognized challenge, also requires more robust and universally applicable solutions beyond the specific strategies currently proposed (Bouzinis et al. 2025).

Underdeveloped explainability, quantitative privacy, and ethical considerations

Deep neural networks, central to many surveyed frameworks, are often treated as "black boxes," making it difficult to understand their decision-making process. This lack of explainability, hinders human oversight and trust. While the trend towards Explainable AI (XAI) is emerging, its integration into FL for CPS security is still in its infancy. Similarly, while many FL solutions integrate PETs, there is still a significant challenge in providing comprehensive and quantitative privacy evaluation metrics. As explicitly noted in Aflaki et al. (2024), a comprehensive privacy evaluation metric tailored to the intricacies of FL is noticeably absent in the current literature. Finally, critical ethical issues such as algorithmic fairness and bias, which could lead to disproportionate monitoring or targeting, remain largely unexplored in the context of FL for CPS security (Mohamed 2025).

Gaps in engineering feasibility, failed assumptions, and real-world deployment

A critical analysis of the current literature reveals a significant disconnect between idealized algorithmic performance and the engineering realities of CPS deployments. Many proposed federated solutions rely on theoretical assumptions—such as high-speed synchronous communication and consistent client availability—that frequently fail in industrial settings. In environments like smart manufacturing or remote maritime monitoring, intermittent connectivity and "straggler" nodes are common, often leading to model stagnation or the total failure of synchronous aggregation protocols.

Furthermore, the engineering infeasibility of complex cryptographic defenses on resource-constrained hardware remains a significant hurdle. While techniques like Secure Multi-Party Computation (SMPC) provide high privacy guarantees, their computational and communication overhead can exceed the millisecond-latency requirements of mission-critical actuation loops. Additionally, most surveyed models struggle with concept drift; when the physical process itself changes due to mechanical wear or environmental shifts, the model's failure to adapt results in "alert fatigue." These negative outcomes highlight that a "one-size-fits-all" approach is often technically infeasible without extensive, domain-specific hardware tuning and fail-safe manual overrides.

Open challenges and future research directions

Having identified the significant research gaps in the current literature, this section outlines the open challenges and proposes specific research directions critical for advancing the field. The following discussion synthesizes the technical, security, and visionary frontiers that will shape the next generation of federated anomaly detection systems for CPS.

A primary technical frontier is the development of truly Resource-Aware FL. This extends beyond simple model compression to encompass a co-design of the learning process with the extreme resource heterogeneity of CPS edge devices. Future work should focus on lightweight, adaptive architectures and move from heuristic-based to optimization-based intelligent client selection mechanisms. Strategies that consider data quality, device resources, and trust are critical for optimizing both

efficiency and accuracy. Building on this, the challenge of Continual and Online Learning in non-stationary CPS environments remains paramount. The stability-plasticity dilemma, where learning new attack patterns can cause catastrophic forgetting, requires the development of online FL frameworks that can seamlessly integrate new knowledge from evolving data streams.

Parallel to these technical challenges are the frontiers of Security and Robustness. While current research addresses common threats, the next generation of defenses must be resilient against more sophisticated adversarial attacks. There is a pressing need for robust defenses against targeted model poisoning, Byzantine threats, and collusive attacks where multiple malicious clients coordinate to compromise the global model. Future research should move beyond statistical anomaly detection towards reputation-based or cryptographically verifiable defense mechanisms. This pursuit of verifiable security culminates in the need for Formal Verification and Certification. For FL systems to be deployed in safety–critical CPS, a higher standard of assurance is required. A significant future direction is the development of methods, such as model checking, to provide provable guarantees of a system's security and privacy properties.

Finally, the "black box" nature of many deep learning models remains a significant barrier to their adoption in high-stakes CPS environments. A critical frontier is resolving the inherent paradox between FL's privacy goals and the transparency requirements of Explainable AI (XAI). Future work should explore methods for creating federated meta-explanations—techniques that aggregate local, privacy-preserving explanations into a meaningful global insight without compromising client confidentiality.

Conclusion

This survey has provided a comprehensive and critical analysis of the state-of-the-art in securing Cyber-Physical Systems (CPS) through Federated Learning (FL). As digital threats increasingly precipitate physical consequences, the shift towards decentralized, edge-native security paradigms is not merely advantageous but essential for protecting the world's critical infrastructure. Our work addresses this imperative by systematically organizing, deconstructing, and evaluating this rapidly evolving domain.

The primary contribution of this paper is the introduction of a novel, multi-axis taxonomy that structures the field by architecture, detection methodology, application domain, and privacy-preservation scheme. Second, we synthesize this analysis into a prescriptive synthetic taxonomy that serves as a practical guide, connecting specific CPS threat models to the most effective FL security architectures and techniques. This framework moves beyond a simple literature review to provide a structured lens through which the complex interplay between competing system requirements can be understood. This is exemplified in our analysis of the performance trilemma—the fundamental trade-off between detection accuracy, communication efficiency, and privacy guarantees—that governs all FL implementations. By synthesizing the dominant trends and, just as importantly, identifying the significant research gaps, this study offers a clear and holistic perspective on the current landscape.

The importance of this work lies in its ability to bring order and direction to a burgeoning and often fragmented field of research. For academics and practitioners, this survey serves as both a foundational reference and a strategic guide, clarifying the key challenges that must be overcome and highlighting the most promising avenues for innovation.

Ultimately, this paper has not only cataloged the present state of FL-based anomaly detection but has also illuminated the path forward. The journey towards truly resilient and trustworthy security for CPS requires a concerted effort to address the identified gaps, particularly in real-world validation and defense against sophisticated adversarial threats. By building upon the foundations and insights presented here, the research community can accelerate the development of the next generation of intelligent security frameworks essential for a safe and interconnected future.

Acknowledgements
Not applicable.

Declaration of generative AI and AI-assisted technologies in the writing process
During the preparation of this work the authors used Gemini AI in order to improve language. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Author contributions
Conceptualization, A.P., Y.D.; methodology, A.P.; validation, Y.D., and G.J.A.; formal analysis, A.P. and Y.D.; investigation, A.P., and Y.D., and G.J.A.; resources, Y.D. and G.J.A.; data curation, A.P.; writing original draft preparation, A.P.; writing review and editing, Y.D., and G.J.A.; visualization, A.P.; supervision, Y.D. and G.J.A. All authors have read and agreed to the published version of the manuscript.

Funding
Andrea Pinto, the corresponding author, has received research support from Universidad de los Andes. The other authors have no relevant financial or non-financial interests to disclose.

Data availability
Not applicable.

Declarations

Competing interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Received: 5 September 2025 Accepted: 24 February 2026
Published online: 12 March 2026

References

Abbas S et al (2023) A novel federated edge learning approach for detecting cyberattacks in IoT infrastructures. IEEE Access 11:112189–112198. https://doi.org/10.1109/ACCESS.2023.3318866

Abdelkader S et al (2024) Securing modern power systems: implementing comprehensive strategies to enhance resilience and reliability against cyber-attacks. Results Eng 23:102647. https://doi.org/10.1016/j.rineng.2024.102647

Abdullahi SM, Lazarova-Molnar S (2025) On the adoption and deployment of secure and privacy-preserving IIoT in smart manufacturing: a comprehensive guide with recent advances. Int J Inf Secur. https://doi.org/10.1007/s10207-024-00951-8

AbuElHassan S, Abo Alian A, AbdelKader T, Badr N (2025) A review on privacy-preserving techniques for spatiotemporal data, 2025, Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s41060-025-00807-x.

Aflaki A, Karimipour H, Gadekallu TR (2024) Privacy-prioritized model aggregation in ICPS: a novel approach to federated learning aggregation with lime and blockchain. IEEE Trans Ind Cyber-Phys Syst 2:370–379. https://doi.org/10.1109/ticps.2024.3419751

Agrawal S et al (2022) Federated learning for intrusion detection system: concepts, challenges and future directions. Comput Commun 195:346–361. https://doi.org/10.1016/j.comcom.2022.09.012

Agrawal S et al (2022) Federated learning for intrusion detection system: concepts, challenges and future directions. Comput Commun. https://doi.org/10.1016/j.comcom.2022.09.012

Ahanger TA, Ullah I, Algamdi SA, Tariq U (2025) Machine learning-inspired intrusion detection system for IoT: security issues and future challenges. Comput Electr Eng 123:110265. https://doi.org/10.1016/j.compeleceng.2025.110265

Ahmad J et al (2025) Cybersecurity in smart microgrids using blockchain-federated learning and quantum-safe approaches: a comprehensive review. Appl Energy 393:126118. https://doi.org/10.1016/j.apenergy.2025.126118

Ahsan MS, Islam S, Shatabda S (2025) A systematic review of metaheuristics-based and machine learning-driven intrusion detection systems in IoT. Swarm Evol Comput 96:101984. https://doi.org/10.1016/j.swevo.2025.101984

Akpolat AN, Kalay MS (2025) Defense mechanism of PV-powered energy islands against cyber-attacks utilizing supervised machine learning. Appl Sci. https://doi.org/10.3390/app15095021

Alabdulatif A (2025) GuardianAI: privacy-preserving federated anomaly detection with differential privacy. Array 26:100381. https://doi.org/10.1016/j.array.2025.100381

Alazab M et al (2025) Adaptive protocols for hypervisor security in cloud infrastructure using federated learning-based anomaly detection. Eng Appl Artif Intell. https://doi.org/10.1016/j.engappai.2025.110750

Alhazmi M, Zhao AP, Li W, Yang C (2025) Federated learning for real-time demand response by data centers toward energy efficiency and privacy preservation. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3583436

Ali S, Li Q, Yousafzai A (2024) Blockchain and federated learning-based intrusion detection approaches for edge-enabled industrial IoT networks: a survey. Ad Hoc Netw 152:103320. https://doi.org/10.1016/j.adhoc.2023.103320

Ali S et al (2025a) CLDM-MMNNs: cross-layer defense mechanisms through multi-modal neural networks fusion for end-to-end cybersecurity—issues, challenges, and future directions. Inf Fusion 122:103222. https://doi.org/10.1016/j.inffus.2025.103222

Ali M, Suchismita M, Ali SS, Choi BJ, Multidisciplinary Digital Publishing Institute (MDPI) (2025b) Privacy-preserving machine learning for IoT-integrated smart grids: recent advances, opportunities, and challenges. Energies. https://doi.org/10.3390/en18102515

Althunayyan M, Javed A, Rana O, Spyridopoulos T (2024) Hierarchical federated learning-based intrusion detection for in-vehicle networks. Future Internet. https://doi.org/10.3390/fi16120451

Alturki B et al. (2025) IoMT landscape: navigating current challenges and pioneering future research trends, Jan. 01, 2025, Springer Nature. https://doi.org/10.1007/s42452-024-06351-w

Belenguer A, Pascual JA, Navaridas J (2025) A review of federated learning applications in intrusion detection systems. Comput Netw 258:111023. https://doi.org/10.1016/j.comnet.2024.111023

Bhardwaj T, Sumangali K (2025) An explainable federated blockchain framework with privacy-preserving AI optimization for securing healthcare data. Sci Rep. https://doi.org/10.1038/s41598-025-04083-4

Bhogal W, Hankins A, Watts J (2025) Predicts 2025: scaling zero-trust technology and resilience, Stamford, CT, USA, Mar. 2025

Blika A et al (2025) Federated learning for enhanced cybersecurity and trustworthiness in 5G and 6G networks: a comprehensive survey. IEEE Open J Commun Soc 6:3094–3130. https://doi.org/10.1109/OJCOMS.2024.3449563

Bouzinis PS et al (2025) StatAvg: mitigating data heterogeneity in federated learning for intrusion detection systems. IEEE Trans Netw Serv Manag. https://doi.org/10.1109/TNSM.2025.3564387

Chen Z, Tian P, Liao W, Yu W (2021) Zero knowledge clustering based adversarial mitigation in heterogeneous federated learning. IEEE Trans Netw Sci Eng 8(2):1070–1083. https://doi.org/10.1109/TNSE.2020.3002796

Chiriac BN, Anton FD, Ioniță AD, Vasilică BV (2025) A modular AI-driven intrusion detection system for network traffic monitoring in industry 4.0, using Nvidia Morpheus and generative adversarial networks. Sensors. https://doi.org/10.3390/s25010130

de Carvalho Bertoli G, Alves Pereira Junior L, Saotome O, dos Santos AL (2023) Generalizing intrusion detection for heterogeneous networks: a stacked-unsupervised federated learning approach. Comput Secur 127:103106. https://doi.org/10.1016/j.cose.2023.103106

Dhanushkodi K, Thejas S (2024) AI enabled threat detection: leveraging artificial intelligence for advanced security and cyber threat mitigation, 2024, Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2024.3493957.

Dragos I (2024) 2024 year in review: OT cybersecurity, Hanover, MD, USA, 2024

Duy PT, Hien DTT, Luong TD, Quyen NH, Pham V-H (2024) Fed-evolver: an automated evolving approach for federated Intrusion detection system using adversarial autoencoder in SDN-enabled networks. Internet of Things 28:101397. https://doi.org/10.1016/j.iot.2024.101397

Elkhodr M (2025) An AI-driven framework for integrated security and privacy in Internet of Things using quantum-resistant blockchain. Future Internet 17(6):246. https://doi.org/10.3390/fi17060246

Friha O, Ferrag MA, Shu L, Maglaras L, Choo K-KR, Nafaa M (2022) FELIDS: Federated learning-based intrusion detection system for agricultural Internet of Things. J Parallel Distrib Comput 165:17–31. https://doi.org/10.1016/j.jpdc.2022.03.003

Friha O, Ferrag MA, Benbouzid M, Berghout T, Kantarci B, Choo K-KR (2023) 2DF-IDS: decentralized and differentially private federated learning-based intrusion detection system for industrial IoT. Comput Secur 127:103097. https://doi.org/10.1016/j.cose.2023.103097

Ghadi YY et al (2025) A hybrid AI-blockchain security framework for smart grids. Sci Rep 15(1):20882. https://doi.org/10.1038/s41598-025-05257-w

Gupta H et al. (2023) Commissioning federated reinforcement learning to envision network security strategies. In: 2023 IEEE 20th India Council International Conference, INDICON 2023, Institute of Electrical and Electronics Engineers Inc. pp. 1007–1012. https://doi.org/10.1109/INDICON59947.2023.10440912

Hamad NA, Bakar KAA, Qamar F, Jubair AM, Mohamed RR, Mohamed MA (2025) Systematic analysis of federated learning approaches for intrusion detection in the internet of things environment. IEEE Access 13:95410–95444. https://doi.org/10.1109/ACCESS.2025.3574672

Hamouda D, Ferrag MA, Benhamida N, Seridi H (2023) PPSS: a privacy-preserving secure framework using blockchain-enabled federated deep

learning for Industrial IoTs. Pervasive Mob Comput 88:101738. https://doi.org/10.1016/j.pmcj.2022.101738
Hayawi K, Sajid J, Malik AW, Trabelsi Z (2025) Revolutionizing electric robot charging infrastructure through federated transfer learning and data route optimization. Cluster Comput. https://doi.org/10.1007/s10586-024-05001-5
He X et al (2025) Artificial intelligence security and privacy: a survey. Sci China Inf Sci 68(8):181101. https://doi.org/10.1007/s11432-025-4388-5
Hossain S, Senouci SM, Brik B, Boualouache A (2025) A privacy-preserving self-supervised learning-based intrusion detection system for 5G-V2X networks. Ad Hoc Netw. https://doi.org/10.1016/j.adhoc.2024.103674
Hossain MT, Islam S, Badsha S, Shen H (2021) DeSMP: differential privacy-exploited stealthy model poisoning attacks in federated learning. In: Proceedings-2021 17th International Conference on Mobility, Sensing and Networking, MSN 2021, Institute of Electrical and Electronics Engineers Inc., pp. 167–174. https://doi.org/10.1109/MSN53354.2021.00038
Huong TT et al (2021) Detecting cyberattacks using anomaly detection in industrial control systems: a federated learning approach. Comput Ind 132:103509. https://doi.org/10.1016/j.compind.2021.103509
Husnoo MA et al. (2023), FeDiSa: a semi-asynchronous federated learning framework for power system fault and cyberattack discrimination. In: IEEE INFOCOM 2023-Conference on Computer Communications Workshops, INFOCOM WKSHPS 2023, Institute of Electrical and Electronics Engineers Inc., https://doi.org/10.1109/INFOCOMWKSHPS57453.2023.10226030
Hussien M, Cheriet M, Nguyen KK, Larabi A, Baek J (2025) GenAI-based privacy-preserving transfer learning. IEEE Trans Ind Cyber Phys Syst 3:329–340. https://doi.org/10.1109/ticps.2025.3556993
Javeed D, Saeed MS, Adil M, Kumar P, Jolfaei A (2024) A federated learning-based zero trust intrusion detection system for Internet of Things. Ad Hoc Netw 162:103540. https://doi.org/10.1016/j.adhoc.2024.103540
Jayanthiladevi A, Natarajan J, Arjun KP, Atlas LG, Arvindhan M, Arockiam D (2025) AI-based cybersecurity frameworks for 7G-enabled virtual therapy platforms. Cyber Secur Appl. https://doi.org/10.1016/j.csa.2025.100099
Kamatchi K, Uma E (2025) Securing the edge: privacy-preserving federated learning for insider threats in IoT networks. J Supercomput. https://doi.org/10.1007/s11227-024-06752-z
Kanyama MN, Bhunu Shava F, Gamundani AM, Hartmann A (2024) Machine learning applications for anomaly detection in smart water metering networks: a systematic review. Phys Chem Earth, Parts a/b/c 134:103558. https://doi.org/10.1016/j.pce.2024.103558
Karampasi A et al. (2024) Towards transparent AI-powered cybersecurity in financial systems: the deployment of federated learning and explainable AI in the CaixaBank pilot. In: IEEE International Conference on Data Mining Workshops, *ICDMW*, IEEE Computer Society, 2024, pp. 270–277. https://doi.org/10.1109/ICDMW65004.2024.00041
Khacha A, Aliouat Z, Harbi Y, Gherbi C, Saadouni R, Harous S (2024) Landscape of learning techniques for intrusion detection system in IoT: a systematic literature review. Comput Electr Eng 120:109725. https://doi.org/10.1016/j.compeleceng.2024.109725
Khan IA, Pi D, Abbas MZ, Zia U, Hussain Y, Soliman H (2023) Federated-SRUs: a federated-simple-recurrent-units-based IDS for accurate detection of cyber attacks against IoT-augmented industrial control systems. IEEE Internet Things J 10(10):8467–8476. https://doi.org/10.1109/JIOT.2022.3200048
Kheddar H (2025) Transformers and large language models for efficient intrusion detection systems: a comprehensive survey. Inf Fusion 124:103347. https://doi.org/10.1016/j.inffus.2025.103347
Kheddar H, Himeur Y, Awad AI (2023) Deep transfer learning for intrusion detection in industrial control networks: a comprehensive review. J Netw Comput Appl 220:103760. https://doi.org/10.1016/j.jnca.2023.103760
Kumar K, Khari M (2025) Federated active meta-learning with blockchain for zero-day attack detection in industrial IoT. Peer-to-Peer Netw Appl. https://doi.org/10.1007/s12083-025-02014-8
Kumar P, Kumar R, Jolfaei A, Mohammad N (2024) An automated threat intelligence framework for vehicle-road cooperation systems. IEEE Internet Things J 11(22):35964–35974. https://doi.org/10.1109/JIOT.2024.3397652
kumar sah D, Vahabi M, Fotouhi H (2025) Federated learning at the edge in Industrial Internet of Things: a review. Sustain Comput Inform Syst. https://doi.org/10.1016/j.suscom.2025.101087
Latif N, Ma W, Ahmad HB (2025) Advancements in securing federated learning with IDS: a comprehensive review of neural networks and feature engineering techniques for malicious client detection. Artif Intell Rev. https://doi.org/10.1007/s10462-024-11082-w
Li B, Wu Y, Song J, Lu R, Li T, Zhao L (2021) DeepFed: federated deep learning for intrusion detection in industrial cyber-physical systems. IEEE Trans Ind Inform 17(8):5615–5624. https://doi.org/10.1109/TII.2020.3023430
Li H, Ge L, Tian L (2024) Survey: federated learning data security and privacy-preserving in edge-Internet of Things. Artif Intell Rev. https://doi.org/10.1007/s10462-024-10774-7
Ma J, Su W (2025) Collaborative DDoS defense for SDN-based AIoT with autoencoder-enhanced federated learning. Inf Fusion 117:102820. https://doi.org/10.1016/j.inffus.2024.102820
Macas M, Wu C, Fuertes W (2022) A survey on deep learning for cybersecurity: progress, challenges, and opportunities. Comput Networks 212:109032. https://doi.org/10.1016/j.comnet.2022.109032
Mahmoodi ABZ, Sheikhi S, Peltonen E, Kostakos P (2023) Autonomous federated learning for distributed intrusion detection systems in public networks. IEEE Access 11:121325–121339. https://doi.org/10.1109/ACCESS.2023.3327922
Makris I et al (2025) A comprehensive survey of federated intrusion detection systems: techniques, challenges and solutions. Comput Sci Rev 56:100717. https://doi.org/10.1016/j.cosrev.2024.100717
Manivannan D (2024) Recent endeavors in machine learning-powered intrusion detection systems for the internet of things. J Netw Comput Appl 229:103925. https://doi.org/10.1016/j.jnca.2024.103925
Mehta N, Bharot N, Breslin JG, Verma P (2025) PPFL-DCS: privacy-preserving federated learning using neural transformer and leveraging dynamic client selection to accommodate data diversity. IEEE Access 13:94225–94238. https://doi.org/10.1109/ACCESS.2025.3572605
Mia J, Hadi Amini M (2024) A secure object detection technique for intelligent transportation systems. IEEE Open J Intell Transp Syst 5:495–508. https://doi.org/10.1109/OJITS.2024.3440876
Min W, Muthanna MSA, Ibrahim M, Alkanhel R, Muthanna A, Laouid A (2025) Privacy-preserving federated UAV data collection framework for autonomous path optimization in maritime operations. Appl Soft Comput. https://doi.org/10.1016/j.asoc.2025.112906
Mohamed N (2025) Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. Knowl Inf Syst. https://doi.org/10.1007/s10115-025-02429-y
Mughal FR et al (2024) Adaptive federated learning for resource-constrained IoT devices through edge intelligence and multi-edge clustering. Sci Rep. https://doi.org/10.1038/s41598-024-78239-z
Namakshenas D, Yazdinejad A, Dehghantanha A, Parizi RM, Srivastava G (2024) IP2FL: interpretation-based privacy-preserving federated learning for industrial cyber-physical systems. IEEE Trans Ind Cyber-Phys Syst 2:321–330. https://doi.org/10.1109/ticps.2024.3435178
Nandanwar H, Katarya R (2025a) Securing industry 5.0: an explainable deep learning model for intrusion detection in cyber-physical systems. Comput Electr Eng 123:110161
Nandanwar H, Katarya R (2025b) Securing Industry 5.0: an explainable deep learning model for intrusion detection in cyber-physical systems. Comput Electr Eng 123:110161
Nguyen G et al (2025) Landscape of machine learning evolution: privacy-preserving federated learning frameworks and tools. Artif Intell Rev. https://doi.org/10.1007/s10462-024-11036-2
Nkoom M, Hounsinou SG, Crosby GV (2025a) Securing the internet of robotic things (IoRT) against DDoS attacks: a federated learning with differential privacy clustering approach. Comput Secur 155:104493. https://doi.org/10.1016/j.cose.2025.104493
Nkoom M, Hounsinou SG, Crosby GV (2025b) Securing the internet of robotic things (IoRT) against DDoS attacks: a federated learning with differential privacy clustering approach. Comput Secur 155:104493. https://doi.org/10.1016/j.cose.2025.104493
Orabi MM, Emam O, Fahmy H (2025) Adapting security and decentralized knowledge enhancement in federated learning using blockchain technology: literature review. J Big Data. https://doi.org/10.1186/s40537-025-01099-5

Peng M, Wang J, Song D, Miao F, Su L (2023) Privacy-preserving and uncertainty-aware federated trajectory prediction for connected autonomous vehicles. In: IEEE International Conference on Intelligent Robots and Systems, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 11141–11147. https://doi.org/10.1109/IROS55552.2023.10341638
Poorazad SK, Benzaïd C, Taleb T (2024) A novel buffered federated learning framework for privacy-driven anomaly detection in IIoT. In: Proceedings - IEEE Global Communications Conference, GLOBECOM, Institute of Electrical and Electronics Engineers Inc., pp. 1725–1730. https://doi.org/10.1109/GLOBECOM52923.2024.10901786
Praharaj L, Gupta D, Gupta M (2025) Efficient federated transfer learning-based network anomaly detection for cooperative smart farming infrastructure. Smart Agric Technol 10:100727. https://doi.org/10.1016/j.atech.2024.100727
Rahdari A et al (2025) A survey on privacy and security in distributed cloud computing: exploring federated learning and beyond. IEEE Open J Commun Soc 6:3710–3744. https://doi.org/10.1109/OJCOMS.2025.3560034
Raza M, Saeed MJ, Riaz MB, Sattar MA (2024) Federated learning for privacy-preserving intrusion detection in software-defined networks. IEEE Access 12:69551–69567. https://doi.org/10.1109/ACCESS.2024.3395997
Razzak I, Xu G, Khan MK (2022) Guest editorial: privacy-preserving federated machine learning solutions for enhanced security of critical energy infrastructures. IEEE Comput Soc. https://doi.org/10.1109/TII.2021.3128962
Rehman T, Tariq N, Khan FA, Rehman SU (2025) FFL-IDS: a fog-enabled federated learning-based intrusion detection system to counter jamming and spoofing attacks for the industrial Internet of Things. Sensors. https://doi.org/10.3390/s25010010
Reis MJCS (2025) Edge-FLGuard: a federated learning framework for real-time anomaly detection in 5G-enabled IoT ecosystems. Appl Sci. https://doi.org/10.3390/app15126452
Rizvi S, Demeri A (2025) Life at risk: uncovering the urgent security gaps in internet of things-integrated cloud infrastructures. Computer 58(8):102–106. https://doi.org/10.1109/MC.2025.3561445
Saha S, Hota A, Chattopadhyay AK, Nag A, Nandi S (2024) A multifaceted survey on privacy preservation of federated learning: progress, challenges, and opportunities. Artif Intell Rev. https://doi.org/10.1007/s10462-024-10766-7
Sakr HA, Fouda MM, Ashour AF, Abdelhafeez A, El-Afifi MI, Abdellah MR (2024) Machine learning-based detection of DDoS attacks on IoT devices in multi-energy systems. Egypt Inform J. https://doi.org/10.1016/j.eij.2024.100540
Sarker A, Jesser A, Speidel M (2024) Enhancing decentralized federated learning with user feedback loops: a novel approach for personalized and adaptive learning in iot environments. In: 2024 IEEE 3rd International Conference on Computing and Machine Intelligence, ICMI 2024 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2024. https://doi.org/10.1109/ICMI60790.2024.10585926
IBM Security (2024) Cost of a Data Breach Report 2024, Armonk, NY, USA, 2024
Selvam P et al (2025) Federated learning-based hybrid convolutional recurrent neural network for multi-class intrusion detection in IoT networks. Discov Internet Things. https://doi.org/10.1007/s43926-025-00130-8
Selvaraj AK, Prathiba SB, Kumar AD, Dhanalakshmi R, Gadekallu TR, Srivastava G (2024) Co-training-based personalized federated learning with generative adversarial networks for enhanced mobile smart healthcare diagnosis. IEEE Trans Consum Electron 70(3):6131–6139. https://doi.org/10.1109/TCE.2024.3460469
Senthil K, Karthikeyan R, Priya SS, Monikaa R, Ramamoorthi S, Hussain SFM (2025) Advanced privacy protection (APP) machine learning model using cryptographic techniques for IoT. Discover Appl Sci. https://doi.org/10.1007/s42452-025-06571-8
Shanmugarasa Y, Paik H, Kanhere SS, Zhu L (2023) A systematic review of federated learning from clients' perspective: challenges and solutions. Artif Intell Rev 56:1773–1827. https://doi.org/10.1007/s10462-023-10563-8
Shao J-M, Zeng G-Q, Lu K-D, Geng G-G, Weng J (2024) Automated federated learning for intrusion detection of industrial control systems based on evolutionary neural architecture search. Comput Secur 143:103910. https://doi.org/10.1016/j.cose.2024.103910
Sharma N, Shambharkar PG (2025) Transforming security in internet of medical things with advanced deep learning-based intrusion detection frameworks. Appl Soft Comput. https://doi.org/10.1016/j.asoc.2025.113420
Sharma V, Kumar A, Sharma K (2025) Digital twin: securing IoT networks using integrated ECC with blockchain for healthcare ecosystem. Knowl Inf Syst 67(3):2395–2426. https://doi.org/10.1007/s10115-024-02273-6
Shawkat M, El-desoky A, Ali ZH, Salem M (2025) Blockchain and federated learning based on aggregation techniques for industrial IoT: a contemporary survey. Peer-to-Peer Netw Appl. https://doi.org/10.1007/s12083-025-01991-0
Shen J, Yang W, Chu Z, Fan J, Niyato D, Lam KY, Effective intrusion detection in heterogeneous Internet-of-Things networks via ensemble knowledge distillation-based federated learning. In: IEEE International Conference on Communications, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 2034–2039. https://doi.org/10.1109/ICC51166.2024.10622262
Shenoy D, Bhat R, Krishna Prakasha K (2025) Exploring privacy mechanisms and metrics in federated learning. Artif Intell Rev. https://doi.org/10.1007/s10462-025-11170-5
Singh K (2025) Industrial internet of things fortify: multi-domain feature learning framework with deepdetectnet++ for improved intrusion detection. Comput Secur 156:104506. https://doi.org/10.1016/j.cose.2025.104506
Singh J, Wazid M, Das AK, Chamola V, Guizani M (2022) Machine learning security attacks and defense approaches for emerging cyber physical applications: a comprehensive survey. Comput Commun 192:316–331. https://doi.org/10.1016/j.comcom.2022.06.012
Soares K, Shinde AA, Patil M (2025) PPFedSL: privacy preserving split and federated learning enabled secure data sharing model for internet of vehicles in smart city. Int J Comput Netw Appl 12(2):154–177. https://doi.org/10.22247/ijcna/2025/11
Su X, Zhang G (2025) APFed: adaptive personalized federated learning for intrusion detection in maritime meteorological sensor networks. Digit Commun Networks 11(2):401–411. https://doi.org/10.1016/j.dcan.2024.02.001
Tabassum A, Erbad A, Lebda W, Mohamed A, Guizani M (2022) FEDGAN-IDS: privacy-preserving IDS using GAN and federated learning. Comput Commun 192:299–310. https://doi.org/10.1016/j.comcom.2022.06.015
Takele AK, Villányi B (2025) Resource-efficient clustered federated learning framework for Industry 4.0 edge devices. AI. https://doi.org/10.3390/ai6020030
Taslimasa H, Dadkhah S, Neto ECP, Xiong P, Ray S, Ghorbani AA (2023) Security issues in Internet of Vehicles (IoV): a comprehensive survey. Internet of Things 22:100809. https://doi.org/10.1016/j.iot.2023.100809
Torre D, Chennamaneni A, Jo JY, Vyas G, Sabrsula B (2025) Toward enhancing privacy preservation of a federated learning CNN intrusion detection system in IoT: method and empirical study. ACM Trans Softw Eng Methodol. https://doi.org/10.1145/3695998
Vemulapalli L, Sekhar PC (2025) A customized temporal federated learning through adversarial networks for cyber attack detection in IoT. J Robot Control (JRC) 6(1):366–384. https://doi.org/10.18196/jrc.v6i1.24529
Verma P, Bharot N, Breslin JG, O'shea D, Vidyarthi A, Gupta D (2024) Zero-day guardian: a dual model enabled federated learning framework for handling zero-day attacks in 5G enabled IIoT. IEEE Trans Consum Electron 70(1):3856–3866. https://doi.org/10.1109/TCE.2023.3335385
Wehbi O et al (2023) FedMint: intelligent bilateral client selection in federated learning with newcomer IoT devices. IEEE Internet Things J 10(23):20884–20898. https://doi.org/10.1109/JIOT.2023.3283855
Wen M, Zhang Y, Zhang P, Chen L (2025) IDS-DWKAFL: an intrusion detection scheme based on dynamic weighted K-asynchronous federated learning for smart grid. J Inf Secur Appl. https://doi.org/10.1016/j.jisa.2025.103993
Xia Z, Zhou H, Hu Z, Jiang Q, Zhou K (2025) Semi-asynchronous federated learning-based privacy-preserving intrusion detection for advanced metering infrastructure. Int J Crit Infrastruct Prot 49:100742. https://doi.org/10.1016/j.ijcip.2025.100742
Xiao Y et al (2025) Privacy protection anomaly detection in smart grids based on combined PHE and TFHE homomorphic encryption. Electronics 14(12):2386. https://doi.org/10.3390/electronics14122386
Yang R, He H, Wang Y, Qu Y, Zhang W (2023a) Dependable federated learning for IoT intrusion detection against poisoning attacks. Comput Secur 132:103381. https://doi.org/10.1016/j.cose.2023.103381

Yang R et al (2023b) Efficient intrusion detection toward IoT networks using cloud–edge collaboration. Comput Netw 228:109724. https://doi.org/10.1016/j.comnet.2023.109724
Yang LT, Zhao R, Liu D, Lu W, Deng X (2023c) Tensor-empowered federated learning for cyber-physical-social computing and communication systems. IEEE Commun Surv Tutor 25(3):1909–1940. https://doi.org/10.1109/COMST.2023.3282264.
Yang Z, Cheng C, Li Z, Wang R, Zhang X (2025) Reliable federated learning based on delayed gradient aggregation for intelligent connected vehicles. Eng Appl Artif Intell. https://doi.org/10.1016/j.engappai.2024.109719
You X, Liu C, Li J, Sun Y, Liu X (2024) FedMDO: privacy-preserving federated learning via mixup differential objective. IEEE Trans Circuits Syst Video Technol 34(10):10449–10463. https://doi.org/10.1109/TCSVT.2024.3408463
Yousefnezhad N, Malhi A, Främling K (2020) Security in product lifecycle of IoT devices: a survey. J Netw Comput Appl 171:102779. https://doi.org/10.1016/j.jnca.2020.102779
Zhang H, Su Q (2025) PJPFL: personalized federated learning with privacy preservation based on sample similarity. Inf Fusion. https://doi.org/10.1016/j.inffus.2025.103221
Zhang C, Yang S, Mao L, Ning H (2024) Anomaly detection and defense techniques in federated learning: a comprehensive review. Artif Intell Rev. https://doi.org/10.1007/s10462-024-10796-1
Zhang H, Ye J, Huang W, Liu X, Gu J (2025) Survey of federated learning in intrusion detection. J Parallel Distrib Comput 195:104976. https://doi.org/10.1016/j.jpdc.2024.104976
Zhu L, Zhao B, Guo J, Ji M, Peng J (2025) A cutting-edge framework for industrial intrusion detection: privacy-preserving, cost-friendly, and powered by federated learning. Appl Intell. https://doi.org/10.1007/s10489-025-06404-6

Publisher's Note