Poisoning Attacks in Federated Learning: An Accountability-Oriented Survey with Centralized Learning as Baseline

Safiia Mohammed *, Dima Alhadidi , Alioune Ngom

Posted Date: 25 May 2026

doi: 10.20944/preprints202605.1674.v1

Keywords: poisoning attacks; federated learning; model poisoning; backdoor attacks; byzantine-Robust aggregation; cryptographic verification; accountability; trustworthy AI

 doi:10.20944/preprints202605.1674.v1

Article

Poisoning Attacks in Federated Learning: An Accountability-Oriented Survey with Centralized Learning as Baseline

Safiia Mohammed *, Dima Alhadidi and Alioune Ngom

Department of Computer Science, University of Windsor, Windsor, Ontario, Canada
* Correspondence: mohamm7d@uwindsor.ca

Abstract

Artificial intelligence (AI) systems are increasingly deployed in critical domains such as healthcare, finance, defense, and transportation. These deployments, however, face growing risks from poisoning attacks that corrupt training data, manipulate model updates, or implant covert backdoors. Such attacks undermine trust, reduce transparency, and challenge the safe and accountable use of AI in high-stakes settings. This survey examines poisoning attacks in federated learning (FL), using centralized learning as a comparative baseline to clarify how the threat landscape changes when data, updates, and control are distributed. Rather than reintroducing a generic poisoning taxonomy as a standalone contribution, we position the paper relative to prior surveys, identify what remains insufficiently covered, and synthesize representative primary studies through an accountability-oriented lens. Because FL introduces additional vulnerabilities, including untrusted servers, non-IID heterogeneity, and limited observability of client behavior, we analyze how these properties expand the poisoning threat surface. We review state-of-the-art countermeasures, including Byzantine-robust aggregation, anomaly detection, validation-based defenses, and cryptographic prevention mechanisms such as malicious-secure aggregation, authenticated update handling, and verifiable aggregation protocols. Particular emphasis is placed on accountability-enabling mechanisms such as auditability, traceability, and forensic readiness, which are essential to responsible AI and regulatory compliance. Our analysis identifies persistent research gaps, including the lack of unified privacy-robustness-accountability frameworks, insufficient defenses against server-side attacks, limited verifiability tools, and scalability challenges in real-world FL systems. The paper's contribution is therefore not to claim that accountability-oriented FL defenses are new, but to clarify the survey scope, position prior surveys directly, and integrate aggregation-based, cryptographic, and governance-oriented strands within an Accountability-Integrated Taxonomy (AIT) and an evidence-oriented discussion of trustworthy federated learning.

Keywords: poisoning attacks; federated learning; model poisoning; backdoor attacks; byzantine-Robust aggregation; cryptographic verification; accountability; trustworthy AI

1. Introduction

Artificial intelligence (AI) is transforming sectors ranging from healthcare and education to finance and public administration. As its adoption accelerates, concerns about reliability, transparency, and safety have become increasingly prominent [1]. In high-stakes settings, poisoning attacks are especially consequential because they target the training process itself: attackers can corrupt data, manipulate model updates, or implant backdoors before a system is deployed [2,3]. These risks are even harder to analyze in federated learning (FL), where training is distributed across parties that do not share raw data and where the aggregation process itself can become an attack surface. This is where accountability

 doi:10.20944/preprints202605.1674.v1

becomes central rather than peripheral. In AI, accountability refers to the ability to identify responsible actors, inspect relevant evidence, and justify corrective action when a system fails or is manipulated [4,1,5]. Closely related concepts such as auditability and traceability are directly relevant to poisoning research because any meaningful response to a poisoning incident depends on evidence showing who acted, what changed, when it changed, and whether that evidence can be independently verified [6,7,8]. In centralized learning, such evidence may be available through dataset provenance, training logs, and unified control of the pipeline. In FL, however, privacy-preserving mechanisms, client heterogeneity, and untrusted or semi-trusted coordination make these questions substantially harder to answer. The survey literature on poisoning is already extensive, but its emphasis remains fragmented. Some surveys develop broad taxonomies of adversarial attacks in machine learning [9]; others examine poisoning attacks and defenses in general ML [3,10]; and more recent work reviews attacks and defenses in FL more specifically [11,12,13,14,15]. What remains comparatively underdeveloped is a direct positioning of these surveys against one another, a clearer explanation of what a new survey still needs to contribute, and a sustained discussion of accountability, untrusted-server threats, and verifiable evidence in FL. Importantly, this underdevelopment should not be interpreted as absence. Recent FL studies already examine verifiable aggregation, privacy-preserving traceability, provenance tracking, auditable participant selection, and zero-knowledge–backed supervision [16,17,18,19,20,21]. The gap is therefore integrative rather than absolute: these technical mechanisms are rarely synthesized alongside aggregation-based and detection-based defenses within a common accountability-oriented comparison. Accordingly, this manuscript is positioned as a scoping review with analytical survey objectives organized in two stages. First, we compare prior surveys to clarify scope, overlap, and remaining gaps. Second, we synthesize representative primary studies to analyze poisoning attacks and defenses through an accountability-centric lens. Federated learning remains the main object of study throughout the paper. Centralized learning is included only as a comparative baseline to help isolate what changes once observability, trust, and control are distributed. This paper makes the following contributions:

1. It positions this survey against nine core prior surveys, clarifying where the literature is already mature and where important gaps remain.
2. It uses centralized learning as a comparative baseline while maintaining federated learning as the primary scope of analysis.
3. It introduces an Accountability-Integrated Taxonomy (AIT) that extends conventional poisoning categories with additional dimensions for observability, attribution granularity, required audit evidence, and trust assumptions.
4. It synthesizes FL defenses—including aggregation-based, detection-based, and cryptographic mechanisms—through the lenses of auditability, traceability, and accountability, while offering a deeper critical discussion of untrusted servers, verifiable aggregation, and standards-relevant evidence.

2. Survey Methodology

This work is positioned as a scoping review with analytical survey objectives rather than as a formal systematic review. Accordingly, the purpose of the methodology is not to exhaustively enumerate all published studies, but to transparently identify, position, and synthesize representative literature that can support a critical discussion of poisoning attacks and defenses in federated learning. The review is structured around four practical questions: (RQ1) how prior surveys delimit the poisoning problem in FL; (RQ2) which threat assumptions remain repeatedly underexplored, particularly with respect to server-side manipulation and collusion; (RQ3) how defense families perform under non-IID data, secure aggregation, and untrusted-server assumptions; and (RQ4) which accountability artifacts can make defenses auditable rather than merely robust. The literature collection process covered the period from 2015 to 2025. Relevant publications were retrieved from major digital libraries, including

 doi:10.20944/preprints202605.1674.v1

IEEE Xplore, the ACM Digital Library, SpringerLink, Elsevier ScienceDirect, and arXiv. The search used combinations of keywords such as "federated learning poisoning," "data poisoning attacks," "backdoor attacks," "model poisoning," "robust aggregation," "secure aggregation," and "defense survey." The objective of this initial search was to capture a broad pool of works related to poisoning threats, defenses, and accountability considerations in both centralized and federated learning settings. The initial search yielded approximately 120 papers, including surveys, empirical studies, and conceptual or framework-oriented works. These papers were first screened to remove duplicates and studies that were out of scope, such as works focused exclusively on application-specific settings (e.g., domain-specific IoT or medical imaging pipelines) without offering generalizable insight into poisoning or defense analysis. Unlike a systematic review, no formal quantitative quality scoring was applied. Instead, papers were evaluated qualitatively on the basis of their relevance, conceptual clarity, and contribution to understanding poisoning mechanisms, defense strategies, accountability implications, or gaps in existing survey coverage. From the screened pool, 18 survey and survey-like works were identified as providing structured overviews, taxonomies, or comparative discussions of poisoning attacks or defenses. A second screening stage then emphasized analytical depth and relevance to federated learning, excluding surveys that lacked a comparative structure, did not address training-time poisoning, or treated federated learning only marginally. This process resulted in a final set of 9 core surveys, which are comparatively positioned in Table 2. These selected works include foundational surveys that established general poisoning taxonomies Pitropakis et al. [9]; Tian et al. [3]; Cin'a et al. [10], as well as more recent surveys focused on FL threats, poisoning defenses, and life-cycle security issues Sikandar et al. [11]; Lianga et al. [12]; Xia et al. [13]; Nowroozi et al. [22]; Li et al. [14]; Zhou et al. [15]. The methodology therefore has two stages. Stage 1 positions this paper relative to prior surveys in order to make the survey's scope and novelty explicit. Stage 2 revisits representative primary studies outside the core comparison set to support the substantive synthesis of attacks, defenses, and accountability implications throughout the rest of the manuscript. These additional works contribute to the interpretation and comparison of defense strategies but are not exhaustively cataloged, consistent with the analytical survey focus of this manuscript. The screened corpus and the final bibliography are therefore not expected to correspond one-to-one. The initial retrieval set was intentionally broad and included duplicates, marginally relevant papers, application-specific studies without transferable insight into poisoning, and papers used only to test the boundaries of the survey scope during screening. Only the subset retained as core surveys or analytically central primary studies is cited in the final manuscript. In this sense, the bibliography reflects the final analytic corpus, whereas the larger figure reported above reflects the screened corpus.

2.1. Review Type, Search Strategy, and Eligibility Criteria

To avoid ambiguity, the manuscript should be read as a scoping review rather than as a formal systematic review. Its purpose is breadth, structuring, and analytical synthesis rather than statistical aggregation or exhaustive evidence appraisal. We therefore report the search process, screening logic, and corpus roles explicitly, while acknowledging that no PRISMA-style quality scoring or meta-analysis was performed. The search strategy combined venue-based retrieval (IEEE Xplore, the ACM Digital Library, SpringerLink, Elsevier ScienceDirect, and arXiv) with topic-based keyword families covering attacks, defenses, trust assumptions, and accountability concepts. Screening was guided by three eligibility principles: (1) poisoning relevance, meaning direct discussion of training-time poisoning, backdoors, or poisoning defenses; (2) FL relevance, meaning substantive treatment of decentralized or federated training rather than only incidental mention; and (3) analytical value, meaning that a work contributed either empirical evidence, a conceptual model, or a governance- or accountability-oriented perspective useful for comparison. We also distinguish source roles more explicitly than in the previous version. Survey and survey-like papers are used to position the manuscript within the existing review literature. Primary studies are then grouped into three functional categories: empirical (attack or defense evaluation), conceptual/framework (architectures, models, or

 doi:10.20944/preprints202605.1674.v1

trust analyses), and governance/standards (audit-, policy-, or compliance-oriented sources). Empirical works support comparative claims about threat models and defenses; conceptual works support architectural and accountability-oriented synthesis; and governance sources support the discussion of evidence, auditability, and standards alignment.

Table 1. Review protocol and study-selection logic used in this manuscript.

Stage	Input	Inclusion logic	Exclusion logic	Output
Discovery search	Broad pool (≈120 works)	Papers addressing poisoning attacks, poisoning defenses, FL trust assumptions, secure aggregation, or accountability-related evidence	Adversarial-example papers focused solely on inference, purely application-specific case studies without transferable poisoning insight, and duplicates	Initial literature corpus
Survey positioning	18 survey and survey-like works	Structured surveys with taxonomies, comparative discussion, or explicit treatment of poisoning in ML/FL	Only marginal treatment of FL, no comparative structure, or no meaningful treatment of training-time poisoning	9 core surveys for direct positioning
Primary-study synthesis	Remaining relevant papers	Empirical, conceptual, or governance works needed to analyze attacks, defenses, trade-offs, and audit artifacts	Papers cited only tangentially or lacking analytical value for the paper's research questions	Evidence base for later sections
Analytical output	Final manuscript corpus	Cross-survey positioning, scenario-based defense comparison, accountability indicators, and standards-oriented discussion	Claims of exhaustive or statistically pooled evidence synthesis	Scoping review with analytical survey contribution

2.2. Positioning Relative to Existing Surveys

Table 2 summarizes the nine core surveys retained in Stage 1 and the role each plays in our analysis. The comparison shows that prior surveys already cover baseline poisoning taxonomies, FL attack-and-defense catalogs, and broader life-cycle security concerns in substantial depth. Our aim is therefore not to suggest that the poisoning literature has gone unsurveyed. Rather, we compare these surveys directly, clarify where their scopes overlap, and identify what remains underdeveloped for an FL-oriented readership concerned with attribution, auditability, untrusted servers, and verifiable evidence. This manuscript likewise does not claim first-in-field status for accountability-oriented FL defense; instead, it brings together previously separate lines of work—including robust aggregation, cryptographic verification, provenance, and governance evidence—and examines where their coverage of poisoning resilience remains partial [17,19,21,16]. This positioning also clarifies the manuscript's scope. It is not intended as a general survey of all poisoning attacks across all ML settings. Instead, it focuses on FL and uses centralized learning only as a comparative baseline. Likewise, the paper is not a meta-review that merely summarizes surveys; rather, it uses survey comparison to motivate a second-stage synthesis of primary studies and to support the accountability-oriented analysis developed in later sections.

 doi:10.20944/preprints202605.1674.v1

Table 2. Positioning this survey relative to representative prior surveys.

Survey	Primary scope	Main strength	Gap relative to this work
Pitropakis et al. Pitropakis et al. [9]	Broad adversarial ML taxonomy	Foundational attack taxonomy across ML settings	Limited FL-specific poisoning detail and no accountability- or audit-evidence perspective
Tian et al. Tian et al. [3]	Poisoning in ML	Strong overview of poisoning attacks and countermeasures in centralized ML	Limited treatment of FL-specific trust boundaries, server-side threats, and forensic readiness
Cinà et al. Cinà et al. (2023)	Training-data poisoning in ML	Detailed synthesis of poisoning threat models and defenses	Emphasis remains on poisoning taxonomy itself rather than on FL accountability, untrusted aggregation, and evidence requirements
Sikandar et al. Sikandar et al. [11]	FL attacks and defenses broadly	Broad security overview of FL attack surfaces	Not poisoning-centric and offers limited critical discussion of accountability or auditability
Lianga et al. Lianga et al. [12]	FL poisoning attacks and defenses	FL-oriented catalog of attack and defense families	Limited direct comparison with other surveys and limited emphasis on accountability under untrusted-server assumptions
Xia et al. Xia et al. [13]	FL poisoning survey	Concise FL-specific poisoning overview	More descriptive than critical, with limited discussion of evidence-bearing defenses and governance implications
Nowroozi et al. Nowroozi et al. [22]	FL data poisoning	Recent FL-focused synthesis with updated attack examples	Narrower focus on data poisoning and a less developed treatment of accountability-oriented design
Li et al. Li et al. [14]	FL life-cycle threats and defenses	Broad life-cycle perspective on FL security and privacy	Poisoning is one component among many; there is limited centralized-baseline comparison and no explicit accountability taxonomy
Zhou et al. Zhou et al. [15]	FL data-poisoning defenses	Updated defense-focused discussion of FL poisoning	Limited attention to traceability, audit evidence, and malicious-server scenarios
This survey	FL poisoning with a centralized baseline	Cross-survey positioning, explicit scope clarification, AIT, and a deeper discussion of accountability and untrusted servers	Analytical rather than exhaustive; uses representative literature instead of claiming complete coverage

3. Poisoning Attacks and Accountability Taxonomy

3.1. Baseline Poisoning Taxonomy

This section establishes the technical background for the remainder of the paper. We first summarize the baseline poisoning taxonomy already established in prior literature, then use that baseline to show what changes in FL when observability, trust, and control are distributed across clients and server-side infrastructure. As machine learning models become more prevalent across application domains, attacks on the training phase have also intensified Zhang et al. [23]; Tian et al. [3]. In poisoning attacks, adversaries inject malicious data into training datasets or manipulate model updates to slow convergence, degrade accuracy, or implant targeted behaviors Srivastava et al. [24]; Biggio et al. [25]. These threats were first studied in centralized learning settings, where data and optimization are managed within a single training pipeline. FL preserves many of the same technical attack goals, but fundamentally changes the operational context: participants train locally, the server aggregates remote updates, and raw data remain hidden Bonawitz et al. [26]. As a result, FL inherits classical poisoning mechanisms while introducing new trust boundaries, new concealment opportunities, and new attribution problems Bhagoji et al. [27]. Most importantly, FL expands the threat surface beyond malicious clients alone. An untrusted or compromised server can manipulate aggregation, alter client weights, fork models, or collude with clients to implant backdoors.

Adversaries inject poisoned data or model updates during training, leading to compromised model behavior during inference.

These server-side possibilities are not merely minor variations of classical poisoning; they change what evidence is available, who can verify it, and how responsibility can be assigned after an incident. For this reason, the baseline taxonomy below is presented as established knowledge, while the manuscript's more distinctive analytical contribution appears later in the Accountability-Integrated Taxonomy (Section 3.4).

Poisoning attacks exploit vulnerabilities in the machine learning (ML) training process to corrupt model integrity, bias predictions, or implant persistent malicious behavior. Prior taxonomies have

 doi:10.20944/preprints202605.1674.v1

typically focused on either data poisoning or model poisoning in centralized settings Tian et al. [3]; Srivastava et al. [24]; Pitropakis et al. [9]; Gao et al. [28]; Li et al. [29]. However, these classifications often understate the broader threat landscape introduced by distributed learning paradigms such as FL, where adversaries may operate on the client or server side and where forensic visibility is limited. In this section, we synthesize the conventional taxonomy widely used in the literature and organize poisoning attacks along three complementary dimensions: attacker knowledge, manipulation target, and intent. This baseline structure consolidates threats applicable to both centralized and federated learning and provides the reference point that later allows us to explain what an accountability-oriented extension must add. Figure 1 illustrates the general poisoning workflow within the ML pipeline.

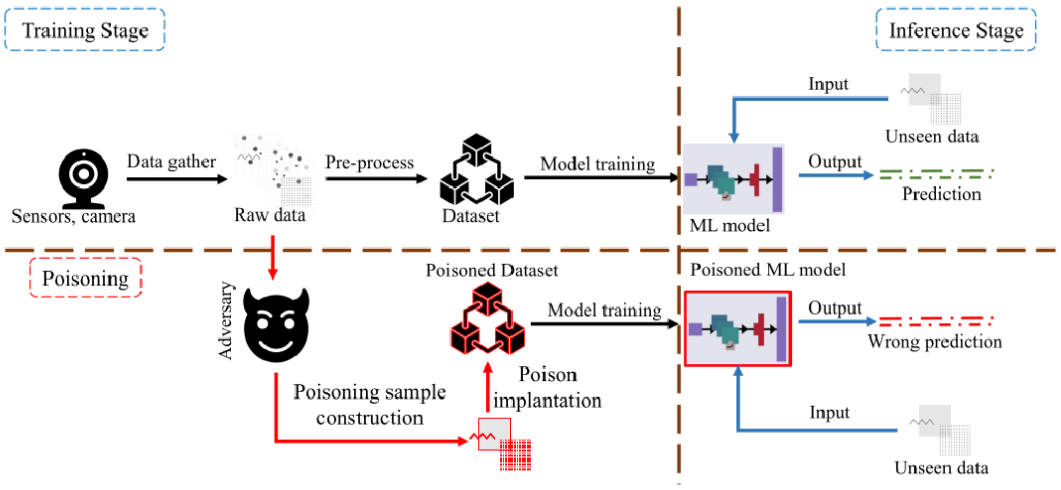

Figure 1. General poisoning attack in the machine-learning training pipeline, adapted from Tian et al. [3].

3.1.1. Taxonomy Overview

Poisoning attacks can be categorized along three orthogonal dimensions, each capturing a distinct aspect of adversarial capability:

1. Attacker knowledge: the degree of visibility into model internals or the training pipeline (white-box, gray-box, black-box) Papernot et al. [30].
2. Manipulation target: the component the adversary modifies (training data, model parameters, or training dynamics).
3. Attack intent: whether the goal is the untargeted degradation of overall performance or the targeted manipulation of specific outputs or behaviors.

 doi:10.20944/preprints202605.1674.v1

Table 3. Unified typology of poisoning attacks.

Access	Target	Intent	Example Mechanism	Representative Refs.
White-box	Data / Model	Untargeted / Targeted	Gradient manipulation, label flipping, and model replacement	Nowroozi et al. [22]; Sikandar et al. [11]; Biggio et al. [2]; Bagdasaryan et al. [31]
Gray-box	Data / Model	Untargeted / Targeted	Limited-knowledge poisoning and adaptive gradient updates (e.g., AgrEvader)	Mazzone et al. [32]; Zhang et al. [33]
Black-box	Data	Untargeted / Targeted	Query-based or transfer poisoning, shadow-model training	Sakhnovych [34]; Wang et al. [35]
White-box	Backdoor	Targeted	Trigger embedding and model-level implants	Lianga et al. [12]; Shafahi et al. [36]; Zhang et al. [37]; Shah et al. [38]

This unified taxonomy shows how attacker knowledge shapes the trade-off between impact and stealth. White-box adversaries exert maximal influence but are also more detectable through forensic logging or gradient fingerprints, whereas gray- and black-box adversaries rely on transferability or surrogate modeling to achieve stealth at the cost of fine-grained control. These distinctions are important for designing accountability-oriented defenses such as verifiable aggregation and provenance-based auditing.

3.1.2. Data Poisoning Attacks

Data poisoning introduces malicious or mislabeled samples into the training set to corrupt the learned decision boundary. Two primary variants are commonly distinguished:

- Clean-label poisoning: the attacker cannot modify labels, but instead crafts adversarial inputs that collide with target-class features Shafahi et al. [36]; Zhang et al. [39]. GAN-generated poisons can operate under non-IID conditions and without full knowledge of the victim model Xia et al. [13].
- Dirty-label poisoning: both inputs and labels are modified to implant malicious behavior or enable targeted misclassification Weng et al. [40]; Pan et al. [41]; Xu et al. [42]. Recent work shows that these categories remain active areas of research rather than closed historical cases. For example, recent defenses against data poisoning in neural networks, together with newer empirical evaluations of FL poisoning, continue to refine how clean-label, dirty-label, and label-flipping attacks are instantiated and measured in practice De Gaspari et al. [43]; Bena et al. [44]. A canonical example is label flipping: Static Label Flipping (SLF) directly alters labels (e.g., "1" to "7"), whereas Dynamic Label Flipping (DLF) targets feature-overlapping samples to increase stealth Zhang et al. [39]. These attacks highlight the risks posed by insufficient data provenance and the lack of audit trails.

3.1.3. Model Poisoning Attacks

Model poisoning manipulates gradients or parameters to distort global learning or implant persistent malicious behavior Bagdasaryan et al. [45]; Xie et al. [46]; Wan et al. [47]. This threat is especially critical in FL, where compromised clients can submit arbitrary updates to the server. Common mechanisms include the following:

- Model replacement: adversaries scale updates to overwrite the global model in a single round Bagdasaryan et al. [31].
- Gradient ascent: adversaries send updates in the opposite optimization direction to amplify divergence Shejwalkar and Houmansadr [48]; Guerraoui et al. [49].

 doi:10.20944/preprints202605.1674.v1

- Stealth poisoning: adversaries craft malicious updates that mimic benign statistics in order to evade anomaly detection Shejwalkar and Houmansadr [48].

These attacks exploit the limited transparency of FL, where server behavior and client updates are rarely auditable and verifiable aggregation is often absent.

3.1.4. Backdoor Attacks

Backdoor poisoning implants hidden triggers that activate only during inference when a specific pattern or condition is present Lianga et al. [12]; Shafahi et al. [36]. Backdoors may operate at:

- Data level: imperceptible patches, pixel patterns, or text perturbations embedded into poisoned samples Chen et al. [50]; Rocha and Conti [51].
- Model level: direct manipulation of model weights or aggregation procedures to implant latent malicious behavior Zhang et al. [37]; Shah et al. [38]. Effective backdoors are both stealthy and semantically valid, making them particularly dangerous in FL, where decentralized updates hinder attribution and forensic reconstruction.

3.1.5. Summary

Data, model, and backdoor poisoning attacks collectively reveal a continuum of adversarial manipulation spanning input perturbation, gradient corruption, and persistent malicious implants. Their success is amplified by gaps in provenance tracking, insufficient forensic logging, and the absence of verifiable aggregation mechanisms. In centralized learning, poisoning threats often arise from contaminated datasets or insider access Muñoz-González et al. (2017), whereas in FL the attack surface expands to include untrusted servers, heterogeneous clients, and cross-round collusion Bhagoji et al. [27]; Bagdasaryan et al. [31]. From an accountability standpoint, these attacks highlight a structural weakness: few existing defenses provide update traceability, cryptographic validation of model lineage, or mechanisms for attributing malicious behavior to specific clients or servers. Recent surveys Sun et al. [52]; Bagdasaryan et al. [31] underscore this fragmentation, in which robustness and privacy are addressed in isolation while accountability remains largely unexamined. These insights motivate the need for forensic-ready, audit-aware defense architectures that integrate robustness, accountability, and traceability. Subsequent sections build on this taxonomy to analyze poisoning in centralized and federated environments and to evaluate defenses that align with these accountability requirements.

3.2. Poisoning Attacks in Centralized Learning

Table 4 summarizes the dominant poisoning mechanisms in centralized learning, organized by attack strategy and intended impact. Although these approaches differ in subtlety and computational sophistication, they share the common assumption that attackers can inspect or directly modify the training dataset or model state.

3.2.1. Observations & Auditability Differences

Centralized poisoning attacks reveal structural properties that distinguish them from attacks in federated learning. Recent work from 2022 to 2025 reinforces these observations:

- Full data visibility enables high-impact, optimization-driven poisoning. In centralized settings, adversaries can directly inspect or manipulate the entire training dataset. This access enables precise bilevel optimization attacks Oprea et al. [53]; Jagielski et al. [54], adaptive gradient poisoning Srivastava et al. [24], and large-scale integrity attacks that are difficult to realize when only partial data visibility is available, as in FL Li et al. [14].
- The stealth-versus-impact trade-off is more pronounced. Recent clean-label and influence-based attacks Xia et al. [13]; Zhang et al. [37] achieve high stealth but smaller global impact, whereas aggressive label-flip and gradient-based poisons Gharib et al. [55]; Zhou et al. [15] cause

 doi:10.20944/preprints202605.1674.v1

substantial degradation but are easier to detect. Centralized settings allow adversaries to manage this trade-off more effectively because they retain full control over the data pipeline.

Table 4. Common poisoning attacks in centralized machine learning.

Attack Type	Description / Mechanism	Impact / Observations
Gradient-based Poisoning	Crafts poisoned samples using bilevel optimization to distort gradient updates Oprea et al. [53].	Can degrade accuracy by up to 60%; enables precise targeted misclassification or divergence.
Label Flipping	Randomly or selectively alters labels to mislead training Paudice et al. [56]; Gharib et al. [55].	Reduces accuracy by 20–40%; effects are visible in class-confusion patterns.
Clean-label Poisoning	Inserts benign-looking samples that collide with target-class features Shafahi et al. [36].	Hard to detect; preserves overall accuracy while enabling targeted misclassification.
Backdoor Poisoning	Embeds imperceptible triggers into poisoned inputs Gu et al. [57].	Produces near-100% targeted misclassification when the trigger is present.
Availability Attacks	Corrupts data or features to prevent convergence Kiss et al. [58].	Leads to accuracy collapse or unstable training failure.
Influence Function Attacks	Identifies and poisons influential training points using Hessian-vector analysis Koh and Liang [59].	Enables stealthy manipulation with a small poisoning budget.
Regression Poisoning	Manipulates continuous-valued regression data to maximize prediction error Oprea et al. [53].	Doubles or triples MSE and affects downstream risk models.
Domain-Specific Poisoning	Uses contextualized attacks targeting healthcare or ICS systems Ma et al. [60].	Causes misdiagnosis or anomaly suppression in 85–90% of industrial monitoring cases.

- High auditability and traceability via centralized logging. Centralized workflows benefit from reproducible pipelines, dataset versioning, and unified audit logs Muñoz-González et al. (2017); Tian et al. [3]. Forensic tools such as gradient fingerprinting and lineage tracing Sun et al. [52]; Gao et al. [28] allow investigators to reconstruct poisoning events. In contrast, FL lacks global visibility, and poisoned updates often blend with benign ones Bhagoji et al. [27]; Bagdasaryan et al. [45].
- Limited adversarial diversity compared with federated learning. Centralized systems involve a single training pipeline, making multi-party collusion and multi-round adaptive poisoning relatively rare. Recent federated attack studies Sikandar et al. [11]; Lianga et al. [12]; Nowroozi et al. [22] demonstrate far greater diversity, including client–server collusion, Byzantine coordination, and cross-round adaptive poisoning.
- Clearer provenance and accountability pathways. Centralized ML benefits from complete ownership over data collection, preprocessing, and training workflows. Accountability tools—dataset fingerprinting, secure provenance tracking, and auditable pipelines Kroll [61]; Miguel and Chen [62]—are easier to enforce. In FL, the absence of unified provenance and verifiable aggregation complicates the attribution of malicious behavior Zhang et al. [37]; Li et al. [14]. These observations highlight that centralized poisoning attacks remain highly damaging, yet comparatively easier to audit and attribute than their federated counterparts. This motivates a separate analysis of poisoning in federated learning, where untrusted participants, opaque aggregation, and cross-round interactions create a significantly broader and more complex threat surface.

3.3. Poisoning Attacks in Federated Learning

Recent findings show that federated learning (FL) introduces structural vulnerabilities that do not arise in centralized machine learning. Adversaries in FL can exploit client–server communication channels, heterogeneous local datasets, and privacy-preserving aggregation protocols to inject, amplify, or conceal poisoning behaviors Sikandar et al. [11]; Lianga et al. [12]; Nowroozi et al. [22]. As recent FL surveys show, adversaries may manipulate local training, bias aggregation rules, or distribute

 doi:10.20944/preprints202605.1674.v1

malicious model updates, thereby significantly expanding the overall attack surface Lianga et al. [12]; Nowroozi et al. [22]; Li et al. [14]. A core challenge is that aggregation in FL acts as a single point of failure. Malicious gradients can easily bias the global model, especially when they fall within the natural statistical variation caused by non-IID client data Bhagoji et al. [27]; Lianga et al. [12]. Non-IID distributions increase divergence among honest client updates, enabling poisoned updates to blend with legitimate outliers and evade detection Cao et al. [63]; Li et al. [14]. Moreover, the privacy–robustness trade-off further complicates defense. FL commonly employs secure aggregation and differential privacy to protect sensitive client information, but these same protections hide update-level anomalies and prevent the server from inspecting individual gradients or weights Bonawitz et al. [26]; Ma et al. [64]. Consequently, many classical anomaly-detection and provenance-based techniques used in centralized learning no longer apply. FL's multi-round training structure also expands the temporal attack surface. Instead of requiring a single high-impact attack, adversaries can inject small, coordinated perturbations across many rounds to carry out slow-drift poisoning, long-term backdoor reinforcement, or collusive update shaping Bagdasaryan et al. [45]; Zhang et al. [37]. These temporal dynamics enable highly stealthy, persistent attacks that accumulate over time without triggering statistical alarms. Together, these challenges help explain why federated learning is significantly harder to secure, audit, and regulate than centralized architectures. Current FL systems lack unified provenance, update lineage, and verifiable aggregation, which limits the ability to perform post-incident attribution or forensic reconstruction. This gap highlights the need for accountability-oriented mechanisms—such as tamper-evident logs, verifiable aggregation proofs, and federated provenance frameworks—to support trustworthy and forensic-ready FL deployments He et al. [19]; Gu et al. [21]; Zeng et al. [16].

3.3.1. Client-Side Poisoning

Client-side poisoning constitutes the most prevalent and well-studied class of adversarial threats in federated learning. Because FL delegates full control of local training to individual participants, any compromised or malicious client can manipulate its data, training procedure, or gradient updates before transmission to the server. This autonomy—combined with the lack of visibility into local computation creates an ideal environment for adversaries to inject harmful behavior without ever accessing the global model directly Bagdasaryan et al. [45]; Zhang et al. [37]. The attack surface therefore includes every component of the local training pipeline, from dataset curation and preprocessing to model optimization and gradient packaging. A large body of work examines data poisoning, in which adversaries tamper with a client's private dataset. Because FL assumes that data remain on-device, the server cannot validate data integrity or detect local manipulation. Malicious clients may alter labels, embed backdoor triggers, insert adversarially crafted samples, or selectively corrupt feature distributions. Dirty-label poisoning—in which both inputs and labels are modified—has emerged as particularly potent in FL because it imposes no constraint on label consistency across clients and can be applied without raising suspicion under non-IID data distributions Xia et al. [13]. Clean-label poisoning is also feasible, especially when adversaries craft feature-space collisions or exploit similarities across client data to achieve targeted misclassification while preserving the overall training-loss profile Gharib et al. [55]. Because the server receives only gradients, rather than raw data, even substantial dataset corruption remains invisible at aggregation time. Beyond data-level threats, model- or gradient-poisoning attacks directly manipulate a client's computed updates. This form of poisoning is highly expressive because the adversary can arbitrarily modify or replace the gradient tensor. One of the most widely studied mechanisms is model replacement, in which the attacker scales a gradient update so that, once aggregated, it overwrites the global model in a single round Bagdasaryan et al. [31]. Other studies highlight gradient-ascent attacks, in which adversarial clients intentionally update the model in the wrong optimization direction, slowing training or inducing divergence Guerraoui et al. [49]. More recent work demonstrates stealth poisoning, in which malicious updates are carefully crafted to mimic benign statistical properties, ensuring that distance-based or clustering-based anomaly detectors fail to identify them Shejwalkar and Houmansadr [48]. These stealth strategies are especially effective under

 doi:10.20944/preprints202605.1674.v1

realistic FL settings, where heterogeneous client data naturally produce high-variance gradients that mask malicious deviations. Overall, client-side poisoning is uniquely powerful because it exploits the fundamental design principles of federated learning—local autonomy, privacy preservation, and decentralized control. Even a small fraction of malicious participants can significantly bias the global model, especially when attackers coordinate their updates or exploit non-IID data distributions Bhagoji et al. [27]. Without mechanisms for provenance tracking, update validation, or verifiable computation, the FL server must assume that incoming gradients faithfully represent honest local training, leaving the system highly susceptible to subtle, persistent, and difficult-to-detect poisoning behavior.

3.3.2. Server-Side and Aggregation Attacks

Although federated learning is commonly portrayed as a client-driven paradigm, the central server remains the most influential component in coordinating training, orchestrating communication, and defining how local updates shape the global model. This makes a compromised or malicious server uniquely powerful: unlike client-side adversaries, who must work within the constraints of their local datasets and limited influence, a malicious server can manipulate the entire training trajectory with unrestricted control Li et al. [14]; Zhang et al. [37]. Server-side poisoning therefore represents a structural vulnerability embedded in the design of FL. One of the most consequential attack vectors in federated learning arises from the server's exclusive authority over update aggregation. The aggregation server functions as a single point of failure, and recent analyses show that an untrusted or malicious server can subtly manipulate updates, bias the global model, or even inject poisoned parameters directly into the aggregation pipeline Li et al. [14]; Zhang et al. [37]. By selectively altering incoming updates, amplifying suspicious gradients, or discarding contributions from honest clients, a compromised server can steer the optimization trajectory with high precision—an ability highlighted in studies of untrusted server behavior in FL Li et al. [14]; ElZemity and Arief [65]. Even minor perturbations to the aggregation rule, such as adjusting client weights or modifying the aggregation function, can produce significant shifts in global behavior, especially under non-IID data, where honest updates naturally diverge Cao et al. [63]; Bhagoji et al. [27]. Because clients lack visibility into aggregation internals and cannot verify how their updates are combined, server manipulations remain invisible, uncontestable, and forensically opaque throughout training Bonawitz et al. [26]; Ma et al. [64]. Beyond aggregation, the server also governs the distribution of global models. A powerful tactic known as model forkingallows the server to send different global models to different subsets of clients. This enables targeted backdoor insertion against specific groups while maintaining apparently benign performance for the rest of the federation Li et al. [14]. Forked models can propagate hidden triggers or biased parameters over multiple rounds without ever appearing in the model version evaluated by other clients, thereby creating a highly asymmetric and difficult-to-detect poisoning channel. Server-side attacks become even more problematic under secure aggregation, where privacy protections prevent the server from inspecting individual client updates but simultaneously give it the freedom to inject arbitrary parameters into the global model. Cryptographic protections designed to safeguard clients therefore do not constrain the server; instead, they remove the client's ability to verify whether the aggregated update is legitimate. As a result, a malicious server can embed backdoor parameters directly into the global model during broadcast, bypassing all client-side defenses and exploiting the assumption that the server behaves honestly ElZemity and Arief [65]. Despite being less studied than client-side poisoning, server-side manipulation is considerably more dangerous and difficult to mitigate. Robust aggregation strategies rely on the assumption that the server is honest; once that assumption is broken, no amount of statistical filtering or majority voting among clients can restore integrity Nowroozi et al. [22]. Server-side poisoning therefore exposes a fundamental limitation of FL: the architecture lacks built-in mechanisms for verifying server behavior, rendering the central coordinator a trusted but unverifiable root of authority.

 doi:10.20944/preprints202605.1674.v1

3.3.3. Collusion and Multi-Round Poisoning

Federated learning's iterative, multi-round structure introduces an additional class of threats that exploit temporal dynamics and distributed control. Unlike one-shot poisoning in centralized learning, adversaries in FL can coordinate across clients, across rounds, or even with the server itself. These distributed and persistent strategies create long-term poisoning trajectories that gradually reshape the global model while evading round-by-round anomaly detection. Collusion among clients is one of the most effective forms of coordinated poisoning. When multiple malicious participants synchronize their updates, they can collectively influence the global model while maintaining statistical similarity to one another. This coordinated behavior undermines robust aggregation methods such as Trimmed Mean, Median, and Krum, which rely on the assumption that malicious updates appear as isolated outliers Fung et al. [66]; Yin et al. [67]. Colluding adversaries can distribute small portions of the malicious perturbation across multiple clients, allowing the aggregated effect to be significant while each individual update remains deceptively benign. Collusion may also occur between clients and the central server. In this scenario, the server amplifies or prioritizes malicious updates, distributes tailored model versions to attacking clients, or subtly adjusts aggregation rules to support the adversaries' objectives Lianga et al. [12]. Such hybrid collusion is especially challenging to detect because it invalidates the foundational trust assumptions of FL: the server is assumed honest, and clients are assumed independent. When these assumptions fail, many FL defenses lose their theoretical guarantees. Multi-round poisoning offers another avenue for stealth and persistence. Instead of injecting a large malicious update at once, adversaries gradually drift the model over many rounds, making each update appear statistically consistent with natural fluctuations caused by non-IID client data. Slow-drift poisoning, adaptive gradient manipulation Zhang et al. [33], and periodic trigger reinforcement enable adversaries to accumulate harmful influence without triggering anomaly detectors that rely on abrupt deviations. The temporal dimension also enables the injection, reinforcement, and preservation of backdoors over time, even if some malicious clients drop out or are intermittently inactive. Sybil attacks further amplify the impact of coordinated poisoning. By creating multiple fake client identities, an adversary can inflate its share of participating clients in each round, overwhelm robust aggregators such as Krum or FoolsGold, and effectively take control of the global model Fung et al. [66]; Cao et al. [68]. These attacks exploit the fact that FL typically lacks strong identity verification, especially in cross-device scenarios with millions of ephemeral participants. Altogether, collusive and multi-round poisoning strategies show how the distributed and iterative nature of FL transforms poisoning from a one-time event into an evolving, long-term adversarial process. The combination of non-IID data, decentralized control, and multi-round training makes coordinated poisoning both highly effective and extremely difficult to detect or attribute without specialized accountability and provenance mechanisms.

Table 5 summarizes the major families of poisoning attacks in federated learning and highlights how their mechanisms, stealth characteristics, and detectability differ across settings. While Section 3.3 describes each attack class in detail, this table provides a consolidated overview that helps illustrate the broader structure of the FL threat landscape. The comparison reflects key insights from recent surveys on poisoning behavior, demonstrating how non-IID data, decentralized control, and multi-round optimization amplify the effectiveness of certain attacks while weakening traditional defenses. By contrasting attack mechanisms with their expected stealth levels and corresponding countermeasures, the table serves as a bridge between the taxonomy presented earlier and the defense strategies analyzed in the subsequent section.

3.3.4. Why Federated Learning Is Harder to Audit

Compared to centralized learning, FL poses major structural barriers to forensic analysis, accountability, and post-incident investigation Sun et al. [52]; Kroll [61].

- Lack of global visibility. The server never sees client data, intermediate states, or, when secure aggregation is used, local gradients, which makes attribution extremely difficult Zhang et al. [37].

 doi:10.20944/preprints202605.1674.v1

- Opaque aggregation. Robust aggregation techniques discard statistical information needed for forensics, while secure aggregation fully hides client updates Ma et al. [64].

Table 5. Comparative analysis of major poisoning attacks in federated learning.

Attack Type	Mechanism	Stealth Level	Detectability	Primary Defense (Examples)
Label flipping	Clients alter labels randomly or target specific classes Paudice et al. [56]; Gharib et al. [55]	Low (large gradient shift)	High under distance metrics	Local/global validation, confusion-matrix checks
Gradient poisoning	Adversarial gradient crafting via bilevel optimization Biggio et al. [2]; Oprea et al. [53]	Medium	Medium (norm and angle metrics)	Krum Blanchard et al. [69], Trimmed Mean Yin et al. [67]
Model replacement	Single-round overwrite of global model via scaled malicious updates Bagdasaryan et al. [31]	Low (extreme deviation)	High (Euclidean distance, cosine)	FLTrust Cao et al. [70], Multi-Krum Blanchard et al. [69]
Backdoor (Visible)	Large, high-frequency or patterned triggers embedded in inputs Gu et al. [57]	Medium	Medium (spectral activation analysis)	Spectral signatures Tran et al. [71]; pruning
Backdoor (Stealthy)	Clean-label triggers or imperceptible perturbations Shafahi et al. [36]; Xia et al. [13]	Very High	Very Low under non-IID noise	Neural Cleanse Wang et al. [72], STRIP Gao et al. [73]

- Untrusted participation. Cross-device FL allows open enrollment, enabling adversaries to create multiple fake clients and amplify their influence through Sybil-style coordination Fung et al. [66]; Cao et al. [63].
- No unified provenance trail. FL often lacks unified dataset lineage, versioning, and update histories, unlike centralized pipelines in which provenance is easier to trace Miguel and Chen [62]. These structural challenges make FL intrinsically harder to secure and audit and motivate the need for accountability-integrated taxonomies and forensic-ready FL architectures, as explored in later sections.

3.4. Accountability-Integrated Taxonomy (AIT)

The Accountability-Integrated Taxonomy (AIT) extends traditional poisoning classifications by highlighting how accountability, attribution, and auditability differ fundamentally between centralized and federated learning. Table 6 operationalizes these dimensions through a concise set of qualitative indicators that can be applied consistently across attack and defense families.

Table 6. Operational indicators used to assess accountability in federated learning defenses.

Indicator	Question answered	Typical evidence or proxy	Interpretation for comparison
Observability	Can poisoning-relevant events be inspected at all?	Visible per-client updates, validation traces, or encrypted-only outputs	Low observability limits anomaly detection and post-incident analysis even if privacy is strong.
Attribution specificity	To what entity can suspicious behavior be linked?	Sample-, client-, round-, or server-level identifiers; committee decisions; signed messages	Higher specificity improves blame assignment and remediation precision.
Aggregation verifiability	Can the claimed aggregate be independently checked?	Commitments, ZK proofs, authenticated transcripts, verifiable computation logs	High verifiability is especially important when the server is not fully trusted.
Tamper evidence	Can later modification or deletion of evidence be detected?	Hash chains, append-only ledgers, signatures, timestamps	Strong tamper evidence improves forensic readiness and institutional trust.
Forensic replayability	Can an incident be reconstructed after the fact?	Versioned models, round metadata, rejection logs, challenge transcripts	Replayability determines whether audits are actionable rather than merely symbolic.
Audit overhead	What extra cost is incurred to preserve evidence?	Communication, computation, storage, and latency overhead	High accountability may be impractical unless the overhead remains deployment-compatible.

 doi:10.20944/preprints202605.1674.v1

In centralized learning, the entire pipeline—including data ingestion, preprocessing, model updates, and training logs—is observable within a single controlled environment. This visibility enables poisoned samples, mislabeled data, or anomalous gradients to be traced back to specific dataset sources or training iterations, a property emphasized in classical poisoning studies Biggio et al. [2]; Jagielski et al. [54]; Koh and Liang [59]. Because centralized learning maintains unified logs, reproducible pipelines, and full parameter visibility, accountability is stronger and forensic reconstruction is generally feasible. Concretely, AIT adds four explicit dimensions to the baseline taxonomy in Section 3.1:

1. Observability of the training process: whether the relevant evidence is fully visible, partially visible, or hidden by decentralization or privacy-preserving mechanisms.
2. Attribution granularity: whether suspicious behavior can be attributed to a sample, a client, a communication round, the server, or only to the federation as a whole.
3. Required audit evidence: which artifacts are needed to support verification, such as dataset provenance, training logs, validation traces, cryptographic commitments, signatures, or zero-knowledge proofs.
4. Trust assumption and control point: whether the dominant threat lies in data contributors, model updaters, the aggregation server, or collusion across these roles. Under AIT, an attack is therefore characterized not only by what is manipulated, but also by what can be seen, who can be blamed, which evidence is needed, and where trust is most brittle. These additional dimensions distinguish the accountability-oriented analysis in this paper from the conventional poisoning taxonomy summarized earlier.

3.4.1. Operational Accountability Indicators

To make accountability analytically usable, we operationalize it through a small set of qualitative but comparable indicators that can be applied across attack and defense families. These indicators do not replace formal benchmarking; rather, they translate otherwise abstract notions such as auditability and traceability into reviewable design questions. In the remainder of the paper, these indicators are used qualitatively to compare defense families and accountability-oriented frameworks. This does not suggest that the field already has universally accepted accountability benchmarks; rather, it provides a transparent vocabulary for comparing methods that are otherwise discussed only descriptively. Federated learning, by contrast, removes these observability guarantees. Local data remain private, client updates are often anonymized, and secure aggregation hides individual gradients from the server and other participants Bonawitz et al. [26]; Ma et al. [64]. As noted in recent FL surveys, this privacy–robustness tension creates an opaque training ecosystem in which neither the server nor the clients can reliably inspect or validate all contributions Sikandar et al. [11]; Lianga et al. [12]. Attacks that leave clear signatures in centralized pipelines—such as label flipping, gradient corruption, or targeted backdoors—become difficult to attribute in FL because the server cannot observe raw data, intermediate features, or per-client loss trajectories Bhagoji et al. [27]. Non-IID data heterogeneity further masks adversarial updates by naturally increasing gradient variance, allowing poisoned updates to blend into benign statistical noise Cao et al. [63]; Li et al. [14]. Additionally, phenomena unique to FL, including model forking, Sybil identities, and cross-round or cross-client collusion, further erode attribution and auditability Fung et al. [66]; Yin et al. [67]; Zhang et al. [33]. Server-side manipulation amplifies these issues: an untrusted aggregator can modify updates, bias aggregation rules, or inject poisoned parameters directly into the global model, and such manipulations may remain invisible to clients Nowroozi et al. [22]; ElZemity and Arief [65]. Consequently, forensic requirements differ substantially between settings. Whereas centralized learning relies on dataset provenance, training logs, and classical statistical analysis, FL systems must incorporate cryptographic proofs, tamper-evident logging, verifiable aggregation, and external auditors to restore accountability Zhang et al. [74]; Li et al. [75]; Ning et al. [76]; Liu et al. [77]. By unifying these observations, AIT provides a framework for evaluating poisoning threats not only by their technical mechanisms,

 doi:10.20944/preprints202605.1674.v1

but also by their accountability exposure, attribution difficulty, auditability requirements, and trust assumptions. This integration bridges Sections 3.2 and 3.3.1, illustrating how the same attack class may be fully traceable in a centralized environment yet practically undetectable in federated settings. AIT therefore supports a holistic assessment of poisoning threats that aligns technical risk analysis with governance, compliance requirements under the GDPR and the EU AI Act, and forensic-readiness demands for high-risk AI systems. To make accountability analytically usable, we operationalize it through a small set of qualitative but comparable indicators that can be applied across attack and defense families. These indicators do not replace formal benchmarking; rather, they translate otherwise abstract notions such as auditability and traceability into reviewable design questions. In the remainder of the paper, these indicators are used qualitatively to compare defense families and accountability-oriented frameworks. This does not suggest that the field already has universally accepted accountability benchmarks; rather, it provides a transparent vocabulary for comparing methods that are otherwise discussed only descriptively. Federated learning, by contrast, removes these observability guarantees. Local data remain private, client updates are often anonymized, and secure aggregation hides individual gradients from the server and other participants Bonawitz et al. [26]; Ma et al. [64]. As noted in recent FL surveys, this privacy–robustness tension creates an opaque training ecosystem in which neither the server nor the clients can reliably inspect or validate all contributions Sikandar et al. [11]; Lianga et al. [12]. Attacks that leave clear signatures in centralized pipelines—such as label flipping, gradient corruption, or targeted backdoors—become difficult to attribute in FL because the server cannot observe raw data, intermediate features, or per-client loss trajectories Bhagoji et al. [27]. Non-IID data heterogeneity further masks adversarial updates by naturally increasing gradient variance, allowing poisoned updates to blend into benign statistical noise Cao et al. [63]; Li et al. [14]. Additionally, phenomena unique to FL, including model forking, Sybil identities, and cross-round or cross-client collusion, further erode attribution and auditability Fung et al. [66]; Yin et al. [67]; Zhang et al. [33]. Server-side manipulation amplifies these issues: an untrusted aggregator can modify updates, bias aggregation rules, or inject poisoned parameters directly into the global model, and such manipulations may remain invisible to clients Nowroozi et al. [22]; ElZemity and Arief [65]. Consequently, forensic requirements differ substantially between settings. Whereas centralized learning relies on dataset provenance, training logs, and classical statistical analysis, FL systems must incorporate cryptographic proofs, tamper-evident logging, verifiable aggregation, and external auditors to restore accountability Zhang et al. [74]; Li et al. [75]; Ning et al. [76]; Liu et al. [77]. By unifying these observations, AIT provides a framework for evaluating poisoning threats not only by their technical mechanisms, but also by their accountability exposure, attribution difficulty, auditability requirements, and trust assumptions. This integration bridges Sections 3.2 and 3.3.1, illustrating how the same attack class may be fully traceable in a centralized environment yet practically undetectable in federated settings. AIT therefore supports a holistic assessment of poisoning threats that aligns technical risk analysis with governance, compliance requirements under the GDPR and the EU AI Act, and forensic-readiness demands for high-risk AI systems.

4. Defenses and Countermeasures in Federated Learning

Federated learning requires multilayered defenses to address diverse poisoning threats arising from decentralized training, untrusted participants, and limited observability. This section integrates category-based defenses, lifecycle-aligned mechanisms, and accountability-centric strategies, offering a unified perspective on how to mitigate adversarial actions across FL's technical, temporal, and governance dimensions.

4.1. Category-Based Defenses in Federated Learning

Federated learning employs a broad spectrum of defense mechanisms to mitigate poisoning attacks, ranging from statistical anomaly detection to cryptographic verification and accountability-oriented aggregation. Because FL decentralizes training and limits global visibility, no single defense is

 doi:10.20944/preprints202605.1674.v1

sufficient. Instead, layered approaches are required to address diverse adversarial strategies Sikandar et al. [11]; Lianga et al. [12]. This subsection organizes defenses into five major categories widely recognized in recent FL surveys: anomaly and statistical detection, performance-based filtering, Byzantine-robust aggregation, cryptographic defenses, and accountability-enabling frameworks.

4.1.1. Anomaly and Statistical Detection

Anomaly-detection methods identify suspicious client updates by analyzing statistical deviations in gradient norms, cosine similarity, weight directions, or parameter trajectories. These methods assume that malicious updates differ significantly from honest ones in magnitude or orientation Cao et al. [70]; Tounsi et al. [78]. Classical metrics include Euclidean distance, cosine similarity, and Pearson correlation, often combined with clustering to separate benign and adversarial groups Li et al. [79]; Ma et al. [28]. More advanced approaches employ deep autoencoders, SVM-based detectors, and GAN-inspired models to learn representations of benign update distributions Chen et al. [80]; Alsulaimawi [81]. While effective in IID settings, these defenses degrade under high non-IID heterogeneity, where benign updates naturally diverge and poisoned updates can blend into statistical noise Cao et al. [63]; Li et al. [14]. Furthermore, secure aggregation obscures individual gradients, reducing the applicability of anomaly-based detection Ma et al. [64].

4.1.2. Performance-Based Filtering

Performance-based defenses evaluate client updates through validation accuracy, loss behavior, or prediction consistency on a trusted validation set Khraisat et al. [82]; Fang et al. [70]. Poorly performing updates are rejected or down-weighted before aggregation. Global validation schemes employ trusted validators or distributed committees to assess aggregated model performance Gambs et al. [83]; Zhang et al. [84]. Although intuitive and easy to implement, these defenses rely on high-quality, representative validation data. Poisoned or biased validation sets can cause honest clients to be misclassified as adversaries, and well-crafted backdoors can preserve clean accuracy while remaining undetected Hakeem and Kim [85].

4.1.3. Byzantine-Robust Aggregation

Byzantine-robust aggregation (BRA) algorithms mitigate the influence of anomalous or adversarial updates during model aggregation Li et al. [86]; Xu et al. [87]. Popular BRA methods include Krum and Multi-Krum, which select updates that are closest to the majority Blanchard et al. [69]; Trimmed Mean and Median, which remove extreme values before averaging Yin et al. [67]; and Bulyan, which combines selection and trimming to improve robustness El Mhamdi et al. [49]. Although these methods are effective against isolated malicious clients, they rely on honest-majority assumptions and degrade under collusive or Sybil attacks Fung et al. [66]. Non-IID data further weakens BRA performance by making honest updates appear inconsistent or adversarial Ashwinee and Natarajan [88]. Beyond robustness- and detection-based approaches, recent work has explored cryptographic and encryption-based mechanisms that provide preventive guarantees against poisoning while enabling verifiable and auditable aggregation.

4.1.4. Cryptographic and Encryption-Based Defenses

Cryptographic and encryption-based defenses aim to prevent or constrain poisoning attacks by enforcing the integrity, authenticity, and verifiability of client updates under explicit trust and threat models. Unlike robustness- or detection-based approaches that operate after aggregation, these mechanisms provide preventive guarantees by restricting the space of admissible updates, authenticating eligible contributors, and enabling verifiable aggregation in privacy-preserving settings. From a poisoning-defense perspective, these schemes intervene at four concrete control points. First, client authenticity and enrollment integrity rely on auditable participant selection, authenticated identities, or signed channels to reduce impersonation, repudiation, and Sybil-style entry points Zeng et al. [16]; Chen et al. [13]. Second, update integrity and provenance rely on signatures, commitments,

 doi:10.20944/preprints202605.1674.v1

or traceable metadata to bind each submission to a round and preserve verifiable lineage Gu et al. [21]; Chen et al. [13]. Third, malicious-secure aggregation protocols impose consistency checks or admissibility constraints before aggregation, thereby limiting what an adaptive adversary or untrusted server can inject into the protocol Rathee et al. [89]; Ma et al. [90]; Lycklama et al. [91]; Jiang et al. [92]. Fourth, verifiable aggregation protocols generate proof transcripts that allow the claimed global update to be checked after the fact without disclosing private local updates Wang et al. [17]; Zhu et al. [18]; Ma et al. [93]; Xu et al. [94]; He et al. [19]. Plain secure aggregation primarily protects confidentiality; by itself, it does not automatically detect or prevent poisoning. Its relevance to poisoning defense emerges when it is combined with malicious-secure consistency checks, authenticated shares, commitment schemes, or MPC-based admissibility enforcement. In this stronger form, recent protocols extend secure aggregation beyond honest-but-curious assumptions and constrain adaptive poisoning attempts even when some participants or the server behave adversarially Rathee et al. [89]; Ma et al. [90]; Lycklama et al. [91]; Jiang et al. [92]. Homomorphic encryption and MPC further enable computation directly over encrypted updates, allowing aggregation without revealing individual contributions, albeit at the cost of higher computational and communication overhead Liu et al. [87]. Beyond confidentiality, recent work has emphasized verifiable aggregation as a means of supporting accountability and auditability in federated learning. Zero-knowledge proofs (ZKPs) and related cryptographic techniques allow an aggregator to prove that the reported global model update is a correct function of client submissions without disclosing the updates themselves. Such mechanisms Cryptographic and encryption-based defenses aim to prevent or constrain poisoning attacks by enforcing the integrity, authenticity, and verifiability of client updates under explicit trust and threat models. Unlike robustness- or detection-based approaches that operate after aggregation, these mechanisms provide preventive guarantees by restricting the space of admissible updates, authenticating eligible contributors, and enabling verifiable aggregation in privacy-preserving settings. From a poisoning-defense perspective, these schemes intervene at four concrete control points. First, client authenticity and enrollment integrity rely on auditable participant selection, authenticated identities, or signed channels to reduce impersonation, repudiation, and Sybil-style entry points Zeng et al. [16]; Chen et al. [13]. Second, update integrity and provenance rely on signatures, commitments, or traceable metadata to bind each submission to a round and preserve verifiable lineage Gu et al. [21]; Chen et al. [13]. Third, malicious-secure aggregation protocols impose consistency checks or admissibility constraints before aggregation, thereby limiting what an adaptive adversary or untrusted server can inject into the protocol Rathee et al. [89]; Ma et al. [90]; Lycklama et al. [91]; Jiang et al. [92]. Fourth, verifiable aggregation protocols generate proof transcripts that allow the claimed global update to be checked after the fact without disclosing private local updates Wang et al. [17]; Zhu et al. [18]; Ma et al. [93]; Xu et al. [94]; He et al. [19]. Plain secure aggregation primarily protects confidentiality; by itself, it does not automatically detect or prevent poisoning. Its relevance to poisoning defense emerges when it is combined with malicious-secure consistency checks, authenticated shares, commitment schemes, or MPC-based admissibility enforcement. In this stronger form, recent protocols extend secure aggregation beyond honest-but-curious assumptions and constrain adaptive poisoning attempts even when some participants or the server behave adversarially Rathee et al. [89]; Ma et al. [90]; Lycklama et al. [91]; Jiang et al. [92]. Homomorphic encryption and MPC further enable computation directly over encrypted updates, allowing aggregation without revealing individual contributions, albeit at the cost of higher computational and communication overhead Liu et al. [87]. Beyond confidentiality, recent work has emphasized verifiable aggregation as a means of supporting accountability and auditability in federated learning. Such mechanisms support third-party verification, post-training audits, and compliance checks in high-stakes or regulated deployments Ning et al. [76]; Liu et al. [77]; Ma et al. [93]; Xu et al. [94].

Table 7 summarizes representative cryptographic defenses and supporting mechanisms published from 2023 to 2025 and highlights how they operationalize accountability, auditability, and poisoning prevention in federated learning. From a forensic-readiness perspective, verification transcripts,

 doi:10.20944/preprints202605.1674.v1

commitments, authenticated aggregation proofs, and round receipts can serve as audit evidence, thereby supporting traceability and post-incident investigation while preserving data privacy. Despite their strong guarantees, cryptographic defenses introduce nontrivial trade-offs. Protocol complexity, computational overhead, and communication costs can limit scalability in large-scale or resource-constrained settings. In addition, the opacity introduced by encryption can hinder fine-grained anomaly detection and statistical inspection, highlighting a fundamental tension between privacy, robustness, and auditability. Consequently, cryptographic mechanisms are best viewed as complements to robust aggregation and detection-based defenses rather than as standalone solutions, making them a critical technical pillar for accountable and trustworthy federated learning.

Table 7. Recent (2023–2025) research on cryptographic, integrity-preserving, and verifiable federated learning that supports accountability, auditability, traceability, and poisoning resilience.

Theme	Representative work (2023–2025)	Relevance to accountability, auditability, and traceability
Verifiable aggregation/integrity	Ma et al. [93]; Xu et al. [94]	Enables verifiable aggregation through cryptographic proofs, such as zero-knowledge or verifiable computation techniques, allowing the server or external auditors to verify aggregation correctness without revealing individual client updates. This directly strengthens auditability in accountable federated learning.
Malicious-secure aggregation (preventive poisoning resilience)	Jiang et al. [92]; Rathee et al. [89]; Ma et al. [90]; Lycklama et al. [91]	Provides cryptographic and MPC-based aggregation protocols that tolerate malicious clients and untrusted servers by enforcing update integrity and admissibility constraints before aggregation, thereby limiting adaptive poisoning attempts and strengthening accountability guarantees.
Client authenticity/participant integrity	Zeng et al. [16]; Chen et al. [13]	Uses auditable participant selection, identity-binding signatures, or traceable communication channels to reduce impersonation, repudiation, and Sybil-style poisoning entry points before model aggregation begins.
Update provenance/round traceability	Gu et al. [21]; Chen et al. [13]	Preserves round-level lineage, commitments, or provenance metadata so that suspicious updates can be linked to a participant, a communication round, and an evidence trail for later audit or incident response.
Secure aggregation with accountability support	Bouamama et al. [95]; Bottoni et al. [96]	Integrates secure aggregation with verification artifacts such as commitments, authenticated logs, or proofs that can serve as forensic evidence, thereby enabling traceability and post-hoc audits while preserving client privacy.

4.1.5. Accountability-Oriented Frameworks

Accountability-driven defenses introduce transparent and verifiable mechanisms for provenance tracking, tamper-evident aggregation, and forensic readiness. Frameworks such as BVDFed and CAFCOR incorporate validation proofs, hash-chained logs, and audit-ledger integration into the aggregation pipeline Zhang et al. [74]; Li et al. [75]. Blockchain-based FL systems record model updates and coordination decisions in immutable ledgers, enabling traceability and post-incident investigation. Digital signatures and cryptographic identity binding prevent impersonation and support attribution during malicious-activity analysis. Although these defenses significantly enhance accountability, they also introduce computational, storage, and governance overhead. They should therefore be treated as evidence-enabling complements to robust aggregation and validation rather than as complete defenses on their own.

While substantial progress has been made in understanding and mitigating poisoning attacks in both centralized and federated learning, significant gaps remain across robustness, privacy, and accountability. Current defenses often operate in isolation, struggle under adaptive adversaries, and lack the forensic capabilities required for regulatory compliance. This section synthesizes recurring limitations in existing approaches, structures the main design trade-offs explicitly, and outlines promising research directions needed to achieve trustworthy, transparent, and attack-resilient learning systems. The literature reviewed above suggests that defense quality in FL is strongly scenario-dependent. A method that performs well when per-client updates are visible may fail once secure aggregation is enabled; a mechanism that is robust against isolated malicious clients may be ineffective when the server is untrusted; and a framework that improves auditability may do little to stop poisoning unless it is paired with stronger aggregation or validation controls. Table 8 summarizes these trade-offs qualitatively so that the review can support design decisions rather than merely list methods. Three

 doi:10.20944/preprints202605.1674.v1

analytical conclusions follow from this comparison. First, no reviewed defense family dominates simultaneously across non-IID heterogeneity, encrypted aggregation, malicious-server settings, and collusion. Second, methods that maximize privacy often reduce observability, which directly weakens anomaly detection and forensic replayability. Third, accountability-supporting mechanisms improve evidence quality and trustworthiness, but they should be understood as complements to robustness rather than substitutes for it.

Table 8. Comparative analysis of defense families under challenging federated learning scenarios.

Defense family	Non-IID	Secure agg.	Server distrust	Collusive attackers	Audit artifact	Decision-support interpretation
Anomaly/statistical detection	Low–Medium	Low	Low	Low–Medium	Suspicion scores, cluster traces, rejected-update logs	Useful when per-client updates are visible and heterogeneity is moderate; it degrades when benign updates naturally diverge or when secure aggregation hides local updates.
Performance-based filtering	Medium	Low–Medium	Low	Low–Medium	Validation outcomes, committee decisions, and rejection records	Practical when trusted validation data exist, but vulnerable to poisoned validation sets and stealthy attacks that preserve clean accuracy.
Byzantine-robust aggregation	Low–Medium	Medium	Low	Low	Aggregation trace, selected-update metadata	Effective against isolated Byzantine clients under honest-majority assumptions, but weak against Sybils, coordinated collusion, and malicious aggregation logic.
Verifiable crypto defenses	High	High	High	Medium	Proofs, signed transcripts, and verification logs	Best suited to settings where server distrust, auditability, or compliance dominate; stronger integrity guarantees come with higher communication and computation overhead.
Audit-oriented frameworks	Medium	Medium	Medium–High	Medium	Tamper-evident logs, signatures, and provenance ledgers	Most valuable as an orchestration layer for evidence and governance; they improve attribution and auditability but must be combined with robust detection or aggregation to block poisoning in real time.

4.2. Structured Trade-off Analysis

The defense landscape reviewed in Section 4.2 reveals three recurring trade-offs that should guide future FL design.

Accuracy/robustness versus privacy. Detection-based and validation-based defenses benefit from visibility into per-client behavior, but secure aggregation and strong privacy controls reduce this visibility. As a result, privacy-preserving deployments frequently lose the very signals that help identify poisoning attempts. Auditability versus scalability. Evidence-rich designs such as authenticated logging, verifiable aggregation, and proof-carrying updates improve accountability, but they increase communication, storage, and verification overhead. This makes them attractive for cross-silo or high-stakes settings, yet potentially burdensome in large-scale cross-device FL. Security coverage versus deployment realism. Many defenses are evaluated against isolated malicious clients, whereas real deployments may face heterogeneous data, adaptive collusion, or partially trusted infrastructure. A defense that is theoretically strong under one threat model may therefore provide limited assurance under more realistic assumptions. These trade-offs imply that FL defenses should not be compared only by attack-success reduction. They should also be compared by what evidence they preserve, what trust assumptions they require, and which deployment constraints they impose.

4.3. Persistent Gaps

Although substantial progress has been made in understanding and mitigating poisoning attacks in both centralized and federated learning, the current defense landscape remains fragmented and

 doi:10.20944/preprints202605.1674.v1

incomplete. Existing mechanisms often address robustness or privacy in isolation, overlooking how adversaries adapt across rounds, persist through multi-step strategies, or exploit the limited auditability inherent in decentralized training environments. Federated learning further introduces structural blind spots—such as untrusted aggregation, severe non-IID heterogeneity, and restricted visibility into client behavior—that exacerbate these vulnerabilities and make attribution especially challenging Bhagoji et al. [97]; Nowroozi et al. [98]. As highlighted in this survey, most defenses lack the ability to provide verifiable provenance, tamper-evident aggregation, or post-incident forensic reconstruction, leaving systems exposed to opaque, stealthy, and collusive attacks that evade detection even under robust aggregation Lianga et al. [99]; Shejwalkar and Houmansadr [100]. These gaps collectively hinder the development of trustworthy, accountable, and resilient machine intelligence systems. The following subsections synthesize the most pressing unresolved challenges, spanning technical, structural, and governance-oriented dimensions.

4.3.1. Poisoning Attack Challenges

Although a wide range of defenses has been proposed, several fundamental weaknesses remain in current poisoning mitigation strategies:

1. Adaptive and Byzantine attackers: most aggregation-based defenses assume static attackers with limited capabilities and that malicious clients constitute only a small minority. Recent studies challenge this assumption by highlighting adaptive poisoning attacks that dynamically adjust their behavior to mimic benign updates or modulate gradient magnitudes to evade statistical filters Chen et al. [101]; Mohamed et al. [102]. This adaptiveness renders fixed-threshold and static rule-based defenses insufficient.
2. Persistent backdoor injection: backdoor attacks remain one of the most enduring threats in federated learning. Malicious clients can embed stealthy triggers in rounds with minimal impact on global accuracy, allowing persistent targeted manipulation Masunda and Ajayi [103]. Their stealthy nature makes them difficult to detect, especially under secure aggregation or encrypted updates.
3. Data heterogeneity exploitation: highly heterogeneous (non-IID) client data distributions enable adversaries to craft model updates that appear statistically consistent with legitimate local training Tallam [104]; Zhukabayeva et al. [105]. This undermines anomaly-detection techniques that assume homogeneous or centrally accessible data, leading to high false-positive rates or undetected attacks.
4. Limited global observability: privacy constraints restrict the central server from accessing client data or intermediate training signals, thereby weakening validation- and verification-based defenses Chen et al. [101]. This creates a structural detection blind spot, particularly for data-centric poisoning attacks executed during the training phase.
5. Untrusted server and aggregation manipulation: most poisoning defenses implicitly assume an honest server, yet recent analyses show that a compromised or malicious aggregator can bias model updates, alter aggregation weights, or fork global models without client visibility Nowroozi et al. [98]; Lianga et al. [99]. Because clients cannot inspect server-side operations, such manipulations remain invisible, creating a critical accountability gap in FL.
6. Secure aggregation as a double-edged sword: while secure aggregation preserves client privacy, it also prevents the server from inspecting individual gradients or update statistics, eliminating many anomaly-based defenses and limiting post-incident forensics Ma et al. [64]; Sikandar et al. [11]. This creates structural blind spots in which data poisoning, backdoors, and collusive behaviors can operate undetected.
7. Lack of standardized forensic and robustness benchmarks: FL poisoning research relies heavily on small image datasets (e.g., MNIST and CIFAR-10) and lacks standardized evaluation protocols that capture realistic heterogeneity, collusion, or accountability requirements Tian et al. [3]; Lianga

 doi:10.20944/preprints202605.1674.v1

et al. [12]. Without benchmarks that incorporate provenance, traceability, and verifiable auditing, comparing defenses or assessing forensic readiness remains difficult.

4.3.2. Anomaly Detection and Traceability

Anomaly detection plays a dual role in FL by improving robustness and enabling post-hoc forensic accountability.

1. Enabling traceability: mechanisms such as gradient fingerprinting, temporal consistency analysis, and inter-round similarity metrics help identify poisoned updates and attribute them to specific clients or rounds Nguyen [106]; Tallam [104]. These capabilities provide a foundation for forensic readiness and auditability in FL.
2. Enhancing explainability: modern anomaly-detection frameworks increasingly incorporate attention mechanisms or other explainable neural components, offering interpretable insights into why updates are flagged as malicious Hamouda [107]; Shaik et al. [108]. Such transparency is essential for trust, human oversight, and alignment with ethical AI principles.
3. Forensic analysis in federated learning: emerging research integrates anomaly detection directly into forensic-analysis pipelines, enabling the reconstruction of poisoning behavior and supporting proactive security strategies Mohamed et al. [102]. This shift from reactive detection to proactive attribution is critical to next-generation accountable FL systems.

4.4. Opportunities for Unified Frameworks

The evolution of federated learning (FL) demands architectural frameworks that jointly optimize privacy, robustness, and accountability rather than treating these guarantees as independent or conflicting objectives. Current research remains fragmented: privacy-preserving mechanisms such as differential privacy (DP) and secure aggregation protect confidentiality but provide limited support for verifiable training or forensic traceability [109,110,111]. Conversely, robust aggregation strategies such as CAFCOR and BVDFed improve resilience to Byzantine and poisoning attacks [75,74] yet often assume semi-honest servers and lack transparent audit mechanisms. Accountability-enhancing tools—such as verifiable computation or blockchain-secured logging [76,77,112]—offer traceability but remain largely decoupled from privacy and robustness objectives. For that reason, we do not claim a fully specified new FL architecture here. Rather, we distill a conceptual design space that future FL systems must satisfy if they are to be simultaneously private, robust, and auditable. Four design requirements emerge from the literature reviewed in this paper:

1. Privacy-preserving observability: systems need minimal but verifiable evidence channels so that poisoning-relevant events can be inspected without exposing raw client data.
2. Evidence-carrying aggregation: aggregation should produce auditable artifacts—for example commitments, signatures, or proofs—rather than only a model output.
3. Attribution across trust boundaries: the design must distinguish client misconduct, server misconduct, and collusion, since each requires different evidence and response mechanisms.
4. Standards-mappable assurance: technical artifacts should be expressible as governance evidence that can support internal audits, external assurance, or compliance review. The interaction between these requirements is illustrated in Figure 2. The three layers should be read as a design decomposition rather than a claim of implementation completeness. At the privacy-preserving layer, techniques such as differential privacy, homomorphic encryption, and secure multiparty computation protect client updates from inference or leakage. The robustness layer incorporates Byzantine-resilient aggregation, anomaly detection, and validation-based filtering to preserve model integrity against poisoning and backdoor attacks [113,114,115]. Above these, the accountability layer introduces tamper-evident audit trails using cryptographic commitments, hash-chained logs, identity-binding signatures, or zero-knowledge proofs to enable verifiable attribution and post-hoc forensic reconstruction [116,117,118]. This layered view also helps deepen the paper's brief mapping to ISO/IEC 42001. At the governance level, FL systems need explicit

 doi:10.20944/preprints202605.1674.v1

role definitions, risk ownership, and incident-handling rules. At the operational level, they need evidence about sampling, aggregation, validation, and model release decisions. At the assurance level, they need artifacts that an independent auditor could inspect without re-running the entire training process. In other words, standards alignment is meaningful only if technical defenses emit reviewable evidence, not merely if they improve accuracy under attack. Future work should therefore formalize composable trust objectives that jointly quantify privacy budgets, robustness guarantees, and auditing costs under unified optimization criteria. Promising directions include: (1) developing quantitative metrics that capture privacy–robustness–accountability interactions; (2) constructing benchmark suites for untrusted-server and multi-round poisoning scenarios; (3) designing hybrid cryptographic–forensic frameworks that combine zero-knowledge verification with blockchain-secured aggregation; and (4) evaluating which evidence artifacts are actually useful to human auditors, regulators, and incident responders. Such integrated approaches would enable the emergence of trustworthy federated intelligence, a next-generation paradigm where security, transparency, and regulatory compliance coexist more coherently.

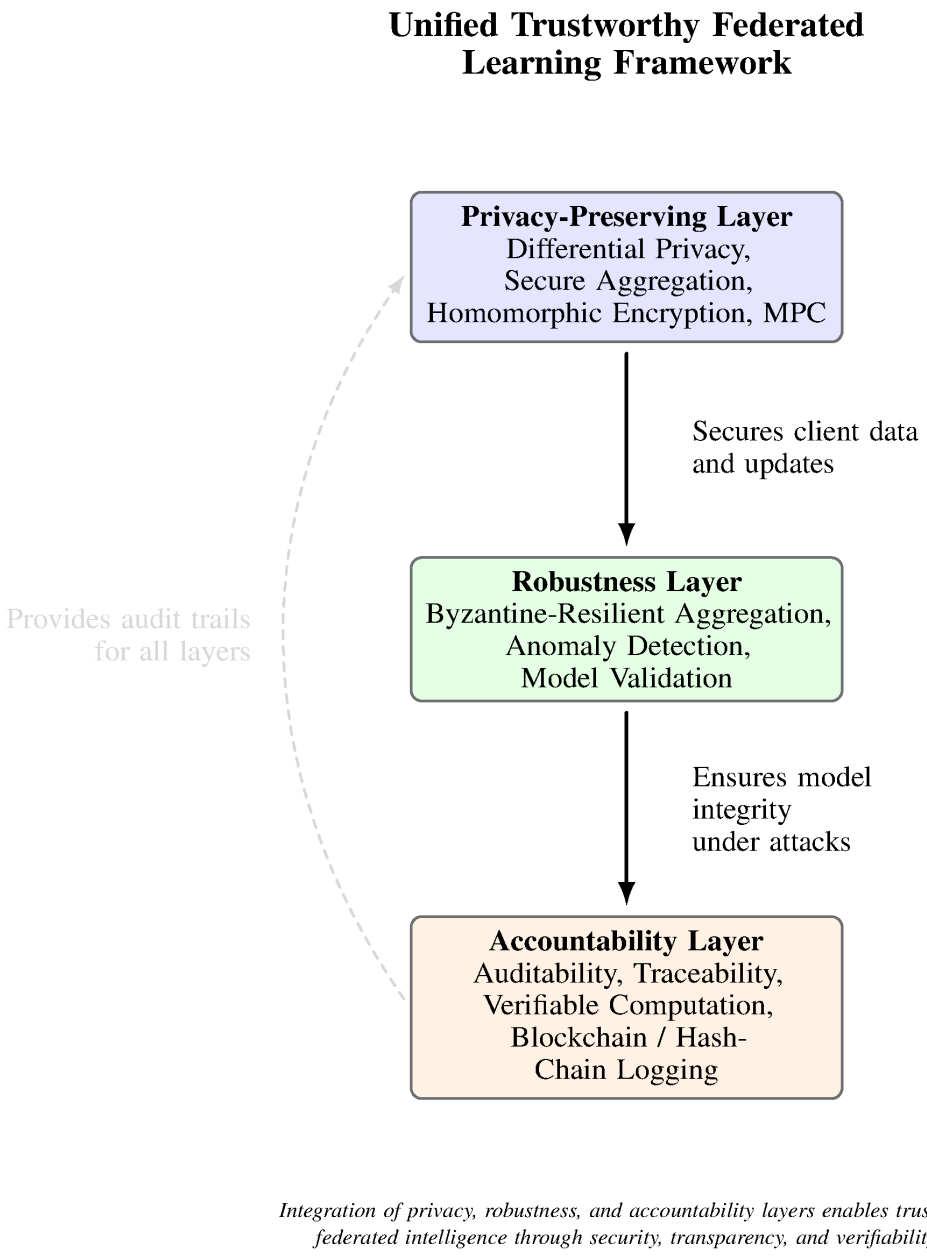

Figure 2. Conceptual architecture for a unified trustworthy federated learning framework integrating privacy, robustness, and accountability layers.

 doi:10.20944/preprints202605.1674.v1

5. Implications for Future Research and Practice

The increasing complexity of federated learning (FL) systems underscores the need for governance and benchmarking frameworks that connect scientific research, industrial use, and emerging policy requirements. The implications of this study span two complementary domains: advancing scientific inquiry and strengthening applied governance.

For researchers, this survey highlights the absence of standardized evaluation protocols that jointly assess privacy, robustness, and accountability. Key research priorities include:

- Developing auditable benchmark suites that model adversarial clients, collusion, and untrusted-server behavior under realistic deployment constraints.
- Embedding forensic auditability and explainability into model evaluation pipelines to support reproducibility in high-stakes or regulated settings.
- Creating open datasets containing poisoning traces, backdoor triggers, and unlearning metadata to accelerate the validation of accountable and verifiable defenses.

For industry and policymakers, the findings align with the global move toward Responsible AI and compliance-by-design paradigms. Future FL deployments should:

- Integrate tamper-evident audit logs and forensic accountability mechanisms to support transparent monitoring across the model lifecycle.
- Map technical safeguards to governance standards such as ISO/IEC 42001:2023 and the NIST AI Risk Management Framework, enabling lifecycle traceability and verifiable compliance.
- Adopt federated explainable AI (fXAI) components, ensuring interpretable reasoning pathways for regulators, auditors, and end-users.

Stronger collaboration among academia, industry, and standardization bodies will be essential to achieve a globally interoperable ecosystem for accountable and trustworthy federated learning.

Future Directions.

1. Federated forensics by design: integrating neuron- and feature-level provenance with cryptographic round receipts to enable explainable reconstruction of poisoning and backdoor behavior.
2. Privacy-preserving auditability: using zero-knowledge–attested robust aggregation and verifiable client sampling to prevent selective aggregation, tampering, and forgery while maintaining confidentiality.
3. Standards alignment: mapping FL artifacts—hashes, provenance metadata, policies, and KPIs—to ISO/IEC 42001 controls, producing machine-verifiable risk and compliance evidence that regulators can query without accessing sensitive data.

Table 9 synthesizes key research efforts on accountability, auditability, and traceability in federated learning. These works span blockchain-backed audit trails, zero-knowledge–based verification, provenance tracking, explainability-driven debugging, and regulatory audit frameworks, collectively illustrating the emerging landscape of accountable FL.

 doi:10.20944/preprints202605.1674.v1

Table 9. Recent research in accountability and traceability in federated learning.

Title	Focus Area
ISO/IEC 42001: Artificial Intelligence Management Systems International Organization for Standardization [119]	Lifecycle governance, auditability, accountability
Blockchain-Based Federated Learning: A Survey and New Perspectives Ning et al. [120]	Auditing, traceability, taxonomy in BCFL
A Comprehensive Survey on Blockchain-based FL Liu et al. [50]	Security and audit frameworks for BCFL
Auditable FL with Blockchain-Based Participant Selection Zeng et al. [16]	Auditable sampling, ledgered evidence
zkFL: ZKP-based Gradient Aggregation for FL Wang et al. [17]	Verifiable aggregation, privacy
Trusted Model Aggregation with Zero-Knowledge Proofs Ma et al. [121]	ZK proofs for trusted aggregation
RiseFL: Secure and Verifiable Data Collaboration with Low-Cost ZKPs Zhu et al. [18]	Low-cost ZK proofs, verifiable training
Publicly Auditable and Privacy-Preserving FL He et al. [19]	Public auditability with robust aggregation
TraceFL: Interpretability-Driven Debugging in FL Gill et al. [122]	Neuron-level provenance, debugging
Enhancing Data Provenance & Transparency in FL Gu et al. [21]	Provenance tracking, reproducibility
PPTFL: Privacy-Preserving and Traceable FL Chen et al. [13]	Traceability with privacy preservation
Interpretable/Explainable FL Survey Li et al. [80]	Explainability, fXAI taxonomy

Federated learning is entering a new phase in which privacy preservation alone is insufficient. As poisoning threats become more adaptive, decentralized, and opaque, future FL systems must integrate verifiable governance, forensic accountability, and explainable anomaly detection into their core architecture. This review does not argue that cryptographic accountability mechanisms are novel in isolation; rather, it shows that verifiable aggregation, provenance, and governance evidence should be treated as a first-class defense category alongside robust aggregation and anomaly detection. This paper therefore highlights how privacy, robustness, and accountability—often treated as separate objectives—can be jointly addressed through unified, multi-layered designs that provide both technical resilience and regulatory transparency. Building on the insights from this survey, several research directions emerge:

- ZK-proof integration: improving the efficiency and deployment realism of zero-knowledge proofs in aggregation workflows so that update correctness can be certified without exposing sensitive data.
- Explainable accountability: evaluating whether interpretable anomaly indicators and provenance signals can serve as reliable, audit-ready evidence in regulated environments.
- Lifecycle governance: operationalizing ISO/IEC 42001 and similar standards through machine-verifiable artifacts that document risk management, training policies, and compliance events.
- Forensic optimization: identifying computationally efficient strategies for tamper-evident logging, dynamic audit triggering, and cryptographic verification that preserve model convergence.
- Human–AI oversight: developing visualization and review tools to support investigators, auditors, and domain experts in understanding attack patterns and audit trails in cross-silo deployments. Advancing these directions will require collaboration across machine learning, cybersecurity, distributed systems, and regulatory domains. Establishing reproducible forensic benchmarks and standard metrics for trust will accelerate progress toward FL systems that are not only secure and privacy-preserving, but also transparent, verifiable, and aligned with global Responsible AI requirements.

 doi:10.20944/preprints202605.1674.v1

Author Contributions: S.M. conceived the study, conducted the literature review, and drafted the manuscript. D.A. contributed to the design of the research framework, supervised the methodological development, and provided critical revisions. A.N. contributed to conceptualization, supervised the general research direction, and reviewed the manuscript for intellectual content. All authors contributed to the revision of the manuscript, read, and approved the submitted version.

Conflicts of Interest: The authors declare that the research was conducted in the absence of commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. Nature Machine Intelligence 1, 389–3999
2. Biggio, B., Nelson, B., and Laskov, P. (2012). Poisoning attacks against support vector machines. Proceedings of the 29th International Conference on Machine Learning (ICML)
3. Tian, Z.; Cui, L.; Liang, J.; Yu, S. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Comput. Surv.* 2022, *55*, 1–35. doi:10.1145/3551636
4. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society 3, 20539517166796791
5. Novelli, C., Taddeo, M., and Floridi, L. (2024). Accountability in artificial intelligence: what it is and how it works. Ai & Society 39, 1871–1882
6. Miguel, B. S., Naseer, A., and Inakoshi, H. (2021a). Putting accountability of AI systems into practice. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. 5276–5278.
7. Cen, S. and Alur, R. (2024a). From transparency to accountability and back: A discussion of access and evidence in ai auditing. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. doi:10.1145/3689904.369471
8. Kroll, J. A. (2021b). Outlining traceability: A principle for operationalizing accountability in computing systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (ACM), 21–31. doi:10.1145/3442188.344593
9. Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E., and Loukas, G. (2019). A taxonomy and survey of attacks against machine learning. Computer Science Review 34, 1001991
10. Cin'a, A. E., Demontis, A., Biggio, B., Pelillo, M., and Roli, F. (2023). Wild patterns reloaded: A survey of machine learning security against training data poisoning. ACM Computing Surveys 55, 1–39. doi: 10.1145/358538
11. Sikandar, H., Waheed, H., Tahir, S., and Malik, S. (2023). A detailed survey on federated learning attacks and defenses. Electronics 12, 2601076
12. Lianga, J., Wang, R., Feng, C., and Chang, C.-C. (2023a). A survey on federated learning poisoning attacks and defenses. arXiv preprint arXiv:2306.03397
13. Xia, G., Chen, J., Yu, C., and Ma, J. (2023). Poisoning attacks in federated learning: A survey. IEEE Access 11, 12345–1236
14. Li, J. et al. (2025). Threats and defenses in the federated learning life cycle: A comprehensive survey and challenges. Frontiers in AI
15. Zhou, Y. et al. (2025). Defending against data poisoning attacks in federated learning: A survey. ACM Computing Surveys
16. Zeng, H. et al. (2024a). A federated learning framework with blockchain-based auditable participant selection. Journal of Information Security and ApplicationsIn press
17. Wang, Z. et al. (2023). zkfl: Zero-knowledge proof-based gradient aggregation for federated learning.1098 arXiv:2310.02554
18. Zhu, Y. et al. (2024). Secure and verifiable data collaboration with low-cost zero-knowledge proofs.1144 PVLDB 17, 2321–2334. doi:10.14778/3665844.3665860
19. He, X. et al. (2024). Enabling privacy-preserving and publicly auditable federated learning. arXiv preprint arXiv:2405.04029
20. Chen, J. et al. (2023). Privacy-preserving and traceable federated learning for industrial iot. Expert Systems with Applicationsdoi:10.1016/j.eswa.2023.xxxxxx

 doi:10.20944/preprints202605.1674.v1

21. Gu, M. et al. (2024). Enhancing data provenance and model transparency in federated learning. arXiv preprint arXiv:2403.01451
22. Nowroozi, E., Haider, I., and Taheri, R. (2025a). Federated learning under attack: Exposing vulnerabilities through data poisoning attacks in computer networks. IEEE Transactions on Information Forensics and Security
23. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., and Song, D. (2020b). The secret revealer: Generative model-inversion attacks against deep neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 253–2611
24. Srivastava, M., Kaushik, A., Loughran, R., and McDaid, K. (2024). Data poisoning attacks in the training phase of machine learning models: A review
25. Biggio, B., Nelson, B., and Laskov, P. (2011). Support vector machines under adversarial label noise. In Asian conference on machine learning(PMLR), 97–1128
26. Bonawitz, K., Kairouz, P., McMahan, B., and Ramage, D. (2021). Federated learning and privacy: Building privacy-preserving systems for machine learning and data science on decentralized data. Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. doi:10.1145/3494834. 3500240
27. Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. (2019a). Analyzing federated learning through an adversarial lens. In International conference on machine learning(PMLR), 634–6438
28. Gao, Y., Doan, B. G., Zhang, Z., Ma, S., Zhang, J., Fu, A., et al. (2020). Backdoor attacks and countermeasures on deep learning: A comprehensive review. arXiv preprint arXiv:2007.10760
29. Li, Y., Jiang, Y., Li, Z., and Xia, S.-T. (2022). Backdoor learning: A survey. IEEE transactions on neural networks and learning systems 35, 5–2298
30. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia conference on computer and communications security. 506–5191
31. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. (2020a). How to backdoor federated learning. arXiv preprint arXiv:1807.00459
32. Mazzone, F., Badawi, A. A., Polyakov, Y., and Everts, M. (2024). Investigating privacy attacks in the gray-box setting to enhance collaborative learning schemes.
33. Zhang, Y., Bai, G., Chamikara, M., and Ma, M. (2023c). Agrevader: Poisoning membership inference against byzantine-robust federated learning. Proceedings of the ACM Asia Conference on Computer and Communications Security. doi:10.1145/3543507.3583542
34. Sakhnovych, Y. (2024). Black-box Model Watermarking in Federated Learning. Ph. D. thesis, TU Wien
35. Wang, Z., Ma, J., Wang, X., Hu, J., Qin, Z., and Ren, K. (2022). Threats to training: A survey of poisoning attacks and defenses on machine learning systems. ACM Computing Surveysdoi:10.1145/3538707
36. Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., et al. (2018). Poison frogs! targeted clean-label poisoning attacks on neural networks. Advances in Neural Information Processing Systems
37. Zhang, R., Hussain, S., Chen, H., and Javaheripi, M. (2023b). Systemization of knowledge: robust deep learning using hardware-software co-design in centralized and federated settings. ACM Computing Surveysdoi:10.1145/3616868
38. Shah, A., Ahmad, A., Ali, B., and Anwer, S. (2025). Guarding the gates: A comprehensive survey of backdoor attacks on neural networks
39. Zhang, J., Chen, B., Cheng, X., Binh, H. T. T., and Yu, S. (2020a). Poisongan: Generative poisoning attacks against federated learning in edge computing systems. IEEE Internet of Things Journal 8,1126 3310–3322
40. Weng, C.-H., Lee, Y.-T., and Wu, S.-H. (2020). On the trade-off between adversarial and backdoor robustness. Advances in Neural Information Processing Systems 33 (NeurIPS)1101
41. Pan, M., Zeng, Y., Lyu, L., Lin, X., and Jia, R. (2023). Asset: Robust backdoor data detection across a multiplicity of deep learning paradigms. Proceedings of the 32nd USENIX Security Symposium
42. Xu, C., Liu, W., Zheng, Y., Wang, S., and Chang, C.-H. (2024). An imperceptible data augmentation based blackbox clean-label backdoor attack on deep neural networks. IEEE Transactions on Circuits and Systems I: Regular Papers. doi:10.1109/TCSI.2024.3365130
43. De Gaspari, F., Hitaj, D., and Mancini, L. V. (2024). Have you poisoned my data? defending neural networks against data poisoning. In Computer Security – ESORICS 2024(Springer). 85–104. doi:10.1007/978-3-031-70879-4

 doi:10.20944/preprints202605.1674.v1

44. Bena, N., Anisetti, M., Damiani, E., Yeun, C. Y., and Ardagna, C. A. (2025). Protecting machine learning from poisoning attacks: A risk-based approach. Computers & Security 155, 104468. doi:10.1016/j.cose.2025.104468
45. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. (2020b). How to backdoor federated learning. AISTATS
46. Xie, C., Huang, K., Chen, P.-Y., and Li, B. (2019). Dba: Distributed backdoor attacks against federated learning. International Conference on Learning Representations (ICLR)1107
47. Wan, Y., Qu, Y., Ni, W., Xiang, Y., and Gao, L. (2024). Data and model poisoning backdoor attacks on wireless federated learning, and the defense mechanisms: A comprehensive survey. IEEE Communications Surveys & Tutorials
48. Shejwalkar, V. and Houmansadr, A. (2021a). Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. USENIX Security Symposium
49. El Mhamdi, E. M., Guerraoui, R., and Rouault, S. (2018). The hidden vulnerability of distributed learning in byzantium. In Proceedings of the 35th International Conference on Machine Learning (ICML) (PMLR), 3521–3530
50. Chen, L., Liu, X., Wang, A., Zhai, W., and Cheng, X. (2024a). Flsad: Defending backdoor attacks in federated learning via self-attention distillation. Symmetry 16, 1497898
51. Rocha, A. and Conti, M. (2025). Weidetect: Weibull distribution-based defense against poisoning attacks in federated learning for network intrusion detection systems. arXiv preprint arXiv:2504.04367
52. Sun, Z., Liu, C., Yang, Q., and Qi, Y. (2021). Data poisoning attacks on federated machine learning. ACM Transactions on Knowledge Discovery from Data (TKDD)15, 1–2510
53. Oprea, A., Li, X., Ma, Y., Rigazzi, G., Bridges, R. A., and Marchal, S. (2022). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. Expert Systems with Applications 204, 1175411
54. Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., and Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. IEEE Symposium on Security and Privacy
55. Gharib, A., Abawajy, J., and Tari, Z. (2022). Poisoning attacks and defenses in machine learning: A survey. IEEE Access
56. Paudice, A., Mu noz-Gonz alez, L., Lupu, E. C., et al. (2018). Label sanitization against label flipping poisoning attacks. IEEE Access 6, 5423–5431
57. Gu, T., Dolan-Gavitt, B., and Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. Proceedings of the IEEE
58. Kiss, I., Guly as, G. G., and Imre, S. (2017). Adversarial machine learning in malware detection: Arms race between evasion attack and defense.2017 IEEE International Conference on Future IoT Technologies
59. Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. International Conference on Machine Learning
60. Ma, Y., Li, X., Rigazzi, G., Oprea, A., and Marchal, S. (2022a). Systematic poisoning attacks on and defenses for machine learning in healthcare. arXiv preprint arXiv:2206.12345
61. Kroll, J. A. (2021a). Accountability in machine learning: Governance, auditability, and responsibility. Communications of the ACM
62. Miguel, J. and Chen, T. (2021). Machine learning provenance for accountability. USENIX Symposium on Operating Systems Design and Implementation
63. Cao, D., Chang, S., Lin, Z., and Liu, G. (2019). Understanding distributed poisoning attack in federated learning. IEEE Conference on Communications and Network Security (CNS)
64. Ma, Z., Ma, J., Miao, Y., and Li, Y. (2022b). Shieldfl: Mitigating model poisoning attacks in privacy-preserving federated learning. IEEE Transactions on Information Forensics and Security
65. ElZemity, A. and Arief, B. (2024). Privacy threats and countermeasures in federated learning for internet of things: A systematic review.2024 IEEE Conference on Communications and Network Security (CNS)
66. Fung, C., Yoon, C. J. M., and Beschastnikh, I. (2018b). Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866
67. Yin, D., Chen, Y., Kannan, R., and Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. In Proceedings of the 35th International Conference on Machine Learning 1 (ICML). 5650–5659
68. Cao, X. et al. (2023). Foolsgold++: Detecting sybil attacks with enhanced gradient similarity. Proceedings of the AAAI Conference on Artificial Intelligence

 doi:10.20944/preprints202605.1674.v1

69. Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems (NeurIPS)
70. Cao, X., Fang, M., Liu, J., and Gong, N. Z. (2020). FLTrust: Byzantine-robust federated learning via trust bootstrapping. arXiv preprint arXiv:2012.13995
71. Tran, B., Li, J., and Madry, A. (2018). Spectral signatures in backdoor attacks. NeurIPS
72. Wang, B. and et al. (2019). Neural cleanse: Identifying and mitigating backdoor attacks in neural networks.1094 IEEE S&P
73. Gao, Y. and et al. (2019). Strip: A defence against trojan attacks on deep neural networks. ACSAC
74. Zhang, Q., Liu, F., and Wang, C. (2023a). Bvdfed: Byzantine- and verifiability-resilient federated learning framework. Pattern Recognition Letters 176, 44–53. doi:10.1016/j.patrec.2023.01.005112
75. Li, W., Zhao, P., and Ahmed, S. (2024a). Cafcor: Covariance-bound aggregation with secret-based local differential privacy for federated learning. IEEE Transactions on Neural Networks and Learning Systemsdoi:10.1109/TNNLS.2024.334598
76. Ning, Z., Li, C., and Xu, Y. (2024b). Blockchain-enabled accountable federated learning for edge ai systems. IEEE Internet of Things Journal 11, 10325–10337. doi:10.1109/JIOT.2023.3312008
77. Liu, Z., Wang, Y., and Tang, H. (2024b). Verifiable and accountable federated learning via permissioned blockchain. Future Generation Computer Systems 155, 521–535. doi:10.1016/j.future.2024.02.01599
78. Tounsi, A., Salem, O., and Mehaoua, A. (2024). Anomaly detection in federated learning: A comprehensive study on data poisoning and energy consumption patterns in iot devices. IEEE Internet of Things Journal
79. Li, D., Wong, W. E., Wang, W., Yao, Y., and Chan, M. C. (2021). Detection and mitigation of label-flipping attacks in federated learning systems with kpca and k-means. In 2 8th International Conference on Dependable Systems and Their Applications (DSA)(IEEE), 551–5599
80. Chen, X., Zan, D., Li, W., and Guan, B. (2024b). A gan-based data poisoning framework against anomaly detection in vertical federated learning. IEEE Transactions on Neural Networks and Learning Systems
81. Alsulaimawi, Z. (2024). Federated learning with anomaly detection via gradient and reconstruction analysis. arXiv preprint
82. Khraisat, A., Alazab, A., Alazab, M., Jan, T., and Singh, S. (2025). Securing federated learning: a defense strategy against targeted data poisoning attack. Cognitive Computation
83. Gambs, S., Zhao, L., and Patel, D. (2021). Client-side validation voting in federated learning. ACM Transactions on Privacy and Security
84. Zhang, X., Kim, M., and Kumar, R. (2022). Flcert: Federated certification via client grouping and voting.1133 arXiv preprint arXiv:2201.XXXXX
85. Hakeem, S. and Kim, H. (2025). Advancing intrusion detection in v2x networks: A comprehensive survey on machine learning, federated learning, and edge ai for v2x security. IEEE Access
86. Li, S., Ngai, E. C. H., and V oigt, T. (2023). An experimental study of byzantine-robust aggregation schemes in federated learning. IEEE Transactions on Dependable and Secure Computing
87. Liu, X., Li, H., Xu, G., Chen, Z., Huang, X., and Lu, R. (2021). Privacy-enhanced federated learning against poisoning adversaries. IEEE Transactions on Information Forensics and Security 16, 4574–4588
88. Ashwinee, K. and Natarajan, V. (2021). Efficient gradient clipping in federated learning defense. Journal of Privacy and Confidentiality
89. Rathee, M., Shen, C., Wagh, S., and Popa, R. A. (2023). Elsa: Secure aggregation for federated learning with malicious actors.2023 IEEE Symposium on Security and Privacy (SP)1059
90. Ma, Y., Woods, J., Angel, S., Polychroniadou, A., and Rabin, T. (2023). Flamingo: Multi-round single-server secure aggregation with applications to private federated learning. 2023 IEEE Symposium on Security and Privacy (SP)
91. Lycklama, H., Burkhalter, L., Viand, A., K uchler, N., and Hithnawi, A. (2023). Rofl: Robustness of secure federated learning.2023 IEEE Symposium on Security and Privacy (SP). doi:10.1109/SP 46.2023. 10179400
92. Jiang, Y., Zarezadeh, M., Dai, T., and K opsell, S. (2025). Alphafl: Secure aggregation with malicious security for federated learning against dishonest majority. Proceedings on Privacy Enhancing Technologies, 348–368doi:10.56553/popets-2025-013
93. Ma, X., Cheng, K., Shen, Y., Li, X., Chang, Z., Zhang, T., et al. (2024b). Trusted model aggregation with zero-knowledge proofs in federated learning. IEEE Transactions on Parallel and Distributed Systems 1 35, 2284–2296. doi:10.1109/TPDS.2024.3455762
94. Xu, B. et al. (2025). Efficient verifiable secure aggregation protocols for federated learning. Journal of Information Security and Applications 80, 104161. doi:10.1016/j.jisa.2025.104161

 doi:10.20944/preprints202605.1674.v1

95. Bouamama, S. et al. (2025). Vesafl: Verifiable secure aggregation for privacy-preserving federated learning. IEEE Transactions on Dependable and Secure Computing
96. Bottoni, P. et al. (2025). Verifiability and privacy in federated learning through distributed ledger technologies and randomized response techniques. Proceedings of the 11th International Conference on Big Data Computing Applications and Technologies (BDCAT '25). doi:10.1145/3773276.377524
97. Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. (2019b). Analyzing federated learning through an adversarial lens. International Conference on Machine Learning (ICML)
98. Nowroozi, E. et al. (2025b). Federated learning under attack: A systematic review of poisoning threats and defenses. Frontiers in Artificial Intelligence
99. Lianga, Z. et al. (2023b). A survey on federated learning poisoning attacks and defenses. IEEE Transactions on Big Data
100. Shejwalkar, V. and Houmansadr, A. (2021b). Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. IEEE Symposium on Security and Privacy Workshops 1 (SPW)1074
101. Chen, C., Liu, J., Tan, H., Li, X., Wang, K., and Li, P. (2025a). Trustworthy federated learning: privacy, security, and beyond. Knowledge and Information Systems
102. Mohamed, H., Koroniotis, N., and Moustafa, N. (2024). Harnessing federated learning for digital forensics in iot. IEEE Transactions on Information Forensics and Security
103. Masunda, M. and Ajayi, R. (2025). Enhancing security in federated learning [Dataset].
104. Tallam, K. (2025). Engineering risk-aware, security-by-design frameworks for autonomous ai. arXiv preprintdoi:10.48550/arXiv.2505.06409
105. Zhukabayeva, T., Zholshiyeva, L., Karabayev, N., and Khan, S. (2025). Cybersecurity solutions for industrial internet of things–edge computing integration: Challenges, threats, and future directions.1147 Sensors 25, 2131148
106. Nguyen, D. (2024). IoT Security: From Context-based Authentication to Secure Federated Learning Anomaly Detection. Ph. D. thesis, TU Darmstadt. doi:10.26083/tuprints-00028827
107. Hamouda, D. (2024). New technologies for security and privacy issues in the era of industry 5.0.PhD Dissertation
108. Shaik, M., Bojja, G., and Gudala, L. (2025). Leveraging artificial intelligence for enhanced threat detection.1068 ResearchGate Preprint
109. Kairouz, P., McMahan, B., and et al. (2023). Advances and open problems in federated learning. Foundations and Trends in Machine Learning 16, 1–210. doi:10.1561/220000008
110. Truex, S., Liu, L., Gursoy, M. E., Yu, L., and Wei, W. (2019). A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security 1 (AISec). 1–11. doi:10.1145/3338501.3357370
111. Yao, J., Shen, J., Wu, Y., and Zhang, R. (2023). Aero: Efficient and verifiable secure aggregation in federated learning. In Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS). 1012–1023. doi:10.1109/ICDCS 57.2023.001131
112. Zeng, Q., Zhou, L., and Li, X. (2024b). Decentralized and auditable federated learning using blockchain and smart contracts. Information Sciences 656, 119982. doi:10.1016/j.ins.2023.119982
113. Fung, C., Yoon, C. J., and Beschastnikh, I. (2018a). Mitigating sybils in federated learning poisoning. In Proceedings of the 35th International Conference on Machine Learning (ICML). 168–1829
114. Xia, Y. and Cheng, L. (2023). Robust federated learning under adversarial attacks: Survey and outlook.1104 ACM Computing Surveys 55, 1–39. doi:10.1145/3576912
115. Rahman, M. and Debnath, B. (2024). A survey on statistical poisoning detection in federated learning.1056 Journal of Network and Computer Applications 238, 104890. doi:10.1016/j.jnca.2024.104890
116. Cen, X. and Alur, R. (2024b). Auditable ai systems: Foundations and techniques. Communications of the ACM 67, 76–88. doi:10.1145/363100
117. Miguel, E., Sanchez, J., and Ortega, P. (2021b). Accountability in artificial intelligence: From principles to practice. AI and Ethics 1, 43–59. doi:10.1007/s43681-021-00010-7
118. Malgieri, G. and Pasquale, F. (2022). The emerging principle of accountability in ai regulation. Computer Law & Security Review 45, 105701. doi:10.1016/j.clsr.2022.105701
119. International Organization for Standardization (2023). ISO/IEC 42001: Artificial Intelligence Management System (AIMS) – Requirements. International Organization for Standardization. ISO/IEC JTC 1/SC 42947

 doi:10.20944/preprints202605.1674.v1

120. Ning, W., Zhu, Y., Song, C., Li, H., Zhu, L., Xie, J., et al. (2024a). Blockchain-based federated learning: A survey and new perspectives. Applied Sciences 14, 9459. doi:10.3390/app14209459
121. Ma, R. et al. (2024a). Trusted model aggregation with zero-knowledge proofs in federated learning. IEEE Transactions on Dependable and Secure Computingdoi:10.1109/TDSC.2024.XXXXXXX
122. Gill, W. et al. (2023). Tracefl: Interpretability-driven debugging in federated learning. arXiv:2312.13632