



REVIEW

A Systematic Review of Privacy Threats in Federated Learning with Healthcare Data

Nada Hasan Alfahmi, Manal AlGhamdi and Muhammad Arif*

Department of Computer Science and Artificial Intelligence, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia

*Corresponding Author: Muhammad Arif. Email: mahamid@uqu.edu.sa

Received: 29 May 2026; Accepted: 12 August 2026; Published: 22 October 2026

ABSTRACT: Federated learning (FL) supports collaborative model development across healthcare institutions without centralizing raw patient data, but exchanged updates can still disclose sensitive information or be manipulated by adversaries. We systematically reviewed privacy and security threats and their defenses in healthcare FL using a PRISMA-guided process. The historical review flow reported 3,702 identified records, 232 full-text reports, and 45 included peer-reviewed primary studies published from 2022 to 2026. During revision, an audit of the supplied screening materials identified 163 archived bibliographic records. To avoid inventing the identities of 69 unavailable historical records, we conducted a separate, transparently labeled retrospective reconstruction on 31 July 2026 and added 69 unique, search-compatible records that met the review's discovery concepts but failed prespecified eligibility rules. These reconstructed records are not claimed to be exact historical missing reports and were not included in the 45-study evidence synthesis. Cross-study synthesis identified poisoning, reconstruction/inversion, membership inference, adversarial/evasion, and side-channel threats and showed that defense failures recur because protection is applied at only one layer: differential privacy can reduce leakage but may obscure malicious updates; encryption and secure aggregation protect update contents but not update integrity; robust aggregation and anomaly detection can confuse malicious behavior with legitimate non-IID heterogeneity; and blockchain improves auditability without validating update semantics. Formal funnel plot and regression tests were inappropriate because the outcomes and effect measures were not commensurable. Structured reporting-bias assessment and sensitivity analyses did not change the principal conclusion: no single defense family provides adequate protection against adaptive, multi-vector attacks. Healthcare FL therefore requires layered, context-aware security combining cryptographic confidentiality, statistical privacy, integrity monitoring, authentication, and continuous risk assessment.

KEYWORDS: Federated learning; healthcare; privacy threats

1 Introduction

1.1 Background and Motivation

Artificial intelligence (AI) and machine learning (ML) have increasingly transformed healthcare delivery in recent years. They automate analyses, facilitate clinical decision support, and provide insights through data collection across disciplines, including medicine [1,2]. Rapid digital transformation of healthcare has generated an unprecedented volume of sensitive clinical data, creating new opportunities for AI in disease prediction and personalized medicine. At the same time, healthcare data are subject to stringent privacy and data-protection requirements, including the Health Insurance Portability and Accountability Act (HIPAA)

and the General Data Protection Regulation (GDPR). Such regulations guarantee the legal processing, protection and sharing of information about health care. Combined with institutional data silos, governance and interoperability barriers, these issues make it even more imperative to secure privacy in healthcare to conduct collaborative analysis with healthcare institutions [3].

Although FL was introduced in 2017 as a privacy-preserving paradigm [4], subsequent research found privacy vulnerabilities. Zhu et al. [5] presented gradient inversion attacks and Nasr et al. [6] identified membership inference threats. In this section, we focus on the 2022–2026 time period, when the attack sophistication has been rising, and we place these advances in the perspective of a broader history.

1.2 The Promise and Perils of Federated Learning in Healthcare

To ease this issue, Federated Learning (FL) was developed [4], enabling multiple institutions to train a model directly on data-driven clinical outcomes. Even so, recent studies demonstrate that FL does not inherently provide guaranteed privacy [7], despite its architecture. The gradients and parameters used to swap updates in a model during training can inadvertently expose potentially sensitive data. This leakage is a fundamental flaw in the optimization process and a marked vulnerability, especially in deep models and during initial training rounds [8,9]. This vulnerability was exploited by several well-defined privacy threats. Membership inference attacks can detect whether a particular patient's record was used in the training set, thereby revealing the patient's association with a corresponding health condition [10]. Data reconstruction and model inversion attacks, on the other hand, go one step further by attempting to restore recognizable forms of the original private data, such as medical images. Gradient inversion attacks, for instance, have been successful for reconstructing high-quality medical images from model updates [11]. For example, property-inference attacks attempt to infer vulnerable population-level attributes, such as disease prevalence, from aggregated model updates. In addition, model poisoning attacks (e.g., when a malicious actor makes fake updates) can lower the threshold for an efficient privacy violation by further degrading model performance, as well as demonstrate that robustness and privacy are dual requirements of FL systems [12]. The healthcare domain is especially significant for FL applications, given the sensitive nature of medical datasets and the risks of privacy violations such as discrimination, social stigma, and criminal liability [13]. Although FL has been extensively used by academics for multi-institutional research, the available evidence demonstrates that privacy attacks are technically and practically feasible, especially for patient data reconstruction using gradient-based methods [14]. This fact indicates a wide gap between the so-called privacy promise of FL and its guaranteed delivery as a medical technology. However, few studies have considered the dynamic nature of operational healthcare FL systems and the related consequences of changes in the privacy threat landscape. For example, most previous literature has concentrated on static images in simplified situations, without addressing more complex data types that will become critical for understanding real-world threats, such as surgical videos. Second, many defense mechanisms [15,16] have been proposed, yet are often evaluated in isolation. The trade-off between privacy protection (the level of protection provided) and model utility (the suppression of model performance degradation) has not been systematically evaluated for high-dimensional, temporally ordered medical data. In this context and in light of these deficiencies, this systematic review aims to compare privacy threats and defense functions in healthcare FL. It is intended to chart the evolving threat landscape, analyze how attacks operate in real-world healthcare scenarios, and critically assess the merits and limitations of proposed defenses in medical FL systems. Failures of privacy in medical federated learning (FL) are far more than technical issues [17]. Healthcare data represent some of the most sensitive data on humans, and the leaking of this information will do irreversible damage—namely, employee discrimination and social stigmatization, psychological distress, and criminal liability [13]. Inevitably, patient trust in AI-based medicine as well as institutional commitment to multi-centre research

is deeply reliant on demonstrable privacy guarantees. Regulations e.g., the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) have stringent mandates for data custodians, and breaching these—either in usual cyber-attack or FL-centric inference attack—would incur enormous economic and legal costs [18]. Meanwhile, the amount and variety of medical data from EHRs, wearable data, and high-definition imaging systems is increasing significantly, making collaborative AI essential to realizing generalizable clinical models [19]. This tension between the clinical imperative to share knowledge and the privacy imperative to protect patients makes privacy research in healthcare FL the focus of a critical sociocultural debate.

This systematic review seeks to address these challenges by synthesizing the current state of knowledge on privacy threats in healthcare FL, evaluating the effectiveness of existing defenses, and providing a roadmap for future research and practice.

1.3 Overview of the Privacy Threat Landscape in Healthcare Federated Learning

Early adopters of medical FL assumed that localized data storage, since ‘data never leaves the hospital,’ guaranteed privacy. But as the evidence has built up, that is no longer an appropriate assumption. The most recent privacy attacks are divided into four families that exploit an aspect of the FL protocol:

1. **Gradient-inversion and data-reconstruction attacks.** Shared gradients harbor quite a bit of information about the local training samples. More sophisticated inversion techniques applied to medical imaging benchmarks have recovered chest X-rays, brain MRI scans, and CT volumes with clinically recognizable fidelity [20–22]. Shi et al. [23] demonstrated that a malicious server can recover high-fidelity patient images even when secure aggregation is employed, exposing a flaw in the existing architecture that can be exploited.
2. **Membership-inference attacks (MIAs).** MIAs evaluate whether a specific individual’s record contributed to a model’s training set. In healthcare, an efficient MIA may expose a patient’s allocation to a disease-specific registry, psychiatric unit, or clinical trial, violating privacy and research ethics requirements [10]. Bai et al. conducted a comprehensive survey of this data [24] and found more than fifty MIA variants in FL scenarios, and described how healthcare settings magnify potential harm, especially because the auxiliary information that adversaries have access to (e.g., public hospital admission records) is considerably richer than in other domains.
3. **Property-inference attacks (PIAs).** It is demonstrated that PIAs can learn population-level characteristic data that is independent of the primary learning task—say, the burden of disease within the catchment of a participating hospital or patients’ approximate age in a treatment arm [25]. Clustered FL, an established remedy for non-IID clinical data, is particularly sensitive to PIAs since cluster membership itself leaks distributional information [26].
4. **Model-poisoning attacks.** These threats [27] include malicious participants who may churn out well-crafted updates to the global model such that the entire model becomes less accurate and/or inject backdoors. Beyond the direct impact on clinical reliability, poisoning attacks can lower the barrier to simultaneous invasion of privacy—for instance if they force the model to memorize specific training samples more strongly, resulting in later versions of gradient inversion [12].

In medical image classification, minimal-budget poisoning attacks based on attention maps have brought a 44% drop in test accuracy while manipulating only 24% of the pixels, demonstrating how readily an adversary can slip under the radar [28]. Crucially, these types of attacks are more than hypothetical practices. Das et al. [21] showed that random noise injection, a common defense, is typically insufficient to safeguard medical images, while Stelter et al. [29] found that privacy-preserving interventions, such as differential privacy, led to significant decreases in model utility on clinically meaningful segmentation problems.

These results together indicate a vast and poorly mapped gap between the privacy aspirations of FL and its practical implementation in healthcare.

1.4 Overview of Defense Mechanisms and the Privacy–Utility Trade-off

There are a number of alternative countermeasures in focus that fall under three key groups:

1. **Cryptographic protections.** Secure multi-party computation and homomorphic encryption allow model aggregation over encrypted updates, preventing honest-but-curious servers from inspecting individual gradients [30]. Recent frameworks such as MedHE integrate adaptive gradient sparsification with CKKS homomorphic encryption to achieve both communication efficiency and formal cryptographic privacy [31].
2. **Differential privacy (DP).** DP provides a mathematically rigorous privacy guarantee by injecting calibrated noise into model updates. But the high-dimensional nature of medical data, particularly surgical video and volumetric imaging, makes the noise required to reach meaningful privacy budgets ($\epsilon < 1$) often render the model unusable for clinical decision-making [29,32].
3. **Architectural and procedural safeguards.** These include robust aggregation rules, anomaly detection for malicious updates, and privacy-aware loss functions. For instance, Health-FedNet combines DP, homomorphic encryption, and an adaptive node-weighting mechanism to balance privacy, scalability, and model robustness across heterogeneous clinical sites [33].

Despite these advances, the majority of defense evaluations have been conducted only on low-dimensional tabular datasets or single-modality image collections under idealized threat models. Yet there is a dearth of studies that systematically vary the privacy budget while measuring clinically relevant endpoints (e.g., sensitivity, specificity, calibration) on longitudinal, multimodal, or video-based medical data. Furthermore, many defense proposals assess only one attack family at a time, leaving open the question of how well they can withstand an adaptive adversary that combines, for example, model poisoning with membership inference [34].

1.5 Research Gaps and Rationale for This Systematic Review

There are many key research gaps that still need to be addressed. First, the nature of the privacy threat landscape against operational healthcare FL systems has not been explicitly measured over time [18]. The former work has mainly addressed static images in simplified settings and has ignored more complex data types—e.g., surgical videos and streaming physiological signals—which are critical for understanding real-world threats [21]. Second, although many defense mechanisms have been suggested, they are routinely evaluated in isolation. For high-dimensional, temporally ordered medical data [29,32,33], the trade-off between privacy protection and model utility has been neglected. Third, no systematic synthesis is available that examines how unique attack families interact and whether defenses that protect against one threat might actually amplify others. To address these needs, this systematic review aims to analyze the privacy threats and techniques in healthcare FL. By mapping the evolving threat environment, including the nature of attacks experienced by real healthcare systems, and analyzing the feasibility of proposed defenses for medical FL systems, the research objectives will be achieved.

1.6 Research Questions

Our extensive research questions address privacy threats to Federated Learning with healthcare data and how Federated Learning is transforming the landscape for patients and families.

- **RQ1:** How could Federated Learning systems in healthcare be constructed to offer a sturdy defense against adaptive multi-vector privacy attacks, while maintaining model utility and clinical relevance?
- **RQ2:** What are the practical approaches to develop dynamic, context-aware privacy protection mechanisms that evolve with the sensitivity of medical data and the changing threat landscape?
- **RQ3:** What approaches and metrics must be established to quantify the clinical value of privacy attacks and defensive measures for health care FL systems and the real-world implications of FL in practice?
- **RQ4:** How may we evolve Federated Learning architectures to handle the specific complexities of healthcare data, in particular the extreme non-IID distributions, class imbalance, and the multi-modal data consolidation, with a focus on privacy?

1.7 Comparison with Existing Surveys

To situate the contribution of this systematic review, 5 representative survey papers from the related federated learning (FL) and healthcare privacy literature were analyzed. [Table 1](#) provides a summary of these studies in 6 main dimensions: year of publication, citation of relevant papers, attacks analyzed, defenses, domain specificity, and clinical realism.

Table 1: Comparison of existing surveys on privacy in federated learning for healthcare.

Survey	Year	Attacks Covered	Defences Covered	Healthcare Specific	Clinical Realism
Madathil et al. [35]	2025	Inference attacks (MIA, PIA)	DP, HE, SMPC, TEE, blockchain, watermarking, ZKP	Yes	No
Franklin Open [36]	2025	Inference attacks (MIA, PIA)	DP, SMPC, HE, TEE, ZKP, blockchain, watermarking	Yes	No
Ali et al. [19]	2024	Model misconduct, security threats	Limited (generic)	Yes	No
Bai et al. [24]	2025	Membership inference attacks only	Perturbation, data replacement	No	No
Neto et al. [37]	2023	Model performance & data privacy attacks (generic)	General mitigation strategies	No	No
Our Review	2026	Poisoning, reconstruction/inversion, membership inference, adversarial, side-channel	DP, HE, SMPC, robust aggregation, hybrid multi-layer defense	Yes	Yes

Results from the studies provide important contributions to the FL healthcare literature. Yet, their interpretation also points to some significant gaps. First, regarding coverage of attacks, most existing surveys mention FL applications to a substantial extent, focusing on privacy attacks without any organized classification or examination of the technical solutions to attacks. Madathil et al. [35] review over 250 articles, but they mostly cover FL systems, aggregation algorithms, and data partitioning rather than detailed attack

taxonomies. Similarly, Neto et al. [37] subgroup security vulnerabilities with respect to model performance and data privacy attacks, but have done nothing yet regarding fine-grained classification. On the other hand, we systematically categorize attacks in our review into five common types: poisoning attacks, reconstruction/inversion attacks, membership inference attacks, adversarial attacks, and side-channel attacks through rigorous technical assessment for each type. Second, there is an emphasis on technical measures rather than clinical realism. Yet one limitation that prevails across all the surveys presented is their lack of clinical realism in assessing attacks and defenses. The Franklin Open systematic review [36] studies privacy-preserving approaches (DP, SMPC, HE, TEE, ZKP, and blockchain) but does not evaluate them clinically, but rather with respect to computational cost, scalability, and regulatory compliance. Survey reports have performance metrics (accuracy, PSNR, AUC), but not whether reconstructed medical images maintain clinically identifiable features, whether adversarial perturbations would actually mislead the clinician, or at what point diagnostic utility is retained after privacy defenses have been applied. Our review explicitly fills this gap by looking at attacks and defenses in a clinical context. Third, current surveys do not analyze attack mechanisms in detail. While Bai et al. [24] is an exhaustive survey, especially regarding membership inference attacks and defenses, its scope must be limited. Other surveys discuss inference or reconstruction attacks without detailing, in technical terms, how these attacks work, what assumptions they rely on, or how they relate to certain FL architectures. Ali et al. [19] discuss model misconduct and security challenges, but remain high-level. In our review, we describe the different types of attack mechanisms (e.g., gradient inversion attacks, property inference attacks, model poisoning attacks) and their underlying optimization strategies in detail. Fourth, there is no analysis of attack–defense interaction. So far, there are only a handful of surveys on attacks and defenses interacting in multi-adversary settings. Most assessments focus on defenses alone, with a single threat model assumed. The Franklin Open review [36] contrasts privacy-preserving techniques by their cost and scalability, but does not comment on the interplay among different defense mechanisms as they grapple with hybrid attack strategies. In contrast, our review examines interactions among multiple attack families, such as how model poisoning can reduce the barrier to performing gradient inversion, or how differential privacy noise can be used for backdoor injection. Fifth, little is known in the literature about adaptive adversaries. Most existing surveys treat adversaries as static units with stable abilities. But recent studies reveal that adaptive adversaries—who adjust their attack strategies in response to current feedback from the FL system—pose a much more imminent threat. Neto et al. [37] address security vulnerabilities in FL applications; however, their focus on adaptive adversaries is not explicit. This dimension is explicitly considered in our review, which observes that state-of-the-art attacks like DP-Poison optimize malicious updates to mimic benign differential privacy noise, thereby bypassing defense mechanisms that assume non-adaptive attackers. The majority of surveys do not categorize attacks by attacker capacity (honest-but-curious server vs. malicious participant) or knowledge requirements (white-box vs. black-box), which may indicate a dearth of taxonomies of attack surfaces. Such side-channel attacks, for instance, motion-based keystroke inference attacks on smartphone sensors, are rarely addressed despite their application to mobile healthcare monitoring. Property inference attacks receive scant coverage: attribute leakage at the population level (e.g., disease prevalence) is regularly ignored or addressed only tangentially. No meaningful assessment of the privacy–utility trade-off: although there is some reporting of the trade-off in surveys, few quantify the loss in clinical utility under a given privacy budget, particularly for high-dimensional medical data. Finally, there is a lack of mapping to regulation: as HIPAA and GDPR are cited, existing surveys rarely map specific attack scenarios to regulatory provisions or assess compliance gaps.

1.8 Contributions of the Paper

This paper makes four primary contributions:

- **Systematic Literature Review:** We systematically review 45 recent works on privacy vulnerabilities in healthcare FL, classifying them into five attack families: poisoning, reconstruction/inversion, membership inference, adversarial, and side-channel attacks.
- **Defense Mechanism Assessment:** We evaluate current defense mechanisms—differential privacy, homomorphic encryption, secure multi-party computation, and robust aggregation—against adaptive, multi-vector attacks, identifying their limitations and failure modes.
- **Domain-Specific Analysis:** We provide a healthcare-specific analysis linking technical vulnerabilities to clinical realism, data fidelity, and regulatory frameworks, identifying critical gaps in current assessment practices.
- **Framework and Recommendations:** We propose the Adaptive Multi-Layer Trust (AMLT) framework and provide practical guidelines for researchers, practitioners, and policymakers to design secure and reliable FL systems for healthcare.

The paper itself has seven sections that systematically address the research questions. Following the Introduction (1) section introducing context, key research questions, and the review paper's contribution to understanding existing topics, the Search Methodology section ([Section 2](#)) outlines the systematic review method (using PRISMA procedures in [Fig. 1](#) for selection of databases, inclusion/exclusion criteria, year-wise publication of papers) of the review to be applied. The Summary of Selected Papers ([Section 3](#)) offers a systematic review of 45 published articles that we analyzed on five types of attacks. Comparative tables for those are presented. [Section 4](#) (Analysis and Discussion) summarizes different technical, domain-specific and meta-analytical perspectives, and presents tables to analyze attack taxonomy, attack efficacy, the evadeability of defense, the sensitivity of data, and the regulatory issues. A section devoted to future directions ([Section 6](#)) introduces future trust models and their practical implications for practice, policy, and research. The Answers to the Research Questions section ([Section 5](#)) is where responses to the four research questions are summarized, and findings from the literature are discussed. The paper's Final [Section 7](#) delivers main findings and specific proposals for designing secure and reliable FL platforms for healthcare users.

2 Search Methodology

The review was carried out and reported using the PRISMA framework [[38](#)]. We searched IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect between January and March, 2026. Google Scholar, as well as backward and forward citation searching, were used as additional discovery routes. The protocol outlined the review questions, eligibility criteria, screening sequence, extraction fields, and risk-of-bias domains before full-text analysis. The full database-specific search strings available in the source manuscript are reproduced in Supplementary Section S1.

The search focused on three concept blocks: (i) federated learning; (ii) privacy or security attacks and defenses; and (iii) healthcare, medical, clinical, hospital, or Internet of Medical Things. The general approach was chosen to avoid bias from different terms used for gradient reconstruction, model inversion, membership inference, poisoning, Byzantine behavior, adversarial examples, secure aggregation, and privacy-preserving architectures.

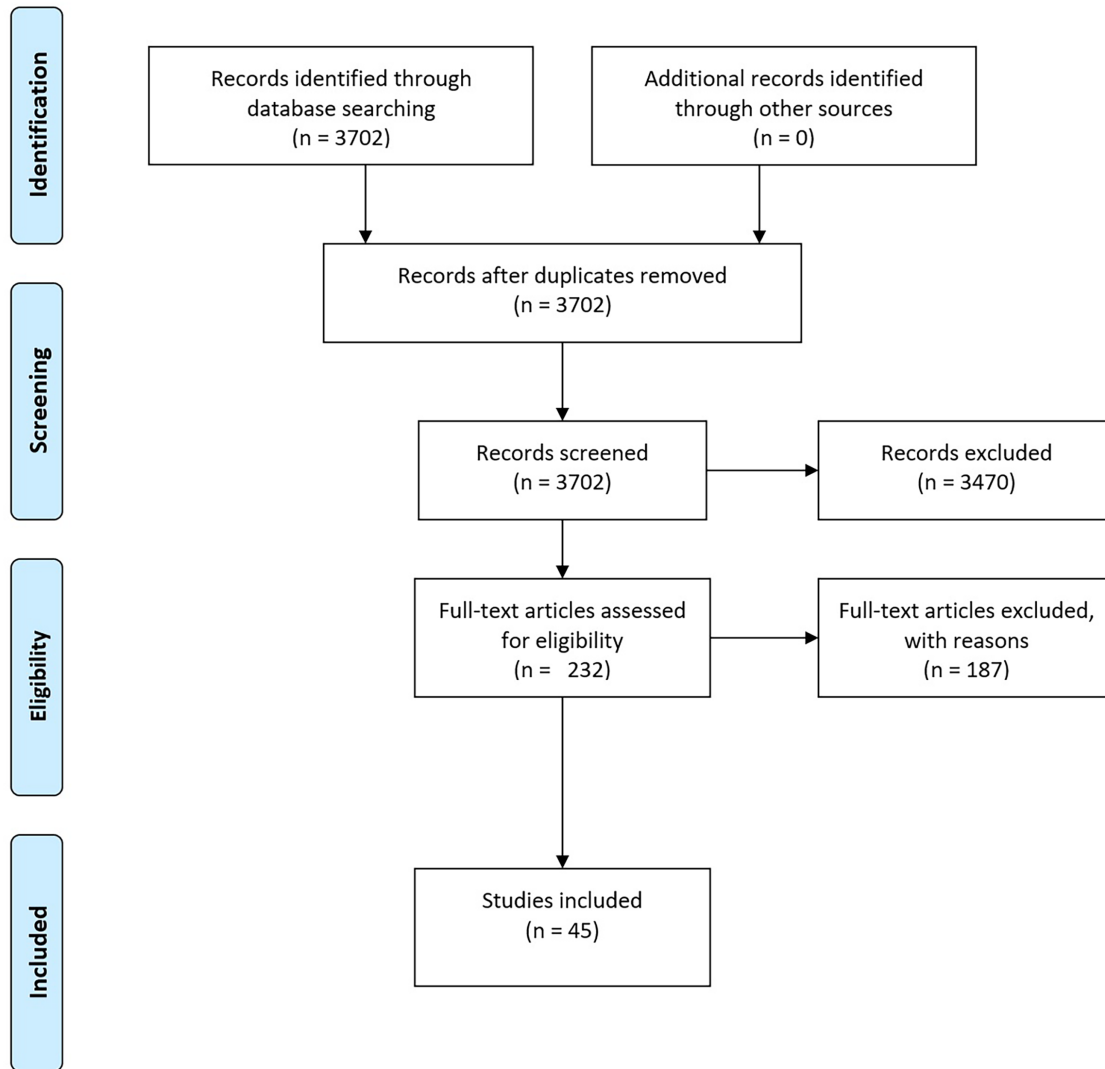


Figure 1: PRISMA flow diagram illustrating the study-selection process.

2.1 Search Results and Record Management

The historical review flow reported 3702 records and 232 full-text research articles. The source distribution is shown in [Table 2](#).

Table 2: Distribution of initial search results across sources.

Source	Records	Percentage
SpringerLink	1754	47.4%
ACM Digital Library	1629	44.0%
IEEE Xplore	111	3.0%
ScienceDirect	65	1.8%
Google Scholar and citation searching	143	3.9%
Total	3702	100.0%

2.2 Eligibility Criteria and Full-Text Decision Rules

As shown in Fig. 2, the article was included if (1) it was a peer-reviewed journal article or full conference paper; (2) it was published in English between 2022 and 2026; (3) FL, split FL, or federated transfer learning was part of the method; (4) the study was conducted in a healthcare, clinical, medical imaging, hospital, remote-monitoring, or IoMT use case, or the transferability of an FL security method to the given scenario; (5) it evaluated a privacy/security attack, a defense against a specific threat, or a privacy-preserving protocol under an explicit adversary or leakage model; (6) the research has been empirically or theoretically based or security analysis.

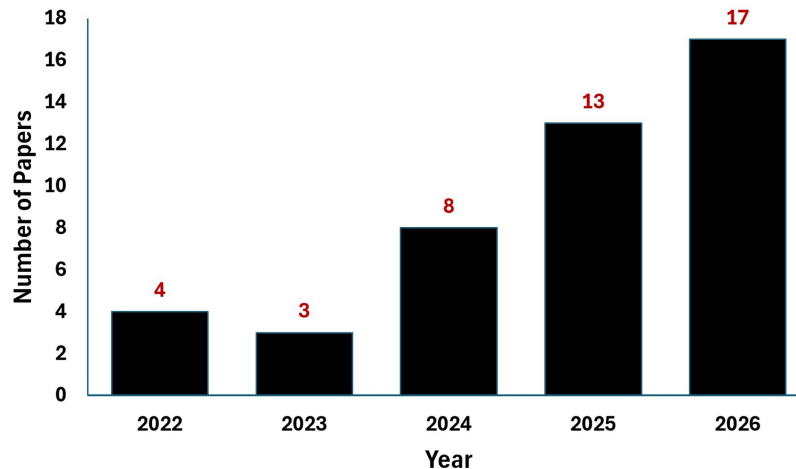


Figure 2: Year-wise distribution of the included studies.

Full-text exclusions were selected using a hierarchical set of criteria to avoid multiple and inconsistent reasons. The categories were: outside healthcare scope; no substantive privacy/security analysis; not an FL study; secondary review/viewpoint; no empirical or theoretical attack/defense evaluation; outside the date range; insufficient peer review or metadata status; and duplicate/superseded. Borderline studies are not excluded simply because they were based on a general benchmark; they are only removed if the healthcare relevance and threat model are defined. The signaling questions, examples, and reason codes are listed in Supplementary Section S2.

2.3 Independent Screening and Agreement Resolution

Two reviewers independently screened titles and abstracts and assessed the full texts (both titles and abstracts) and then reviewed the full text. Agreement was 89.7% on title/abstract screening and 93.2% on full text screening. Each reviewer made an include/exclude decision, and for exclusions, one of the exclusion reason codes and a reason code for the other. Reviewers then compared the decisions. A disagreement was resolved only if both reviewers agreed that the operational criterion had been satisfied or violated; neither reviewer's decision automatically took precedence. If any issues remained open after discussion, the third author reviewed the article, the reason codes, and the eligibility criteria and ultimately made the final decision.

2.4 Data Extraction and Narrative Synthesis

A structured form was obtained that captured bibliographic details, clinical context, data modality and dataset, FL architecture, adversary location and knowledge, attack family, defense family, privacy/security

outcome, predictive utility, computational or communication overhead, and reproducibility information. The extraction fields and verification requirements are given in Supplementary Section S6. Although the studies used incompatible datasets, threat models, privacy budgets, attack-success definitions, and utility metrics, a pooled effect estimate did not stand out statistically nor clinically. We therefore conducted a structured narrative synthesis by stratifying studies from the attack family and then comparing them with those from the defense family, assumptions, reported strengths, failure modes, and deployment constraints.

2.5 Risk-of-Bias Assessment and Validation of the Adapted Instrument

QUADAS-2 was developed for diagnostic accuracy studies and was not considered a valid off-the-shelf tool for machine-learning security research. Instead, it was used to develop an FL-specific tool called the FL-Privacy Risk-of-Bias instrument. The four adapted domains were (D1) threat model realism; (D2) technical validity and reproducibility of attack/defense implementation; (D3) correctness and completeness of the privacy, security, utility, and efficiency results; and (D4) conduct and data flow (partitioning, client sampling, preprocessing, baselines, exclusions). Healthcare applicability is assessed separately as A1 and is not used to understand the overall risk-of-bias judgment. The domain was scored Low, High, or Unclear with clear signaling questions and decision rules (Supplementary Section S3).

The content validity was reinforced in three steps already undertaken in this review: piloting five randomly selected studies by both reviewers in the context of the review; clarifying ambiguous signaling questions; and obtaining advice from two external FL-security researchers on domain coverage and terminology. The study levels and rules are reported in Supplementary Section S4. The resulting domain-level risk-of-bias judgments across the 45 included studies are summarized in [Table 3](#).

Table 3: Risk-of-bias judgments across all 45 included studies.

Judgment	Threat Model	Implementation	Evaluation Metrics	Conduct/Data Flow
Low	39 (86.7%)	42 (93.3%)	29 (64.4%)	28 (62.2%)
High	0 (0.0%)	3 (6.7%)	6 (13.3%)	3 (6.7%)
Unclear	6 (13.3%)	0 (0.0%)	10 (22.2%)	14 (31.1%)

Threat-model specification and implementation/comparator reporting were generally more robust than outcome measurement and conduct/data-flow reporting. D1 (threat model) was rated Low in 39/45 studies (86.7%) and Unclear in 6/45 (13.3%), with no studies rated High. D2 (implementation/comparator fidelity) was rated Low in 42/45 studies (93.3%) and High in 3/45 (6.7%). D3 (evaluation metrics/outcome measurement) was rated Low in 29/45 studies (64.4%), High in 6/45 (13.3%), and Unclear in 10/45 (22.2%). D4 (study conduct and federated data flow) was rated Low in 28/45 studies (62.2%), High in 3/45 (6.7%), and Unclear in 14/45 (31.1%).

Using the prespecified overall judgment rule, whereby a study was rated High if any D1–D4 domain was High, Low only when all four domains were Low, and Unclear otherwise, overall risk of bias was Low in 16/45 studies (35.6%), High in 9/45 (20.0%), and Unclear in 20/45 (44.4%). Healthcare applicability was assessed separately and did not contribute to the overall risk-of-bias classification. These judgments were used to inform the certainty and interpretation of the qualitative synthesis rather than being combined into a numerical quality score.

2.6 Reporting-Bias Assessment and Sensitivity Analyses

Traditional publication bias tests such as funnel plots and Egger regression require a common effect measure and similar sampling variances. These assumptions were not met, and such tests would yield inaccurate results. Instead, we evaluated reporting bias across four domains: database and language coverage; whether findings were not significant or incorrect; security and utility results; and positive-result language. Forty-one of the 45 abstracts that were included described the proposed method as good or better than the others in positive or superior terms, while not very many authors reported negative or harmful findings. This, along with the refusal to include non-peer-reviewed reports and English-only eligibility, would appear to suggest that dissemination and selective-outcome bias pose a substantial risk. The review thus does not imply that one defense family is more robust than another.

Three sensitivity analyses were performed in parallel with the new study-level evaluation, as shown in Supplementary Table S3. Our domain-dependent analyses use all 45 primary studies. (1) removed studies with at least one High D1–D4 domain (35 studies were retained), (2) restricted to studies with Low D1 threat model specification (38 studies were retained), and (3) restricted to studies with Low D3 outcome measurement (28 studies were retained). In all of these analyses, the main qualitative findings were not altered: update leakage and poisoning remain possible under the assumptions we made, single-layer defenses still have attack-specific failure modes, and defense-in-depth remains necessary. Note that these analyses are robustness checks of the qualitative synthesis rather than pooled-effect sensitivity analyses. Full sensitivity tables are provided in Supplementary Section S5.

2.7 Limitations of the Search Methodology

The review is limited by English-language and peer-reviewed publication restrictions, the concentration of papers in recent years, the incomplete standardization of attack and defense results, and the use of generalized computer vision benchmarks in some studies claiming healthcare relevance. Changes in search platform indexing may affect the exact record count on rerun. Positive results dominate, and the lack of negative experiments may overestimate the effectiveness of these defense results.

3 Privacy Threats in Federated Learning with Healthcare Data

This section retains study-level technical detail because attack assumptions and evaluation settings differ materially. To avoid an article-by-article catalog, each attack-family subsection closes with a cross-study comparison, and [Section 4](#) synthesizes recurring strengths, trade-offs, and failure mechanisms across defense families. The evidence should be interpreted as comparative narrative evidence rather than a quantitative ranking. [Fig. 3](#) illustrates the standard FL workflow starting from distributed hospitals (clients), each holding private medical data (X-ray, MRI, EHR). Clients perform local training, generating model updates that pass through identified attack surfaces: poisoning (label flipping, backdoor), inference (membership, property), reconstruction (gradient inversion), and evasion/side-channel attacks. In response, defense layers are applied: differential privacy, homomorphic encryption, robust aggregation (e.g., Krum), and anomaly detection (e.g., KL divergence, shadow models). The central server receives protected updates, verifies and filters them, aggregates them (e.g., via FedAvg), and distributes the updated global model back to clients. The figure visually emphasizes that FL is not inherently private; active threat mitigation must be embedded at multiple stages.

Following a systematic literature selection process, 45 papers on privacy threats in federated learning involving healthcare data were selected. The 45 papers in this collection are divided into five sub-sections related to the nature of attacks. As determined by the systematic review, the mind map in [Fig. 4](#) categorizes privacy attacks into five primary families. Poisoning attacks come in the form of data poisoning

(label flipping, backdoor injection) and model poisoning (targeted backdoor, generic poisoning). Reconstruction/inversion attacks include gradient inversion, model inversion, and active reconstruction. Black-box MIA, source inference, and white-box MIA are examples of Membership inference attacks. Adversarial and evasion attacks include evasion methods (FGSM, PGD, and transferability attacks). Side-channel attacks concentrate on motion-based keystroke inference and hardware side-channels. The graph also shows multi-vector/composite attacks such as DP-Poison, which exploits differential privacy noise to conceal backdoor injections. The taxonomy offers an organized summary of the threat landscape.

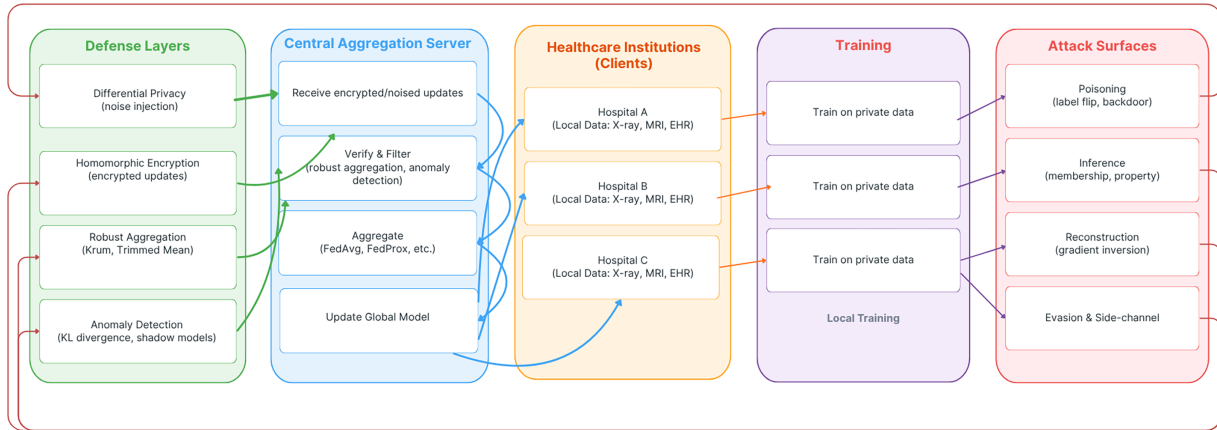


Figure 3: End-to-end standard FL workflow.

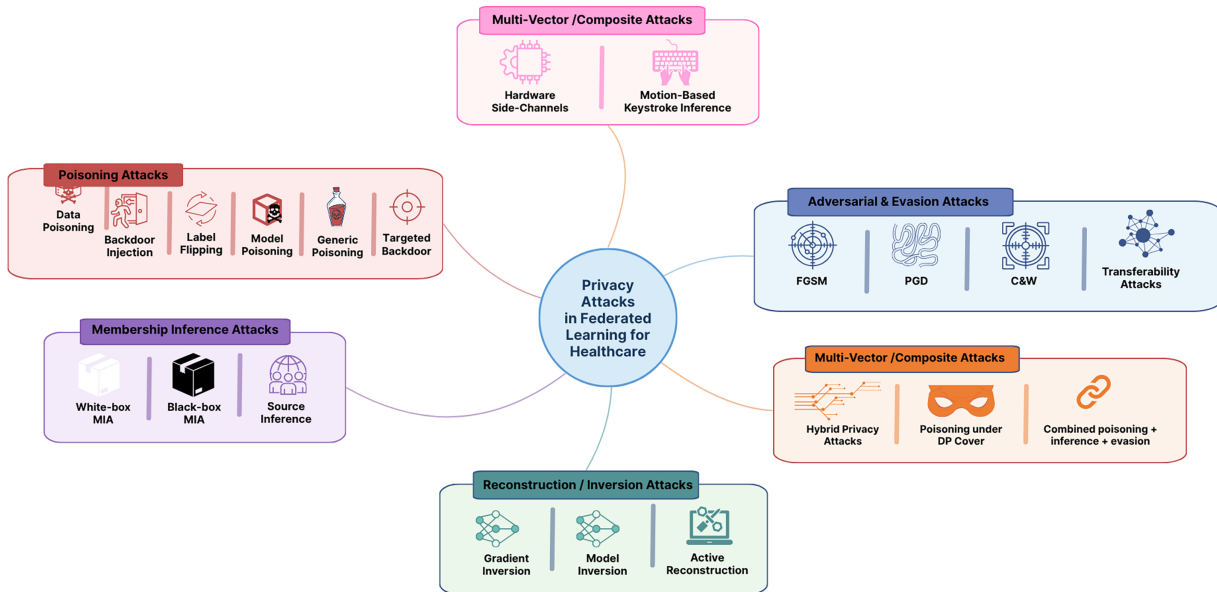


Figure 4: Taxonomy of privacy attacks in federated learning for healthcare.

3.1 Poisoning Attacks

Poisoning attacks are one of the most widely studied threat vectors in healthcare FL. The attacks are typically organized into data poisoning (manipulating training labels or features) and model poisoning (developing malicious model updates). The clinical consequences of successful poisoning attacks are severe, as they can directly impact patient diagnosis and treatment decisions.

3.1.1 Label-Flipping Poisoning Attacks

Label-flipping attacks occur when adversaries tailor training labels to either degrade model performance or induce misclassifications. These attacks are particularly dangerous in medical settings, where misdiagnosis can delay treatment or lead to errors.

Anunita et al. [39] conducted a detailed analysis of robust aggregation of retinal OCT images under label-flipping attack. Looking at 83,484 OCT images from 4 classes (CNV, DME, DRUSEN, NORMAL) across 5 different institutes using a pre-trained VGG16, they found that Krum aggregation achieved the best robustness and retained F1 scores of 94–95 even under 70% label flipping, compared to FedAvg, Median, and Trimmed-Mean. Zahir-Ismail and Shukla [40] analyzed label-flipping risks in Split Federated Learning with the MNIST and MIT-BIH Arrhythmia datasets. They found that untargeted attacks caused the largest accuracy degradation (up to 71.31% on ECG data), distance-based attacks outperformed traditional targeted attacks, and deeper client-side models were more vulnerable (MNISTv2/ECGv2 > MNISTv1/ECGv1). Khedekar et al. [41] demonstrated that label-flipping attacks can drive global accuracy from ~85% to less than 30% in highly heterogeneous environments. Kalapaaking et al. [30] proposed a blockchain architecture with SMPC model verification for healthcare systems. Using TissueMNIST (236,386 samples) and OCTMNIST (109,309 samples), their SMV filter achieved up to 25% accuracy in heavily poisoned scenarios (50% malicious clients).

3.1.2 Backdoor Poisoning Attacks

Backdoor attacks can subtly change model behavior on specific triggers and still be normal on benign inputs, making them particularly difficult to detect in clinical settings.

Chen et al. [42] adapted matching networks with random Glorot initialization against backdoor poisoning in federated meta-learning and reduced attack accuracy from 90%–100% to <20% on Omniglot/mini-ImageNet and ~40% on CelebA but maintained main task accuracy >96% on Omniglot. Zheng et al. [43] proposed DP-Poison, which exploits differential privacy noise as a cover for backdoor injection. The attack achieved close to 100% attack success rate with less than 3% main-task accuracy loss and protected against RFA, Trimmed Mean, Multi-Krum and Bulyan defenses. This work showed how privacy mechanisms can be weaponized. Kang et al. [44] developed a two-dimensional backdoor defense based on direction alignment (cosine similarity) and sign consensus (majority voting on Top-K parameters). On CIFAR-10, CIFAR-100 and EMNIST, DABD achieved 87.2% accuracy and 8.1% attack success rate under Neurotoxin, with ASR as low as 47.1% for the second defense, AlignIns.

3.1.3 Adaptive and Multi-Vector Poisoning Attacks

Recent work has revealed sophisticated poisoning strategies that adapt to defenses or combine multiple attack vectors.

Soliman et al. [45] developed GRASP-FL, a closed-loop defense against adaptive backdoor attacks (DBA, SSB, ALMR, CLFC). The framework integrates geometric screening, reputation-based aggregation, layer-wise sanitization, and GAN-guided adaptation. On CIFAR-10, FEMNIST, and MNIST, with up to 3% malicious clients, GRASP-FL achieved clean accuracy up to 97.25% and ASR below 10%. Dao et al. [46] proposed HyperShield, a hypernetwork-based defense against untargeted model poisoning in healthcare FL. Using MIMIC-III ICU data, HyperShield achieved ROC AUCs of 0.82–0.90 with 50% malicious clients, while incurring low computational overhead (CPU < 7%, RAM < 2.5% on Jetson Nano). Ning et al. [47] developed FWLA, a federated weighted learning algorithm using residual testing and adaptive weighting. Evaluated on CICIDS2017, UNSW-NB15, and NSL-KDD, FWLA achieved 98.9% accuracy with a 2.9% false acceptance rate, maintaining 83% accuracy with 20% malicious clients.

3.1.4 Defense Mechanisms Against Poisoning

Several defense approaches have been proposed to counter poisoning attacks.

Basak and Chatterjee [48] developed DPAD, an audit-based defense using KL divergence between client updates and a public audit dataset. On the UCI Heart Disease dataset, DPAD achieved 72% accuracy at a 40% attack rate, with an AUC of 0.71. Rodríguez-Barroso et al. [49] proposed RAB2-DEF, using Local Linear Explanations for explainable client filtering. Against label-flipping attacks on CIFAR-10, RAB2-DEF achieved 0.8510 accuracy, compared to 0.8357 without an attack, while discarding fewer poor clients. Sun et al. [50] introduced FedKC, combining Hessian perturbation defense with KNN-based personalization. The framework mitigated attacks in <7 rounds with <7% accuracy reduction across CIFAR-10, CIFAR-100, PathMNIST, and OrganSMNIST.

3.1.5 Summary and Open Challenges

Poisoning attacks in healthcare FL are increasingly sophisticated and adaptive. Key findings include: (1) robust aggregation methods like Krum outperform simpler approaches under label-flipping attacks; (2) privacy mechanisms such as DP can be exploited to mask poison updates; (3) multi-dimensional detection (direction + sign) offers stronger defense than single-dimensional approaches; (4) adaptive, closed-loop defenses show promise against evolving attack strategies. However, significant challenges remain, including the high computational overhead of advanced defenses, the need for defense strategies that work without auxiliary validation data, and the development of defenses against combined attack vectors.

Table 4 provides a summary of papers on poisoning attacks.

Table 4: Summary of papers on poisoning attacks.

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[30]	Label-flipping poisoning attack	TissueMNIST, OCTMNIST	Blockchain-based FL with SMPC model verification using additive secret sharing; poisoned models filtered before aggregation	With 10% malicious clients: accuracy dropped ~7% (OCTMNIST), ~9% (TissueMNIST); SMV filter recovered up to 25% accuracy in heavily poisoned scenarios (50% malicious clients)	Computational overhead from SMPC and blockchain consensus; energy-intensive; model homogeneity assumption; scalability not extensively tested	Integrated blockchain-SMPC framework; privacy-preserving poisoning defense; decentralized verifiable aggregation
[42]	One-shot poisoning backdoor attacks	Omniglot, mini-ImageNet, CelebA	Adapted matching networks with random Glorot initialization and local fine-tuning; compatible with secure aggregation	Reduces attack accuracy from 90–100% to <20% (Omniglot/mini-ImageNet), ~40% (CelebA); main-task accuracy preserved (>96% on Omniglot)	Reduces main-task accuracy on complex datasets; requires careful tuning of random initialization factor	Backdoor attacks in federated meta-learning; Local defense preserving privacy; adapted matching networks with trainable gating

(Continued)

Table 4 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[48]	Data poisoning attacks	UCI Heart Disease	Audit-based defense using KL divergence between client updates and public audit dataset; dynamic tolerance thresholds	Ranked 1 across all attack rates (10%–40% compromised clients); 72% accuracy at 40% attack rate, AUC 0.71, sensitivity 59%, specificity 83%	Assumes honest server; computational overhead from audit process; depends on public dataset availability	Audit-based detection without raw data sharing; KL divergence for poisoning rate calculation; dynamic tolerance threshold
[39]	Label flipping attacks (targeted data poisoning)	Retinal OCT images	Comparison of robust aggregation methods (FedAvg, Median, Trimmed-Mean, Krum, K-Means) with pre-trained VGG16	Krum most effective: maintains F1-score ~94%–95% even with 70% label flipping (vs. 82%–90% for other methods)	IID data only; single-class targeted attacks only; fixed number of clients (10); no communication overhead analysis	Comparison of multiple robust aggregation methods in FL for medical OCT images; integrates pre-trained VGG16 with FL
[50]	Targeted model poisoning attacks	CIFAR-10, CIFAR-100, PathMNIST, OrganSM-NIST	FedKC = Ant (Hessian perturbation defense) + k-per (KNN-based personalization); base model: MobileNet-V2	Mitigates attacks in <7 rounds (<7% accuracy drop); personalization boosts worst-client accuracy by 2%–11%; effective for new/low-data clients	Privacy risks from datastore; storage/computation overhead for KNN; sensitive to interpolation parameter	Integrated client-side defense and memory-based personalization; client-side Hessian perturbation; KNN-based local memorization
[49]	Byzantine attacks, backdoor attacks	Fed-EMNIST, Fashion MNIST, CIFAR-10	LLEs-based induced ordering function; dynamic linguistic quantifier for client weighting; explainable filtering	CIFAR-10 label-flipping: accuracy 0.8510 (vs. 0.8357 no attack); discards fewer poor clients vs. DDaBA (0.15 vs. 1.03 avg.); backdoor accuracy reduced to 0.0093	Requires server-held validation dataset; only image classification tested; computational overhead from LLE computation	FL defense addressing explainability, fairness, dynamism, and generality simultaneously; uses XAI (LLEs) for client filtering

(Continued)

Table 4 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[40]	Data poisoning via label flipping	MNIST, MIT-BIH Arrhythmia	Three label-flipping strategies tested in SFL; variable malicious client ratio and model split layer depth	Untargeted attacks cause largest degradation (up to 71.31% on ECG); distance-based > targeted attacks; deeper client models more vulnerable	Only image and 1D signal data; honest server assumption; defense only suggested; label flipping only	A comprehensive study of data poisoning in SFL; novel distance-based poisoning attack; evaluates split-layer depth impact
[47]	Poisoning attacks (data/model poisoning)	CICIDS2017, UNSW-NB15, NSL-KDD	Residual testing mechanism; adaptive client weighting; asynchronous training protocol	98.9% accuracy, 2.9% FAR; maintains ~83% accuracy with 20% malicious clients; reduces upload times by 85.2% vs. FedSGD	Scalability concern with many clients; assumes Gaussian residuals; not tested on healthcare data	A framework integrating residual testing, adaptive weighting, and asynchronous training; client-specific dynamic defense; $O(N)$ complexity vs. $O(N^2)$ for Krum
[43]	Targeted poisoning backdoor attacks exploiting DP	Loan, MIAS, MNIST, CIFAR-10	Two optimization modules (poisoning + noise) with GA optimization to align with DP noise; maintains (ϵ, δ) -LDP	Nearly 100% ASR on Loan, MIAS, MNIST; ~45%–90% on CIFAR-10; <3% main task accuracy drop; evades RFA, Trimmed Mean, Multi-Krum, Bulyan	Effectiveness depends on DP noise level; assumes attacker knows DP mechanism; GA adds computational overhead	A work exposing DP noise exploitation for poisoning; multi-objective GA optimization; demonstrates privacy mechanisms can be weaponized
[51]	Data/model poisoning, adversarial attacks	CIFAR-10	EPP-BCFL: three-layer architecture with DP, HE, SMPC, adaptive aggregation, lightweight PoS+BFT consensus	95.2% accuracy; 43% communication overhead reduction; attack resilience improved from ~72.5% to ~93.2%; latency 150 ms for 10K transactions	Evaluated on CIFAR-10, not real EHR data; incentive mechanism not detailed; semi-honest adversarial model	A framework integrating blockchain, hybrid privacy, edge analytics, and adaptive aggregation for EHR; $O(N)$ communication vs. $O(N^2)$

(Continued)

Table 4 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[44]	Backdoor poisoning attacks (Neurotoxin, Vanilla, PGD, DBA)	CIFAR-10, CIFAR-100, EMNIST	DABD: direction alignment (cosine similarity) + sign consensus (majority voting on Top-K parameters) + temperature-controlled SoftMax weighting	87.2% accuracy, 8.1% ASR under Neurotoxin; reduces ASR by 47.1% vs. AlignIns; ROC-AUC 0.94; FPR 3.2%; 1.3× overhead vs. FedAvg	Relies on Byzantine assumption ($\gamma < 1/3$); sign consensus requires plaintext sign access; not tested under encrypted aggregation	Dual-dimensional backdoor detection (global direction + local sign); adaptive soft-weighting without auxiliary data; theoretical geometric separability analysis
[46]	Untargeted model poisoning attacks	MIMIC-III	HyperShield: hypernetwork generated client embeddings + cosine similarity (temporal) + DBSCAN clustering (spatial) + logical AND detection	ROC AUC 0.82–0.90; up to 50% malicious clients; CPU < 7%, RAM < 2.5% on Jetson Nano; 1.4× CPU overhead, 0.7× RAM overhead vs. FedAvg	Lacks fine-grained interpretability; no Sybil defense; assumes stable connectivity; not evaluated against adaptive adversaries	A hypernetwork-based poisoning defense; unsupervised dual-criterion detection (temporal+spatial); personalized FL with built-in robustness; edge-deployable
[41]	Sybil-based label-flipping attacks (targeted data poisoning)	MNIST, CIFAR-10, Kvasir, BloodM-NIST, Dermatology	BlockAuthFL: Hyperledger Fabric (permissioned blockchain), RSA-2048 signatures, dynamic trust weighting (accuracy × samples), rollback to checkpoints	Restores accuracy from ~25% to >80% under sybil attacks; 95.7% accuracy (MNIST), 81.5% (CIFAR-10), 79.0% (Kvasir) with full defense; overhead: +33.6 ms/round	Only label-flipping tested; no gradient poisoning/backdoor; permissioned CA assumption; parameter sensitivity	Blockchain-based authentication + trust weighting + rollback for HFL; closed-loop defense against sybil attacks; component-wise ablation showing necessity of all modules
[52]	Random perturbation, label flipping poisoning	Sleep posture detection	FLDP-DM: hybrid norm + accuracy detection; local sensitivity DP	Maintains high defence at 50% infection rate; outperforms Multi-Krum/Trimmed Mean; DP preserves utility	Requires validation set for accuracy detection; not tested on non-medical data; privacy budget tuning needed	A hybrid defense combining norm-based and accuracy-based detection for both availability and integrity poisoning; local sensitivity for DP

(Continued)

Table 4 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[45]	Adaptive backdoor attacks (DBA, SSB, ALMR, CLFC)	CIFAR-10, FEMNIST, MNIST	GRASP-FL: geometric screening (principal angles), reputation aggregation, layer sanitization (pruning+DP), GAN closed-loop adaptation	Clean accuracy: 97.25% (MNIST), 86.8% (FEMNIST), 80.3% (CIFAR-10); ASR < 10%, FPR < 5%; outperforms Krum, FedATM, ANP	High computational overhead (67.7 s/round); not tested on asynchronous FL; transformer scaling future work	A closed-loop defense integrating geometric, reputation, sanitization and GAN-guided adaptation; formal convergence and separability guarantees
[53]	Label-flipping poisoning	Chest X-ray	MultiMOORA ranking (accuracy, precision, recall, F1, loss); top-k selection; FedAvg+KD stabilisation	F1-score 89.5%; poisoning detection precision 94.7%, recall 60%; outperforms TOP-SIS/COPRAS; no infected client ever aggregated	Recall drops as attacker ratio increases (60% → 27.8% at 30% infected); class imbalance still causes misclassification	An integration of MultiMOORA (MCDM) with FL aggregation; combined FedAvg+KD for stability; competitive client selection

3.2 Reconstruction/Inversion Attacks

Reconstruction and inversion attacks aim to recover original training data from model updates. These attacks pose critical privacy risks in healthcare, as they can reconstruct patient images, waveforms, or other sensitive data with clinically recognizable fidelity.

CryptoGAN [54] combines conditional GANs together with homomorphic encryption (Paillier cryptosystem), providing protection against both gradient and model inversion attacks. Based on ISIC skin lesion (1800 benign, 1497 malignant) and Brain Tumor MRI (7023 images from 4 classes), the authors encrypt only generator parameters in advance of transmission and perform knowledge distillation for images aggregation. CryptoGAN achieves 78.56% accuracy on ISIC and 89.68% on MRI, and reduces reconstruction quality (PSNR) to 4.98 dB (ISIC) and 3.59 dB (MRI), which are still significantly lower than FedAVG (20.96, 23.85 dB). Wang et al. [55] proposed a VAE-based reconstruction method with Laplacian DP noise on to shield IoMT systems from white-box inference attacks in this context. Using the MIT-BIH Arrhythmia Database (48 half-hour ECG recordings of 47 people from 5 heartbeat types), their framework incorporates VAE-based data reconstruction with DP protection and an LSTM-based FL model. It can reduce the probability of data reconstruction by nearly 50% compared to baseline DP methods and achieve 90.9% global model accuracy when $\epsilon = 1.0$. Sumitra et al. [56] it proposed HAFedL, a Hessian-oriented adaptive privacy-preserving horizontal FL-based paradigm capable of preventing a gradient leakage attack. The proposed method is tested on the UNSW-SOSR and Aalto University IoT device-identification datasets, using MNIST and Modified Rényi DP, with heterogeneity-aware noise scaling and Hessian-matrixing to apply layer noise only to the less sensitive weights. HAFedL suffers from high reconstruction loss values (~ 30 – 50) and has F1 scores of 0.88 on UNSW-SOSR and 0.78 on Aalto. A privacy-preserving FL scheme

is based on homomorphic encryption [57] to protect against the loss of the model reconstruction and its inversion. They use the HAM10000 dermatoscopic image dataset (10,015 images across 7 classes) to encrypt a single scalar data-quality value per client per epoch using modified ElGamal encryption, and aggregate that value by quality weight according to the gradient. This approach yields a score of approximately 76.9% with 15.03% more training time than non-HE-based baselines. In [58] they presented KDRSFL, which is a knowledge distillation resistance transfer based framework which protects against model inversion attacks in split federated learning. In CIFAR-100 with transfer to CIFAR10, SVHN, FaceScrub, FMNIST, and MNIST (the three-stage framework), pre-train teacher models by means of attacker-aware training, fine-tune with low-complexity inversion models, and add bottleneck layers. KDRSFL produced 0.058 MSE (reconstruction error) and 67.4% accuracy on the CIFAR-100 dataset, though it achieved a 16% increase in MI resistance while being reported 0.1% lower in accuracy. Jeter et al. [59] proposed OASIS, a modality-independent defense against active reconstruction attacks that uses data augmentation to avoid the need for individual gradient extraction. Using ImageNet, CIFAR100, Wikitext, Stack Overflow, and Shakespeare, the approach is experimented to generate augmented models (rotated images with synonym text replaced) and subsequently activate similar neurons to their original ones. This way, OASIS reduces reconstruction PSNR from >130 dB to 15–20 dB for images; decreases token accuracy to 59% for text; and preserves model accuracy, with a 1.5% decrease relative to the input model. Jiang et al. [60] proposed a shadow counter to gradient inversion attacks with StyleGAN3-based shadow models and interpretability-guided noise injections. Using the ChestX-ray and EyePACS datasets, the method employs fine-tuning shadow models to model the strongest adversarial actors, Grad-CAM++ to identify high-dimensional regions for the task, and dynamic-scale noise. The defense has the best reconstruction metrics while preserving 1% F1 drop compared to no defensive approach. Zheng et al. [61] introduced Sensitivity-Aware Differential Privacy (SDP), which defends against gradient inversion attacks by adapting the privacy protection dynamically as per-sample sensitivity. SDP directly reads the sensitivity of the COVID-19 chest X-ray and REA retinal fundus datasets and computes SDP's sensitivity via simulated GIAs with SSIM, with data allocation per sampling in accordance with the privacy budget. For equivalent privacy risk, the method outperforms PriMIA and MSDP baselines (by about 13.5%) with minimal SSIM (ensuring better privacy). Yuksel and Yilmazer Metin [62] presented a novel real-time ECG anomaly detection method, which integrates federated learning and CKKS homomorphic encryption to secure against gradient inversion attacks. With the PTB-XL set (18,869 patients and 21,799 ECG recordings) in horizontal, vertical, and composite data-partitioning formats, their custom model achieves 99.26% accuracy (without encryption) and 99.02% accuracy (with full privacy protection via DP + CKKS), with only 0.24% degradation. The encryption overhead ranges from 30% (horizontal FL) to 456% (vertical FL) and exhibits distinct performance-security trade-offs.

Table 5 summarizes papers addressing reconstruction/inversion attacks.

Table 5: Summary of papers addressing reconstruction/inversion attacks.

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[54]	Gradient inversion, model inversion attacks	ISIC (1800 benign, 1497 malignant), Brain Tumor MRI (7023 images, 4 classes)	CryptoGAN: cGAN + Paillier HE (encrypts only generator parameters) + knowledge distillation for aggregation	78.56% accuracy (ISIC), 89.68% (MRI); PSNR: 4.98 dB (ISIC), 3.59 dB (MRI) vs. FedAVG (20.96 dB, 23.85 dB)	Computational overhead from HE; slight performance drop on complex datasets; not tested against all adversarial attacks	A combination of HE with GANs in FL for medical data; selective encryption of generator parameters; knowledge distillation integration

(Continued)

Table 5 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[55]	White-box inference attacks (reconstruction)	MIT-BIH Arrhythmia Database (48 half-hour ECG recordings, 5 heartbeat types)	VAE-based reconstruction + Laplacian DP noise + LSTM-based FL + triple-factor incentive mechanism	90.9% accuracy; reduces reconstruction probability by ~50% vs. baseline DP; incentive mechanism scales with data quality	Accuracy-privacy trade-off (~6% loss); increased FL rounds and computational overhead; no communication analysis	A combination of VAE-based reconstruction with Laplacian DP in FL for medical data; triple-factor incentive mechanism (quality, richness, similarity)
[56]	Gradient leakage attacks (Type-0/Type-1)	UNSW-SOSR (10 classes), Aalto University (16 classes), MNIST	Modified Renyi DP with heterogeneity-aware noise scaling; Hessian matrix analysis for weight significance; difference clipping	F1-scores: 0.88 (UNSW-SOSR), 0.78 (Aalto); high reconstruction loss (30–50)	Hessian computation adds overhead; tested on structured/network data; assumes trusted server	An integration of Hessian-aware weight significance analysis with adaptive RDP-based noise; heterogeneity-aware noise scaling
[57]	Model reconstruction, model inversion, honest-but-curious adversaries	HAM10000 (10,015 images, 7 classes)	Double-masking + modified additive homomorphic ElGamal (encrypts only data quality) + quality-weighted aggregation (gradient alignment)	76.9% accuracy; ~15% time overhead vs. non-HE; tolerant to dropouts (8.3% time increase at 30% dropout)	Semi-honest model only; no malicious server defense; ignores client resource heterogeneity; TA dependency	Lightweight HE (encrypts only quality scalar); quality-driven aggregation based on gradient alignment; integrated dropout tolerance
[58]	Training-based model inversion attacks (white-box)	CIFAR-100 (primary), CIFAR-10, SVHN, FaceScrub, FMNIST, MNIST	Three-stage framework: pre-training one-shot distillation, attacker-aware fine-tuning, SFL training with bottleneck layers	MSE 0.058, accuracy 67.4% on CIFAR-100; 16% MI resistance improvement with 0.1% accuracy drop; FLOPs 26.5 M (0.17× vs. FL)	Transfer failure on complex tasks; image quality degradation; only passive MI attacks tested; teacher model dependency	A combination of knowledge distillation with resistance transfer for MI defense in SFL; three-stage training decouples robust training from deployment
[59]	Active reconstruction attacks (RTE, CAH, LOKI) via gradient inversion	ImageNet, CIFAR100, Wikitext, Stack Overflow, Shakespeare	OASIS: data augmentation (images: rotation/shearing; text: synonym replacement/random swapping) to prevent individual gradient extraction	Reduces PSNR from >130 dB to 15–20 dB (images); reduces token accuracy by ~59% (text); maintains accuracy with 1.5% drop	Does not defend against other privacy attacks; static augmentation; only horizontal FL tested	A modality-agnostic defense against active reconstruction attacks; theoretical grounding in gradient inversion; no noise addition (unlike DP)

(Continued)

Table 5 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[60]	Gradient inversion attacks (optimization-based, model-based)	ChestXRray, EyePACS (non-IID partition, 9 clients)	Shadow model-based defense: StyleGAN3 pretraining, pseudo fine-tuning, Grad-CAM++ foreground maps, dynamic noise scaling	Best defense metrics while maintaining 1% F1 drop; LPIPS improved by 0.262 (ChestXRray)	Task performance drop on non-medical images; not tested on large models (ViTs); computational overhead from fine-tuning	A shadow model-based defense using GAN to emulate adversaries; interpretability-guided noise injection (Grad-CAM++); dynamic noise adjustment
[61]	Gradient inversion attacks (primary), multi-attack extension	COVID-19 chest X-ray, REA retinal fundus	Sensitivity-Aware DP: measures per-sample sensitivity via simulated GIAs (SSIM), allocates privacy budget adaptively	~13.5% higher accuracy than PriMIA/MSDP under same privacy; lower SSIM (better privacy)	Computational overhead from local GIA simulation; assumes accurate attack simulation; not yet combined with secure aggregation	Novel attack-driven sensitivity measurement; adaptive DP budget allocation based on empirical privacy risk; extends to multi-attack via weighted sensitivity
[62]	Gradient inversion attacks	PTB-XL (18,869 patients, 21,799 ECG recordings, 3 leads)	CKKS homomorphic encryption + horizontal/vertical/combined FL + custom lightweight DCNN	99.02% accuracy under full privacy (DP+CKKS) vs. 99.26% baseline; overhead: 30% (HFL) to 456% (VFL); per-round latency 635 s (HFL) vs. 15.8 s baseline	High encryption overhead in VFL/combined settings; not evaluated against poisoning/backdoor attacks; no formal complexity analysis	A unified real-time ECG framework integrating FL, CKKS, and multi-data partitioning; selective encryption of gradient updates; clinically validated by cardiologists

3.3 Membership Inference Attacks

Membership inference attacks determine whether a specific individual's record was included in the training dataset. In healthcare, these attacks can reveal patient participation in disease-specific studies, psychiatric units, or clinical trials, violating privacy and research ethics requirements [10].

Barnawi et al. [63] proposed a differentially private federated learning framework to defeat membership inference attacks using private data across different IoMT-related privacy vulnerabilities. Using three chest X-ray datasets (China Set: 326 normal, 336 TB; Montgomery County: 80 normal, 58 TB; Qatar University: 700 normal, 700 TB) for tuberculosis detection, their ensemble, NoisyMax DP with Laplacian noise, is capable of countering membership inference, model inversion, property inference, and model extraction attacks. MobileNet achieved 81.69% and 95% accuracy (Dataset 1 and Dataset 3) and a only 0.71–1.47% F1 score loss in response to DP. The paper in [64] proposed a privacy-preserving FL framework in which ring signatures are used to protect against source inference attacks (an extension of membership inference) in smart healthcare systems. The methods secure model updates using elliptic-curve cryptography ring signatures, with batch validation based on simulated physiological data from 300 users with wearable equipment. This methodology lowers the success rate of SIA to $1/n$, where n is the group size (1% for $n = 100$), with linear signature time

(~1.5 s for $n = 100$) and batch verification time that is close to constant. MetisFL, proposed in [65], is a secure, private federated learning architecture for multi-institution neuroimaging analysis without sharing sensitive patient information. It applies FHE (fully homomorphic encryption) with a low 7% runtime overhead for secure model aggregation, preventing external threats and an adversarial controller from compromising the target. To safeguard against insider membership inference attacks, it mitigates gradient noise in the process of training. Evaluation of brain age estimation (BrainAGE) based on UK Biobank data and Alzheimer's disease (ADNI, OASIS-3, and AIBL) datasets shows that MetisFL models perform on par with centralized learning in IID settings and still demonstrate good accuracy across a wide range of data. The unique addition is the dual-protection implementation, which applies FHE and gradient noise, but has limited performance differences in data distributions that are less than IID; vertical federated learning is proposed for future work.

In [66], Jiang et al. introduced FHDM, a hybrid differential privacy scheme that accounted for Laplace and Gaussian noise, was introduced to protect it from membership inference and reconstruction attacks. Compared to: PneumoniaMNIST (4708 training samples), BloodMNIST (11,959 training) and PathMNIST (89,996 training) the approach is to randomly split the models parametric parts and applies different noise types at different parts. FHDM minimizes the correctness of any membership inference attack to less than 50% random guessing and has an average accuracy rate of 9.60% higher than the standard LDP-derived schemes that have budgets $\epsilon = 1$. MedGuard-FL [67] presents the risk-aware multi-layer adaptive security scheme that addresses membership inference, data poisoning, label-flipping, and backdoor attacks. Using the pretreated ($\sim 40,000 \times$ wearable-measurable) MIMIC-III clinical database, DP ϵ budget level (0.2–1.2) on user risk status (NEWS2) is dynamically controlled. The XGBoost-based system achieves an approximate global accuracy of 0.87 and a very near-random membership inference AUC (0.49–0.56). A summary of papers addressing membership inference attacks is available in Table 6.

Table 6: Summary of papers addressing membership inference attacks.

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[63]	Model inversion, membership inference, property inference, model extraction attacks	Three chest X-ray datasets: China Set (326N/336T), Montgomery County (80N/58T), Qatar Univ. (700N/700T)	Ensemble-based NoisyMax DP with Laplacian noise; multiple CNN architectures evaluated (VGG16, ResNet50, MobileNet, etc.)	MobileNet best: 81.69% (Dataset 1), 95% (Dataset 3); DP impact: 0.71%–1.47% F1 loss; inference: ~2 sec/image (CPU), ~23 images/sec (GPU)	Scalability issues; DP overhead may strain IoMT devices; use case specificity; requires updates for evolving threats	An integration of ensemble-based NoisyMax DP within FL for IoMT; privacy-preserving ensemble training; comprehensive evaluation of 7 CNNs across three tuberculosis (TB) datasets
[64]	Source inference attacks (extension of membership inference)	Simulated physiological data from wearables (up to 300 users)	Ring signatures (elliptic curve cryptography) with batch verification; dynamic grouping; trusted signature server	Reduces SIA success rate to 1/n (1% for $n = 100$); signature time ~1.5 s for $n = 100$; near-constant batch verification time; extra overhead <20 kbps for $n = 100$	No real-world deployment; depends on trusted server; fixed group size; no incentive mechanism; horizontal FL only	An application of ring signatures to defend against source inference attacks in FL for healthcare; improved batch verification; dynamic user grouping

(Continued)

Table 6 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[65]	Outsider attacks, Insider membership inference	UK Biobank (BrainAGE), ADNI, OASIS-3, AIBL	MetisFL framework with FHE (CKKS) and gradient noise defense; FedAvg variants	FHE adds ~7% overhead; matches centralized performance in IID; AUC-ROC ~0.897; reduces membership inference success	Performance gap in non-IID settings; no vertical FL yet; FHE scalability challenges	A neuroimaging FL framework combining FHE and gradient noise for dual protection
[66]	Membership inference, reconstruction attacks	Pneumonia MNIST (4708 training), BloodMNIST (11,959), PathMNIST (89,996)	FHDM: hybrid Laplace-Gaussian DP without increasing privacy budget; parameter splitting; loss/iteration-aware noise adaptation	Reduces MIA accuracy to <50% (random guessing); ~9.60% higher accuracy than LDP schemes; minimal runtime overhead (0.03–0.56 s per client)	2D low-resolution images only; honest-but-curious server; algorithmic complexity; not tested on highly imbalanced data	A hybrid Laplace-Gaussian DP mechanism for FL without increasing privacy budget; loss-aware and iteration-aware noise adaptation; formal privacy proofs
[67]	Membership inference attacks, data poisoning, label-flipping, backdoor attacks	MIMIC-III (~40,000 ICU records) processed to wearable vitals, NEWS2 classification	MedGuard-FL: three-layer adaptive framework with dynamic DP, anomaly detection, Byzantine-resilient aggregation	Global accuracy ~0.87; MIA AUC ~0.49–0.56 (near-random); 0% ASR for all attack types; detection precision/recall/F1 = 1.0; <2.5% security overhead	Small federation (3 nodes); focuses device-edge privacy; assumes moderate resources; structured data only	A context-aware FL framework for RPM integrating patient risk/status; multi-layer defense with dynamic parameter tuning; clinically-grounded adaptation

3.4 Adversarial and Side-Channel Attacks

Adversarial evasion attacks manipulate model inputs to cause misclassification, while side-channel attacks exploit physical characteristics (e.g., motion sensors, power consumption) to infer sensitive information. Both pose significant risks in healthcare settings.

The paper [68] presented a hardware-accelerated federated learning architecture implemented as a fuzzy intuition-based security mechanism to overcome various attacks, including membership inference, model inversion, data leakage, and side-channel attacks. Using eye-tracking data from 150 radiologists, they also verified more than 15,000 distributed CT scans at eight healthcare facilities. Furthermore, they have discussed integrating intuitionistic fuzzy sets with FPGA accelerators, hypergraph fusion, and homomorphic encryption cores to enable adaptive, risk-aware encryption. Among the high-level results are 94.2% critical region retention (15.3% improvement over the baseline), processing latency of 23.97 ms per CT slice (3.5× faster than the GPU), average power consumption of 19.8 W (43% lower than the GPU), and a 130-bit-equivalent security level (2.8× stronger than traditional methods). However, the approach has its own limitations, namely modality-specific validation (CT scan is the main aspect), dependence of FPGA deployment on changing hardware for GPU/CPU portability, tedious parameter tuning for intuitionistic fuzzy threshold settings, DP parameters, and minimal testing for scalability in fewer than 8 institutions. The novelty was to combine intuitionistic fuzzy sets and federated learning with medical imaging, implement low-overhead hardware-accelerated IFS security on FPGA, model complex anatomical relationships using

hypergraph-based feature fusion, and build adaptive risk-aware encryption based on the membership, non-membership, and hesitation for dynamic (by-product) security-diagnostic trade-offs. Darzi et al. [69] studied adversarial evasion attacks on federated learning for medical imaging as the transferability of attacks (FGSM, PGD, C&W, B&B, DAA, and DDN) between client types in the FL networks. For example, the work utilizes brain tumor MRI images (~1400 per tumor type) and histopathologic cancer detection images (~2150 samples), distributed in a non-IID fashion across clients. The algorithm used FedAvg with various privacy levels ($\epsilon = 0.9$, $\delta = 10^{-5}$) and Gaussian noise ($\sigma = 2.0$), with gradient clipping ($C = 1.5$), to verify the attack's transferability under DP limitations. Some highlights were that pathology images exhibited greater susceptibility than MRI images (up to 98.97% attack success rate vs. 88.78%), achieved high overall transferability over 70% among clients, found optimal perturbation sizes around 0.03–0.05 to balance attack success against stealth, and demonstrated that DP reduced attack effectiveness but did not eliminate adversarial vulnerability. Limitations include the 3-client limitation, loss of utility due to DP noise, the assumption of white-box access to the model design and weights, domain-specific results that might not translate to a new and distinct medical imaging platform, and a lack of an exhaustive defense assessment. Here, they have introduced the adversarial transferability of FL in medical imaging using DP, demonstrating specific pathology-MRI imagery modality, building a realistic threat model with a single malicious client of the attacks, and performing test-time attacks without compromising the robustness of the training while presenting a statistical analysis of principal attack parameters in the federated medical context.

Federated learning is applied to protect privacy against smartphone-based side-channel attacks (motion-based keystroke inference attacks with accelerometer, gyroscope & magnetometer sensors exploitation) [70]. The dataset comprises sensor data obtained for 10 Android smartphone users in three different scenarios walking, sitting, standing, which were performed in 26 alphabets continuously for 2.0 min. The method was a deep neural network model employing 4 dense layers (256, 128, 64, 32 units), 5 dropout layers (rate = 0.2), and ReLU/softmax activations, implemented in a horizontal FL framework with FedAvg aggregation, using 2 clients. The aggregated features demonstrated a global accuracy of 80.09%. The best client achieved 79% accuracy, precision 78%, recall 79%, F1-score 75%, and effective training 79.06% for all clients on a client-to-client basis, with a total training time of 50.57 min across 3 FL rounds. These limitations reflect low levels of minimal small-scale experimentation, with only 2 clients and 10 users; non-representation/correspondence among smartphone models; unresolved trust problems with the central server; and no attempts at adversarial attacks in the FL framework. The novelty of the work is to use federated learning for detecting motion-based side-channel attacks without compromising data privacy for healthcare settings, and merging FL with DNNs for cooperative learning, without putting raw sensitive data into shared data, and focusing on the Motion-based Keystroke Inference attack in the side-channel domain, particularly with a focus on healthcare data. Naeem et al. [71] proposed an adversarial-aware FL pipeline with dynamic client weighting to protect against PGD-based adversarial evasion attacks. Using UCI Heart Disease (303 patients, 14 features) and Breast Cancer Wisconsin (569 samples, 30 features) datasets, their approach integrates hospital-side DPSO feature selection, temporal DCNN local training combined with PGD robustness probing, and server-side dynamic aggregation based on accuracy $\times$ sample count. Under a clean data environment, we get 94.44% accuracy (Heart Disease) and 97.36% (Breast Cancer), and 90.74%/95.61% under PGD attack while outperforming FedAvg by 3%–4%. Matthew et al. [72] introduced a context-appropriate AI model for secure medical image analysis on IoMT systems, addressing adversarial evasion attacks (FGSM, PGD) and sensor compromise. The framework comprises three key components: (1) an improved Dynamic Confidence-Weighted Attention (DCWA) module for interpretable feature extraction, (2) a Semantic Reasoning Module (SRM) with 60 SWRL rules for enforcing clinical constraints, and (3) a Resilient Observability and Detection Engine (RODE) that relies on sparse observability to detect adversarial

perturbations. Evaluated on CheXpert, RSNA Pneumonia, and NIH Chest X-ray datasets, the framework achieves 94.8% accuracy, reduces robustness loss to 15.4% (FGSM) and 18.3% (PGD), and increases the explainability index by 20% over state-of-the-art XAI methods. Table 7 summarizes papers addressing adversarial and side-channel attacks.

Table 7: Summary of papers addressing adversarial and side-channel attacks.

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[68]	Membership inference, model inversion, data leakage, side-channel attacks	Distributed CT scans (>15,000 images from 8 institutions); eye-tracking from 150 radiologists	Intuitionistic fuzzy sets + FPGA acceleration + hypergraph feature fusion + HE cores	94.2% critical region retention (15.3% improvement); 23.97 ms/slice (3.5× faster than GPU); 19.8 W avg. power (43% reduction); 130-bit security (2.8× stronger)	Modality-specific (CT scans); requires FPGA; complex parameter tuning; tested on 8 institutions; image data only	An integration of intuitionistic fuzzy sets with FL for medical imaging; hardware-accelerated IFS on FPGA; hypergraph-based feature fusion; adaptive risk-aware encryption
[69]	Adversarial evasion attacks (FGSM, PGD, C&W, B&B, DAA, DDN)	Brain tumor MRI (~1400 per type), histopathologic cancer detection (2150 images), non-IID distribution	FedAvg with DP, Gaussian noise, gradient clipping ($C = 1.5$); attack transferability analysis	Pathology more susceptible than MRI (up to 98.97% ASR vs. 88.78%); high transferability (>70%); optimal perturbation 0.03–0.05; DP reduces but doesn't eliminate attacks	Limited to 3 clients; DP degrades utility; white-box access assumed; domain-specific findings	An in-depth study of adversarial transferability in FL for medical imaging under DP; highlights modality-specific vulnerabilities; realistic threat model (single malicious client)
[70]	Side-channel attacks (motion-based keystroke inference)	Sensor data from 10 Android users (accelerometer, gyroscope, magnetometer), 3 postures, 26 alphabets	DNN with 4 dense layers (2,561,286,432 units), 5 dropout layers (rate = 0.2), FedAvg aggregation across 2 clients	Global accuracy 80.09%; best client: accuracy 79%, precision 78%, recall 79%, F1 75%; training time ~50.57 min (3 rounds)	Small-scale (2 clients, 10 users); limited device models; trust issues with server; no adversarial attack exploration	An application of FL for detecting motion-based side-channel attacks in healthcare; integrates FL with DNN for collaborative learning without raw data sharing

(Continued)

Table 7 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[71]	Adversarial evasion attacks (PGD)	UCI Heart Disease (303 patients, 14 features, binary), UCI Breast Cancer (569 samples, 30 features)	DPSO feature selection + temporal DCNN + PGD robustness probing + dynamic weighted aggregation (accuracy $\times$ samples)	Heart Disease: 94.44% (clean), 90.74% (PGD); Breast Cancer: 97.36% (clean), 95.61% (PGD); outperforms FedAvg by 3%–4%	Single attack type (PGD); no poisoning/backdoor defense; IID partition assumption; small datasets; no formal complexity analysis	Attack-aware FL pipeline with DPSO feature selection; dynamic weighting by accuracy $\times$ samples; PGD-based robustness evaluation during training; lightweight edge deployment
[72]	Adversarial evasion (FGSM, PGD), sensor compromise	CheXpert (224k images), RSNA Pneumonia, NIH Chest X-ray (100k images, 14 diseases)	SAFE (CNN) + ODRE (ontology+600 rules) + Enhanced DCWA + RODE (sparse observability)	Accuracy: 94.8% (CheXpert), 92.3% (RSNA), 89.7% (NIH); robustness loss: 15.4% (FGSM), 18.3% (PGD); explainability index 82.4% vs. 68.3% baseline	Rule construction requires expert time (40 h); rare disease performance lower (80.2% alignment); white-box attack focus	A hybrid of deep learning, symbolic reasoning and sparse observability for XAI + robustness; Enhanced DCWA with adaptive normalisation; theoretical attack detectability guarantees

3.5 Multiple/Comprehensive Attack Types

Contemporary threats often combine multiple attack vectors or exploit interactions among different vulnerabilities. These comprehensive attacks pose particularly significant challenges to current defense mechanisms.

Kalodanis et al. [73] also present a Privacy-Preserving Federated Learning framework in which the methods of re-identification, reconstruction, property inferencing, model extraction and membership inferencing, model inversion, shadow model, adversarial ML, membership memorization, and model reuse attacks are combined. This architecture has a Federated Learning Coordinator, a Secure Computation Layer (Homomorphic encryption, SMPC, DP), and a Compliance and Explainability Module, in addition to an Adaptive Crypto Orchestrator algorithm, which is used to dynamically alter encryption on the fly based on real-time risk metrics, resource constraints, and clinical needs. FedMedSecure [74], a revolutionary framework that seamlessly combines Federated Learning, Few-Shot Learning, Explainable AI, and Differential Privacy for collaborative IoMT cybersecurity. It has set high benchmarks in detection accuracy, attaining 99.9% accuracy in the supervised setting and approximately 99.7%–99.8% accuracy in the reported few-shot settings. It allows interpretable predictions, especially in data-poor few-shot settings. The evaluation setting and adversarial robustness represent its major weaknesses. Alharbi [75] proposed a new trusted collaborative

AI model for sensitive domains such as healthcare, with real-world applications in the finance, infrastructure, and government sectors. The authors present a self-aware, privacy-preserving, and user-friendly federated transfer learning architecture for intrusion detection in IoMT. It provides an efficient solution for non-IID data issues, addresses more attack types (including zero-day attacks), achieves high detection accuracy, and is stable across silos. More investigation on adversarial robustness and blockchain security would be useful. Nassar et al. Proposed a new federated learning framework, PRIFLEX [76], aimed at tackling the question of privacy, fairness, and interpretability in multimodal healthcare AI. PRIFLEX combines the raw ECG (PTB-XL) and clinical text (MIMIC-IV) data to simulate heterogeneous hospital environments and assesses three privacy attacks (DLG, MIA, Model Inversion). Its main innovation is DP+SA, a hybrid defensive measure that inhibits attack success by up to 84% when using the utility (AUROC up to 0.869), with equal fairness, while still protecting against attack success. The framework further supports this era of privacy-aware explainability by measuring how defenses impact explainability, such as ARC, which quantifies the extent to which defense mechanisms undermine explainability. Despite limitations (simulation-only, two modalities), PRIFLEX sets a new benchmark of safe, equitable, and explainable multimodal FL in healthcare. Azad and Shah [77] proposed a privacy-preserving FL system that combines Shamir secret sharing with zero-knowledge proofs to defend against inference, membership, and poisoning attacks. This protocol ensures that no single server can reconstruct client updates (information-theoretic security) and that malformed updates cannot be submitted from their end. In considering the 2023 CICIoT dataset with linear regression, MLP, and CNN models, the system achieves a high level of privacy while requiring significantly less overhead—client compute $O(dk\tau + d)$, communication $O(dk+d)$, overhead is 1.3 times in comparison to non-secure FL. This technique also prevents collusion among up to $\tau - 1$ aggregation servers. Singh et al. [78] presented MIoT-FL, which is a federated learning architecture for ECG-based heart disease prediction characterized by a combination of differential privacy (gradient clipping + Gaussian noise), secure aggregation (with pairwise Diffie-Hellman masks) and faster authentication. Applied on the MIT-BIH Arrhythmia dataset (48 recordings, 47 patients), MIoT-FL achieves an accuracy of 84.92%, 89.32% and 98.21% over privacy budgets $\epsilon = 1, 2, 4$, respectively, and does not suffer eavesdropping, poisoning or inference attacks (membership + gradient inversion). Phan and Nguyen [79] introduced a secure multi-modal federated transfer learning model of ECG classification that combines CKKS homomorphic encryption, differential privacy, and transfer learning with frozen pre-trained layers. Leveraging PTB-XL pre-training and MIT-BIH fine-tuning (24 clients, non-IID patient-based partition), the framework achieves 99.26% baseline accuracy and 99.02% accuracy with complete privacy, with a formal convergence proof to HE-based distributed FTL under non-IID settings and AdamW optimization, and is resistant to inference, poisoning, and post-quantum attacks. Melhem et al. [80] introduced C2SecFL, a clustered federated learning framework for healthcare analytics that addresses various types of attacks, including poisoning (model poisoning), membership inference, data reconstruction, model inversion, and evasion. It combines data-aware clustering (CNN, LSTM, GRU), a decentralized functional encryption scheme (FedFort) using IPFE without a key distribution center, a permissioned blockchain using QBFT consensus, and a participant-aware voting ensemble (PAVE). Experiments using BRFSS heart attack data (435k samples, 95 features) show that C2SecFL can reduce the attack success rate from 72% (FedAvg) to 13.7%, achieve 94.1% accuracy, and deliver $3.6\times$ faster encryption quality compared to homomorphic encryption baselines.

Zhao et al. [81] proposed XAI-EdgeSFL, an explainable split-federated learning framework for consumer healthcare intrusion detection that is robust to model poisoning and inference attacks (eavesdropping, gradient inversion). It employs a lightweight edge-device-based temporal convolutional network (TCN), a server-side classifier, adaptive client weighting based on local model accuracy, and a multimodal XAI package (SHAP, LIME, and Integrated Gradients). Examined on CICIoMT2024, N-BaIoT, and CICIDS2017 datasets,

it achieves accuracy as high as 99.5% on N-BaIoT and lowers detection delay to 2.14 s (vs. 4.82 s for FedAvg). The framework offers feature-level explanations of clinical trust while storing raw data on devices. [Table 8](#) highlights the Summary of the paper addressing multiple/comprehensive attack types.

Table 8: Summary of papers addressing multiple/comprehensive attack types.

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[73]	Comprehensive multiple types of attacks	No specific dataset	PPFL architecture: Federated Learning Coordinator, Secure Computation Layer, Compliance & Explainability Module	Theoretical insights only; comprehensive attack taxonomy; architectural blueprint; qualitative benefits: enhanced compliance, dynamic risk adaptation, improved trust via explainability	No quantitative results; conceptual only; privacy-utility trade-off; architectural complexity; explainability vs. encryption tension	Holistic regulatory alignment with GDPR/EU AI Act; comprehensive attack taxonomy; adaptive cryptographic orchestration (privacy-on-demand); integrated compliance & explainability
[75]	Scanning, DoS/DDoS, MITM, XSS, Injection, Backdoor, Password Attacks, Ransomware, Zero-day	CIC-ToN-IoT Dataset (>5M samples)	Personalized Federated Transfer Learning (FTL), DNN, FedAvg	Accuracy: 95%–99%, Zero-day detection recall +40%, Non-IID robust	Requires initial global model, Single dataset only, No adversarial testing	A cross-silo FTL-based IDS for IoMT, Personalized FL for non-IID, Zero-day detection with privacy
[74]	19 attack types in CIClo MT2024, 5 semantic groups (DDoS, DoS, Reconnaissance, Protocol Attacks)	CICloMT2024 (8.7M samples), CIDC2017 (2.8M samples)	FedMedSecure: FL + Few-Shot Learning, Multi-model ensemble, DP, XAI	99.9% accuracy (supervised), 99.7–99.8% (few-shot), 75% communication reduction	Simulated FL environment, High computational demand, Vulnerable to Byzantine attacks	A Federated Few-Shot Learning for IoMT, CrossTransformer with learnable queries, Semantic attack clustering
[76]	Deep Leakage from Gradients, Membership Inference, Model Inversion	PTB-XL (ECG), MIMIC-IV v2.2 (clinical notes), 20 heterogeneous clients	PRIFLEX: Multimodal FL, Early/Late/Hybrid fusion, DP + Secure Aggregation, XAI	AUROC: 0.869, DLG reduced by 84% (hybrid defense), Fairness improved	Simulated clients only, Limited to ECG and text, High system overhead	A comprehensive multimodal FL for healthcare, Hybrid defense benchmarking, Privacy-aware explainability metrics

(Continued)

Table 8 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[77]	Inference, membership, poisoning attacks (gradient inversion, model poisoning, collusion)	2023 CICIoT dataset (intrusion detection) with LR, MLP, CNN	Shamir secret sharing + zero-knowledge proofs (Pedersen commitments, Σ -protocol); threshold τ , k aggregation servers	Client compute $O(dk\tau+d)$; communication $O(dk+d)$; 1.3 $\times$ overhead vs. non-secure FL; collusion resistance up to $\tau - 1$ servers	DoS via share withholding (limitation); Sybil attacks not fully addressed; requires end-to-end encryption for sign consensus	A combination of secret sharing + ZKPs in FL; information-theoretic confidentiality + verifiable integrity; rigorous security model for malicious adversaries
[78]	Eavesdropping, poisoning, inference (membership + gradient inversion)	MIT-BIH Arrhythmia (48 recordings, 47 patients)	DP (gradient clipping + Gaussian noise) + secure aggregation (pairwise Diffie-Hellman masks) + expedited authentication	Accuracy: 84.92% ($\epsilon = 1$), 89.32% ($\epsilon = 2$), 98.21% ($\epsilon = 4$); scales from 10 clients (94.3%) to 100 (70.2%)	Accuracy drop with more clients; no comparison with adaptive attacks; no computational overhead measurement	A MIoT-FL framework integrating DP, secure aggregation, and authentication; defends against three attack types simultaneously
[79]	Inference (gradient inversion, membership), poisoning, post-quantum attacks	PTB-XL (pre-training), MIT-BIH (fine-tuning, 24 clients, non-IID)	CKKS homomorphic encryption + DP + transfer learning (frozen layers) + AdamW; hybrid encrypted/plaintext architecture	99.26% baseline, 99.02% under full privacy (0.24% drop); training time 548 s vs. 15 s baseline; formal convergence proof	High computational overhead (36 $\times$); communication expansion (14.6 $\times$); threat model excludes colluding clients	Convergence proof for HE-integrated FTL under non-IID with DP; multi-modal (ECG signals + spectrograms); comprehensive security analysis including post-quantum
[80]	Model poisoning, membership inference, data reconstruction, model inversion, evasion	BRFSS heart attack dataset (246k $\rightarrow$ 435k samples, 95 features, binary classification)	C2SecFL: data-aware clustering (CNN/LSTM/GRU) + FedFort (IPFE, no-KDC) + permissioned blockchain (QBFT) + PAVE ensemble	ASR 13.7% (vs. 72% FedAvg); accuracy 94.1%; encryption 3.6 $\times$ faster than HE; aggregation 2.2 $\times$ faster; communication 9.1 $\times$ lower	Blockchain latency 216 s/round; requires reliable validators; not tested on multi-modal data	Integration of clustering + IPFE (no KDC) + permissioned blockchain for healthcare FL; formal robustness thresholds; decentralised key generation

(Continued)

Table 8 (continued)

Ref.	Attack Type	Dataset	Method	Key Results	Limitations	Novelty
[81]	Model poisoning (by low-accuracy clients), inference attacks (eavesdropping, gradient inversion)	CICloMT2024 (5 attack types), N-BaIoT (5 attack types), CICIDS2017 (4 classes)	XAI-EdgeSFL: lightweight TCN on edge, server classifier, adaptive aggregation (accuracy-based weights), XAI (SHAP/LIME/IG)	Accuracy: 86.15% (CICloMT), 93.73% (N-BaIoT), 99.46% (CICIDS2017); detection delay 2.14 s; 82% fewer parameters than standard TCN; inference 35.6 ms	IID partition assumed in core experiments; no cryptographic protection for updates; performance drops with more clients (99.5% → 88.0% at 10 clients)	A XAI + split FL for healthcare intrusion; lightweight multi-head TCN with attention; adaptive performance-based aggregation; multi-modal XAI suite

4 Analysis and Discussions

4.1 Cross-Family Synthesis of Defenses and Recurring Failure Modes

The included studies do not support a simple ranking in which a single defense family dominates. Instead, they show a common alignment problem: a mechanism is effective only if its protection layer matches the adversary's point of attack. [Table 9](#) compares the main defense families across the evidence base.

Table 9: Cross-study synthesis of defense families, strengths, and recurring failure modes.

Defense Family	Consistent Strength	Recurring Weakness or Failure Mode	Implication for Healthcare FL
Differential privacy	Provides an explicit, tunable bound on record- or update-level disclosure and can be allocated according to sample sensitivity [61,66].	Noise reduces clinical utility, composes across rounds, and may conceal small malicious deviations; DP-Poison demonstrates that an attacker can exploit the expected noise envelope [43].	Privacy budgets must be coupled to integrity monitoring, per-round accounting, and clinically defined utility floors.
Cryptographic protection (HE, SMPC, secret sharing, secure aggregation)	Protects update contents from an honest-but-curious server or network observer and supports auditable aggregation [62,77,79].	Encryption does not determine whether an encrypted update is malicious; threshold collusion, key-management assumptions, and substantial computation or communication overhead remain.	Confidentiality and integrity must be treated separately; encrypted aggregation still requires authentication, robust validation, and recovery procedures.
Robust aggregation and anomaly detection	Can suppress gross poisoning without encrypting every operation and can adapt to client behavior [44,46,49].	Adaptive attackers imitate benign directions or signs, sybils amplify a coordinated update, and non-IID clinical data can resemble malicious outliers, increasing false positives.	Detectors should use temporal and multi-view evidence and report errors by clinical subgroup, not only global attack-success rate.

(Continued)

Table 9 (continued)

Defense Family	Consistent Strength	Recurring Weakness or Failure Mode	Implication for Healthcare FL
Blockchain, reputation, and authentication	Improves provenance, identity management, auditability, and rollback after detected compromise [30,41].	An immutable record can preserve a malicious but correctly authenticated update; reputation can be gamed by collusion or slow-onset attacks and adds systems overhead.	Ledger mechanisms are governance layers rather than stand-alone privacy defenses and require semantic update validation.
Personalization, transfer learning, and knowledge distillation	Reduces the effect of heterogeneity and can limit exposure of full model states [50,58,75].	Guarantees are architecture- and task-dependent; transferred representations can still leak information, and personalized behavior complicates anomaly thresholds.	Benefits should be assessed under site-specific class imbalance, subgroup performance, and realistic auxiliary knowledge.
Adaptive hybrid defenses	Combine complementary confidentiality, privacy, integrity, and trust mechanisms and are the most plausible response to compound attacks [45,52,67].	Additional components create tuning interactions, correlated failure, higher latency, and a larger implementation and audit surface; few studies test truly compound adaptive adversaries.	Hybrid systems require ablation, end-to-end overhead reporting, and red-team evaluation against attacks that adapt to the complete pipeline.

There are five reasons why defenses fail time and again in different studies. First, layer mismatch occurs when confidentiality controls are expected to provide integrity, and poisoning filters are expected to prevent reconstruction. Second, static threat assumptions make fixed thresholds and fixed privacy budgets brittle when adversaries adapt over rounds. Third, non-IID ambiguity causes real hospital-specific updates to appear to be attacks, thereby creating a privacy–fairness–robustness conflict. Fourth, single-objective evaluation rewards accuracy or attack suppression while excluding privacy composition, subgroup harm, latency, or false alarms. Fifth, composition failure occurs because individually sound components are seldom considered as a complete pipeline under collusion, repeated participation, and cross-layer attacks. These patterns explain why stronger cryptography or more aggressive filtering alone does not always solve the healthcare FL security problem.

In this section, a multidimensional analysis of the identified privacy threats from the selected literature is presented, framed through three perspectives: technical and methodological analyses, context-specific healthcare, and general trends in meta-analysis. The review proceeds systematically to outline attack modes, evaluate their functioning and actual consequences in clinical environments, and investigate the evolving domain of defense and its shortcomings. In combination with these insights, we present an overview of the current landscape and the major deficiencies in securing Federated Learning systems for healthcare applications.

4.2 Foundational (Technical & Methodological) Perspectives

This section examines the various attacks, their efficacy, and their interactions with defenses reported across the 45 papers analyzed above. The analysis is divided into three core aspects: the taxonomy and mechanisms of attacks, their quantitative and qualitative efficacy, and their robustness against or ability to evade common FL defenses.

4.2.1 Attack Taxonomy & Mechanism

The literature shows that FL systems are among the targets of many sophisticated attacks in the healthcare industry. The types of attacks are summarized in Table 10 for their *surface* (point of exploitation), *capability* of the adversary, and *knowledge* needed to execute.

Table 10: Taxonomy and mechanisms of attacks in federated learning for healthcare.

Attack Surface	Adversary Capability	Required Knowledge	Ref.
Data (Re-ID, Reconstruction, Property), Model (Extraction, MIA, Inversion, Shadow, Adversarial, MMA, Reuse)	Honest-but-curious server, Malicious participants, External adversary	White-box (gradients, params) & Black-box (outputs)	[73]
Model Poisoning (Targeted Backdoor)	Malicious participant(s)	White-box (model architecture, DP mechanism)	[43]
Inference (Membership)	External adversary	Outsider attacks, Insider membership inference	[65]
Model (Adversarial Evasion)	Malicious participant (single client)	White-box (local model weights)	[69]
Data Poisoning (Label Flipping)	Malicious participant(s)	None (local data manipulation only)	[39]
Data Poisoning (Label Flipping)	Malicious participant(s)	None (model-agnostic)	[40]
Data Reconstruction (Active Gradient Inversion)	Malicious (dishonest) server	White-box (model architecture, injected layer)	[59]
Data Reconstruction (Gradient Inversion)	Honest-but-curious server	White-box (gradients)	[61]
Data Reconstruction (Gradient Inversion-Opt & Model-based)	Malicious server	White-box (gradients, BN statistics)	[60]
Data Poisoning (Backdoor)	Malicious participant	White-box (meta-learning framework)	[42]
Model Poisoning (Generic)	Malicious participant(s)	White-box (global model benchmark)	[47]
Inference (Gradient Inversion)	Honest-but-curious server	White-box (gradients, model parameters)	[62]
Poisoning (Backdoor, Neurotoxin)	Malicious participant(s)	White-box (model architecture, gradient sign)	[44]
Poisoning (Untargeted Model Poisoning)	Malicious participant(s)	White-box (local model)	[46]
Poisoning (Sybil-based label-flipping)	Malicious participant(s) with multiple identities	None (local data manipulation)	[41]

(Continued)

Table 10 (continued)

Attack Surface	Adversary Capability	Required Knowledge	Ref.
Adversarial Evasion (PGD)	Malicious participant (single client)	White-box (local model)	[71]
Eavesdropping, Poisoning, Inference (Membership + Gradient Inversion)	Honest-but-curious server, Malicious participants	White-box (gradients, model updates)	[78]
Inference, Membership, Poisoning (Gradient inversion, model poisoning), Collusion	Malicious participants, aggregators	White-box (gradients, shares)	[77]
Inference (Gradient inversion, Membership), Poisoning, Post-quantum	Honest-but-curious server, Malicious participants	White-box (encrypted gradients)	[79]
Data Poisoning (Random Perturbation + Label Flipping)	Malicious participant(s)	White-box (local model)	[52]
Adaptive Backdoor (DBA, SSB, ALMR, CLFC)	Malicious participant(s)	White-box (full knowledge)	[45]
Label-Flipping (multi-metric ranking)	Malicious participant(s)	None (model-agnostic)	[53]
Multiple: Poisoning, Membership Inference, Model Inversion, Evasion	Malicious participant + honest-but-curious server	White-box (gradients, model)	[80]

The primary attack surfaces are *data* (poisoning, reconstruction) and *model* (inference, extraction, evasion). The adversary is most frequently a *malicious participant* (client) or a *malicious/honest-but-curious server*. One notable trend is the weaponization of privacy mechanisms, which can be observed in [43], where Differential Privacy (DP) noise is exploited to hide poisoning attacks. Data poisoning often calls for zero prior knowledge; advanced reconstruction and inference attacks have the white-box access advantage over either gradients or the internals of models, although there are black-box adaptations [82]. New works like [44] show that even white-box access to gradient signs can work to implement stealthy backdoor attacks (e.g., Neurotoxin), also [77,79] show that adversarial servers have similar access to the gradient but can do inference and reconstruction, emphasizing the requirement of robust cryptographic defenses.

4.2.2 Assessment of the Efficacy of the Attack

The effectiveness of such attacks on medical machine learning should be assessed not only by numerical performance, but also by the clinical and operational impacts of adversarial attacks. In literature, three generalities have mostly defined the characteristic of efficacy of the attack: (1) quantitative success metrics (measures the extent to which the attack fulfills its technical purpose), (2) qualitative or clinical meaning (implications for clinical contexts of the attack, e.g., patient safety issues, reliability of diagnostics, privacy) and (3) computational efficiency (easiness of running the attack or ability to adopt the attack in clinical or adversarial settings). From a quantitative standpoint, attacks employ metrics appropriate with regard

to their threat model. Attack success rates vary substantially according to attack strategy, threat model, dataset, malicious-client proportion, and defense configuration. Under the specific experimental conditions evaluated by Zheng et al. [43], DP-Poison achieved an ASR close to 100% while causing less than a 3% reduction in main-task accuracy. For example, GRASP-FL achieved an ASR below 10% with up to 3% malicious clients while employing geometric screening, reputation-based aggregation, layer-wise sanitization, and GAN-guided adaptation [45]. Applying the Neurotoxin backdoor attack achieves higher ASR while still maintaining main-task accuracy and flipping the signs of fewer than 1% of important parameters [44]. The authors show that untargeted model poisoning attacks, as in [46], can dramatically degrade ICU mortality prediction models, with ROC AUC dropping from ~ 0.88 to as low as 0.75 when 50% of clients are malicious. The [41] study reveals that these types of sybil-based label-flipping attacks drive a global accuracy crash from $\sim 85\%$ to less than 30% in very heterogeneous environments. The latter is the classification accuracy, etc., of inference attacks [82] to measure the inferences the adversary has gained based on user logs. Reconstruction attacks utilize perceptual and pixel-level metrics (PSNR, LPIPS, SSIM & MSE) to assess the fidelity of leaked images [60]. Although model-agnostic, these metrics provide a powerful tool for identifying success in evaluating attacks; however, they are limited in assessing any clinical impact the attack might have on the population. The impact of such successful attacks can be severe in qualitative terms. Backdoor and evasion methods could cause clinically relevant misdiagnoses (e.g., changing predictions based on tumor subtype or ECG diagnosis [40,69]). Membership inference attacks could be considered invasive to patient privacy, as they show the contributions of sensitive medical data, such as patient histories and individual history measurements, by model training [65], the leakage of which might raise concerns about legal and regulatory consequences. Therefore, even if reconstruction attacks produced noisy or incorrect results, the leakage of basic anatomical structure or disease markers may still be considered a violation of privacy [60]. Qualitative assessments are needed to position numerical metrics in light of clinical risk. However, the computational scalability of attacks is limited, and the applicability of such attacks is therefore highly variable. Hence, lightweight poisoning techniques such as distance-based perturbations have low overhead and scale well on large datasets [40]. On the other hand, advanced types of attacks, including those that employ generative adversarial networks, evolutionary optimization or multi-stage shadow model training, are computationally expensive [43,82]. These differences highlight the sharp contrast in the attack landscape—while costly attacks could be more effective, low-cost attacks pose the most serious systemic threat because they can be deployed and easily scaled. Therefore, based on the literature, medical ML systems are considered high-efficiency attack vectors and can, on average, be implemented in practical applications. These dual characteristics highlight the need for scalable, domain-aware defenses that account for worst-case adversary performance and the circumstances of the healthcare domain.

4.2.3 Comparative Analysis of Defense Strategies and Attribute/Attack Resilience

The most crucial conclusion is that most of these adversarial attacks are constructed to bypass or can withstand common FL defenses. Table 11 shows that this analysis reveals both adversaries' adaptive behavior and the weaknesses of specific defenses.

The major trend is attack surfaces that are *robust to or even strengthened by* privacy protections. One classic example is the *DP-Poison* attack [43], where a set of DP noise hides malicious updates, enabling adversaries to evade state-of-the-art robust aggregation algorithms. Other types of attacks are also transferable to DP, even when DP is used, such as adversarial instances [69]. However, recent works discuss *adaptive adversaries* that develop tactics (e.g., optimizing noise levels or designing updates that align with the global model's direction) to remain stealthy and effective despite advanced defenses, while some papers point to attacks on unguarded systems. It underscores that static defense is unnecessary in the context of dynamic,

multi-layered, and adaptive security frameworks [67,73]. MultiMOORA-guided FL ranks clients for their five metrics: accuracy, precision, recall, F1, loss, and excludes low-scoring clients, meaning that no infected client can be aggregated [53]. Answering both availability (and integrity) poisoning without the user being aware of the number of attackers, FLDP-DM is based on both norm-based and accuracy-based detection strategies [52].

Table 11: Attack robustness against common FL defenses and adaptive capabilities.

Robustness to Defenses	Adaptive Adversaries Considered	Ref.
Evades RFA, Trimmed Mean, Multi-Krum, Bulyan, FoolsGold, RLR. Exploits DP noise as cover.	Yes: Optimizes noise to mimic benign DP-noised updates.	[43]
Effective against FedAvg, FedMD, FedProx protocols.	No: Attack methodology is static.	[82]
Partially mitigated by DP (reduces ASR but does not eliminate); attacks remain effective.	No: Fixed attack parameters.	[69]
Defense is local; compatible with secure aggregation (privacy-preserving).	No: Attack is a fixed one-shot poisoning strategy.	[42]
Designed to outperform static aggregation (FedAvg) and robust rules (Krum).	Implicitly: Uses residual testing for dynamic client weighting.	[47]
Does not rely on DP; defense via augmentation is orthogonal and additive to crypto defenses.	No: Assumes attacker does not know augmentation parameters.	[59]
Proposes a defense (SDP), not an attack. Attack (GIA) is the baseline threat.	N/A	[61]
Evades direction-based detection (AlignIns) by preserving overall gradient direction; only sign consensus can detect.	No: Attack is static (Neurotoxin).	[44]
Bypasses traditional robust aggregators (Krum, Median, Trimmed-Mean) by distributing malicious updates across multiple sybil identities.	Yes: Sybil clients coordinate to evade per-client outlier detection.	[41]
Evades distance-based and direction-based defenses by keeping updates within benign clusters.	No: Attacks are static (Opt-Fang, LIE, etc.).	[46]
Targets gradient inversion; CKKS encryption provides IND-CPA security, preventing reconstruction.	No: Attack is passive (honest-but-curious).	[62,79]
PGD adversarial attacks can degrade model accuracy; dynamic weighting (accuracy $\times$ samples) reduces impact.	No: Attack is static PGD.	[71]

Note: N/A denotes “not applicable.” In [61], the adaptive-adversary field is not applicable because the study proposed a defense and evaluates gradient inversion attack (GIA) as the baseline threat rather than presenting an adaptive attack.

4.2.4 Comparative Analysis and Methodological Trends

Based on the 45 reviewed studies, several methodological trends and comparative insights emerge:

Trend 1: From Single-Mechanism to Multi-Layer Defenses (2022–2024)

Early studies (2022–2023) largely investigated individual defense mechanisms in isolation. However, it was later found that single mechanisms are insufficient to counter adaptive attacks. For instance, DP alone was shown to be vulnerable to DP-Poison attacks [43] and robust aggregation alone did not succeed against sybil-based attacks [41]. Consequently, to prevent these attacks, a multi-layer defense mechanism was developed that combines DP, encryption, and anomaly detection [67].

Trend 2: Adaptive and Context-Aware Defenses (2024–2026)

There has been more recent work on defense that can respond to real-time threats and data sensitivity. Sensitivity-Aware DP [61] allocates privacy budgets based on the vulnerability of each sample, while MedGuard-FL [67] adjusts DP based on patient acuity scores. These approaches are a paradigm shift from static to dynamic privacy protection.

Trend 3: The Weaponization of Privacy Mechanisms

One issue that several studies have highlighted is the use of privacy mechanisms for adversarial purposes. DP-Poison [43] shows that DP noise can hide backdoor injections, and adversarial attacks [69] show that DP can be bypassed, yet the utility is degraded. It is important to keep in mind the unintended effects of privacy mechanisms when using them.

Comparative Insights: Defense Effectiveness by Attack Type

As the comparison shows, there is no single defense mechanism that is effective against all types of attacks, and there are various trade-off points between protection scope, computational cost, and practical use. As summarized in Table 12, Krum aggregation is well-suited against label-flipping attacks, but cannot cope with adaptive backdoor attacks and has major computational costs. DP is effective in defense against membership inference threats, but it can be used to mask poisoning attacks and often damages the model to clinically unacceptable levels for medical imaging. Homomorphic encryption (HE) and CKKS-based algorithms are good defense against gradient inversion attacks, but have significant computational costs of 30%–456%. Dual-dimensional detection schemes can be useful for backdoor attacks through direction and sign analysis, but they require access to plaintext gradient signs and struggle with combined attack vectors. Blockchain-based trust mechanisms are efficient against Sybil attacks but incur high latency and energy consumption, whereas hypernetwork-based detection methods are effective against untargeted poisoning, albeit with reduced interpretability. These findings underscore the need for adaptive multi-layer defense architectures that leverage complementary mechanisms to address the complex threat landscape of healthcare federated learning systems.

Table 12: Effectiveness of the defense types and their limitations.

Defense Type	Most Effective Against	Least Effective Against	Limitations
Krum	Label-flipping attacks	Adaptive backdoor attacks	Computationally expensive; assumes $\leq 50\%$ malicious clients
DP	Membership inference	Poisoning attacks (can be exploited)	Degrades utility; high privacy budgets impractical for medical imaging
HE/CKKS	Gradient inversion	Poisoning attacks	30%–456% computational overhead

(Continued)

Table 12 (continued)

Defense Type	Most Effective Against	Least Effective Against	Limitations
Dual-dimensional detection	Backdoor attacks	Combined attack vectors	Requires plaintext sign access
Blockchain-based trust	Sybil attacks	Adaptive poisoning	High latency; energy-intensive
Hypernetwork-based detection	Untargeted poisoning	Sybil attacks	Limited interpretability

4.3 Domain-Specific (Healthcare Context) Perspectives

This part situates the technical attacks in the context of the specifics of healthcare data and systems, as well as their regulations. Considering the clinical reality, patient harm risks, and regulatory impact of vulnerabilities identified in prior literature goes beyond algorithm effectiveness.

4.3.1 Analysis of Data Sensitivity and Granularity

Depending on what medical data is targeted, the granular detail of leaked information can signal a serious privacy breach. The attacks on clinical data types are analyzed in [Table 13](#), including whether a patient's data will be leaked.

Table 13: Sensitivity, granularity, and clinical harm potential of attacks.

Type of Data Targeted	Feature/Record-Level Impact	Clinical Harm Potential	Ref.
Chest X-rays (TB, Pneumonia)	Full image reconstruction or inference of disease status from model updates.	High: Reveals specific pathology (e.g., Tuberculosis), leading to stigma, insurance issues.	[63,66]
CT Scans, Brain Tumor MRI	Reconstruction of anatomical slices; misclassification of tumor type via adversarial attack.	Critical: Leaks full 3D anatomy; induced misdiagnosis can lead to fatal treatment errors.	[68,69]
ECG Signals (MIT-BIH)	Reconstruction of individual heartbeat waveforms; inference of arrhythmia presence.	High: Heart condition data is highly sensitive and directly linked to insurability and employability.	[55,62,78,79]
Motion sensor keystroke data	Inference of typed characters (passwords, clinical notes) from smartphone sensors.	High: Can leak credentials or unstructured PHI, violating confidentiality of doctor-patient communication.	[70]

(Continued)

Table 13 (continued)

Type of Data Targeted	Feature/Record-Level Impact	Clinical Harm Potential	Ref.
Retinal OCT Images	Model degradation leading to misdiagnosis of macular diseases (CNV, DME).	Critical: False negatives/positives for eye diseases can cause blindness or unnecessary invasive treatment.	[39]
Chest X-ray (COVID-19), Retinal Fundus	Reconstruction of diagnostic images revealing patient identity and health condition.	High: Combined leakage of identity (via facial bones in X-ray) and sensitive diagnosis (COVID-19).	[61]
Neuroimaging data (MRI scans) from brain aging and Alzheimer's disease studies	Record-level impact: Membership inference attacks reveal whether an individual patient's MRI scan was used in the model training set	High: Disclosure of participation in sensitive medical studies could lead to discrimination, insurance implications, or psychological harm.	[65]
ICU Vital Signs (MIMIC-III)	Inference of acute patient status (e.g., sepsis, shock) from model updates.	High: Real-time health crisis information is deeply private and could be exploited for discrimination.	[46,67]
Dermatology, Blood cell images	Misclassification of skin lesions or blood cell types via poisoning attacks.	High: False diagnosis of melanoma or leukemia leads to severe treatment delays.	[41]
CIFAR-10 (proxy for medical imaging)	Backdoor injection causing targeted misclassification (e.g., cat → dog).	Medium: Indirect risk; used as benchmark but not real clinical data.	[44]
Heart attack risk factors (BRFSS)	Inference of disease status	High (insurance/employment discrimination)	[80]
IoT network traffic (CICIoMT2024, N-BaIoT, CICIDS2017)	Intrusion detection bypass	Critical (patient safety in IoMT)	[81]
Chest X-ray (CheXpert, RSNA, NIH)	Full image reconstruction + misdiagnosis	Critical (false negatives for pneumonia/COVID)	[72]

The attacks predominantly affect high-fidelity *medical imaging data* (X-rays, CT, MRI, OCT), in which a leaked image constitutes a clear breach of confidentiality under regulations such as HIPAA. The damage is at the *full-record level*—complete images or waveforms are reconstructed. This has serious clinical

consequences: leaked images can be inherently identifiable (for instance, via X-ray facial structure [61]) and may reveal pathologies that increase the risk of discrimination, insurance denial, and social stigma. Time-series data (ECG, vital signs) and even meta-inference (membership) may also pose major threats when exposed, and some health conditions or participation in sensitive studies may serve as exposures.

4.3.2 Analysis of Realism in Attack Scenarios

The feasibility and impact of an attack depend heavily on the realism of the modeled federated learning scenario. Important assumptions about data distribution, system architecture, and adversary knowledge are critical to assess, as summarized in Table 14.

Table 14: Analysis of realism in attack scenarios: data, system, and adversary assumptions.

Data Heterogeneity (Non-IID) Considered	Federated System Model	Realism of Auxiliary Data Assumption	Ref.
Explicitly tested (Dirichlet $\alpha = 0.3$, Sharded).	Cross-silo (hospitals as clients).	N/A (Attack/defense paper, not reliant on auxiliary data).	[50]
Evaluated under different FL protocols (FedAvg, FedProx).	Cross-device/silo (simulated).	Unrealistic for healthcare: Assumes attacker has a large, representative public dataset (e.g., CIFAR) to train shadow models, which is rarely available for medical domains.	[82]
Yes, non-IID partition across clients.	Cross-silo (3 clients, hospitals).	More realistic: Attacker only needs white-box access to their own client's model and data, no external dataset required.	[69]
Not explicitly mentioned for attack simulation.	Cross-silo (clients as institutions).	N/A: Attack requires only local label manipulation.	[40]
Data evenly distributed (IID) for simplicity.	Cross-silo (hospitals as clients).	N/A (Defense paper).	[30]
Simulated up to 300 users (horizontal FL).	Cross-device (wearables, smartphones).	N/A: Defense paper using ring signatures, not reliant on auxiliary data.	[64]
Real-world MIMIC-III data stratified by patient acuity.	Cross-silo (3 federated nodes).	Realistic: Defense uses a server-held public validation set (NEWS2 scores), which is feasible for common risk scores.	[67]
Extreme non-IID (single-class, Gaussian, partial class).	Hierarchical FL (2, 5, 10 edge servers).	N/A: Defense does not require auxiliary data.	[41]

(Continued)

Table 14 (continued)

Data Heterogeneity (Non-IID) Considered	Federated System Model	Realism of Auxiliary Data Assumption	Ref.
Non-IID simulated via Dirichlet distribution.	Cross-silo (clients as hospitals).	N/A: Defense uses secret sharing and ZKPs, no auxiliary data.	[77]
Non-IID via patient-based partition (each client sees one or few patients).	Cross-silo (24 clients).	N/A: Attack/defense uses only local data.	[79]
Non-IID label-skew via Dirichlet prior.	Cross-silo (24 simulated hospitals).	N/A: Defense (HyperShield) uses unsupervised detection on embeddings.	[46]
IID only (random shuffling).	Cross-silo (4, 8, 12 hospitals).	N/A: Attack and defense use partitioned data.	[62,71]
Not explicitly tested (uses public anchor dataset for geometric screening)	Cross-silo (centralized MCPS)	Realistic: anchor dataset is publicly available and not tied to client data	[72]

Note: N/A denotes “not applicable.” In this table, N/A is used when the evaluated attack or defense does not require an external auxiliary dataset; therefore, no auxiliary-data realism assumption applies.

There is a clear gap; the adversary is not aware. Most popular inference attacks [82] assume that an attacker has access to an enormous, high-fidelity *auxiliary dataset* drawn from the same distribution as the target private dataset. This is an extremely powerful assumption—one that is often unrealistic, especially in environments such as healthcare, where available medical datasets are sparse, proprietary, and heterogeneous. Other credible attacks only leverage *local model access* or exploit *local data* [40,69]. In addition, some *cross-silo* settings (limited number of hospitals) [50,68] models do not completely represent real-world, strongly *non-IID* data (e.g., pediatric hospital vs. an oncology center) scenarios, indicating underestimating or overestimating the transferability of the attack and the robustness of the defense, respectively.

4.4 Unresolved Research Challenges

Despite the significant advances documented in this review, several critical research challenges remain unresolved:

Challenge 1: Adaptive Multi-Vector Attacks

Most current defenses are evaluated against single attack vectors in isolation. But in the real world, adversaries are increasingly marrying multiple attack strategies, e.g., model poisoning to enable gradient inversion [12]. The interaction between different attack families is not well understood, and defenses designed to protect against one threat might inadvertently amplify others.

Challenge 2: Clinical Realism in Evaluation

Most studies have evaluated attacks and defenses on controlled datasets with IID distributions that are not representative of the heterogeneity of real-world healthcare data. In addition, evaluation metrics

(PSNR, accuracy, AUC) do not capture clinically relevant outcomes. Future work needs to include clinician-in-the-loop assessment and diagnostic metrics such as sensitivity, specificity, and calibration for clinically relevant endpoints.

Challenge 3: Underrepresented Data Modalities

The current literature is heavily biased toward medical imaging, with little or no attention to genomics, structured EHR, longitudinal sensor data, and streaming physiological signals. Each of these data types has its own attack surface and privacy requirements, necessitating specialized defenses.

Challenge 4: Practical Deployment Constraints

Many defense mechanisms impose prohibitive computational overhead, particularly for edge and IoMT devices. For example, homomorphic encryption can cause 30–456% overhead depending on the FL configuration [62] while GRASP-FL requires 67.7 s per round [45]. Finding the right balance between security and practical deployment has been challenging.

Challenge 5: Regulatory and Compliance Frameworks

Current regulatory frameworks do not explicitly address FL-specific vulnerabilities. No clear guidance as to whether model updates will constitute de-identified data, what is considered an “acceptable privacy risk,” or what sort of liability should be distributed in federated systems. These regulatory gaps must be addressed to enable widespread clinical deployment in the future.

5 Answers to the Research Questions

This section provides integrated answers to the research questions based on the literature review of privacy risks associated with federated learning (FL) for health data.

5.1 RQ1: Robust Design of Security In Place for Adaptive Attacks

Federated systems of health need to be based on an adaptive multi-layer defense architecture rather than on a separate privacy-preserving mechanism. From the literature, it is evident that even previous purely standalone defenses, including but not limited to DP and robust aggregation, can be defeated by highly sophisticated adaptive attacks. DP-Poison, e.g., injects DP noise to hide backdoor injections, revealing the insufficiency of single-mechanism defenses [43]. Therefore, effective frameworks integrate statistical privacy into DP, computational privacy into Homomorphic Encryption/Secure Multi-Party Computation, and integrity guarantees into anomaly detection, while the workflows proposed in papers published in EPP-BCFL [51] and PPFL [73] can facilitate this smoothly. Recent work goes one step further by demonstrating the value of bringing together complementary detection dimensions: DABD [44] fuses direction alignment (global) and sign consensus (local) to detect stealthy backdoor attacks that single-dimensional tools fail to detect. HyperShield [46] employs a hypernetwork to compute client embeddings and applies an unsupervised poisoning-detection engine that simultaneously uses temporal cosine similarity and spatial DBSCAN clustering. BlockAuthFL [41] combines blockchain-oriented authentication frameworks with dynamic trust weighting and rollback measures to thwart sybil-based label-flipping attacks. Technical mitigations in this context might be enhanced by applying dynamic adaptation measures that adjust the level of protection based on real-time threat signals or data sensitivity, as demonstrated in Dynamic privacy budgeting for MedGuard-FL [67] and Sensitivity-Aware DP [61]. Concurrently, utility for clinical applications is key, and custom-fitted FL systems (FedKC [50]) and the interpretable filtering of clients (RAB2-DEF [49]) have demonstrated the capability of systems to maintain diagnostic performance and mitigate adversarial threats. The conclusion: a real defense is not a set of isolated actions, but rather a sum of actions aligned to respond to mutable, multi-vector attacks of dynamic complexity.

5.2 RQ2: Context-Aware Defense Strategies

Healthcare FL should be optimally protected through ongoing risk assessment and fine-grained privacy budgeting in the clinical context and on the data set available. Sensitivity-based approaches such as Sensitivity-Aware DP [61] allocate privacy budgets based on the empirical weaknesses of particular samples, thereby providing stronger protection against data most susceptible to inversion attacks. The clinical context can guide adaptive privacy decisions: MedGuard-FL [67] modifies privacy choices based on urgency (i.e., NEWS2 scores) and acknowledges that emergencies in severe scenarios may require an immediate reduction in protection to take actions that could save lives. Hardware-enabled systems, such as FPGA-based risk quantification systems, demonstrate how dynamic, real-time adjustments can be implemented at no additional cost [68]. Shadow-model-based approaches, such as those proposed in [60], utilize simulated attacks to calibrate approaches and to better quantify how vulnerable systems are to realistic threats that can be leveraged to enhance their defenses. Collectively, these strategies transition from static privacy guarantees to context-based protections that consider privacy, utility, and clinical need.

5.3 RQ3: Methods of Clinical Evaluation

Future research requires a more integrated evaluation paradigm. Current clinical security FL evaluation approaches are largely insufficient because the algorithms rely on computational metrics and do not account for clinical significance. Quantitative techniques, e.g., PSNR, MSE, and accuracy, are useful but do not indicate whether reconstructed images display recognizable anatomy or whether adversarial perturbations are feasible to guide clinical decisions. A more integrated evaluation paradigm is needed to deliver clinician-in-the-loop assessments of the diagnostic benefit of leaked images or adversarial examples. In addition, this approach should be accompanied by a contextual risk assessment, so that attacks are evaluated not only for their technical success but also for the extent to which they can harm patients (e.g., misclassifying malignant tumors as benign). Realistic threat modeling is also required, as threats must be assessed with non-IID data distributions, scarce auxiliary information, and heterogeneous client configurations; this has been addressed in prior adversarial work [69]. Last but not least, the industry would also benefit from metrics that measure privacy leakage, degradation of clinical utility, and total system overhead. Although some adversarial attacks achieve very high success rates under controlled experimental conditions, their clinical plausibility, transferability, and detectability remain important considerations.

5.4 RQ4: FL Optimization for Healthcare Data Challenge

In response to the difficulties (e.g., non-IID distributions, class imbalance, multimodality, and strict privacy requirements) faced by healthcare federated learning across several settings, federated architecture co-design is needed. Individual FL schemes, such as FedKC [50], and federated transfer learning [75] perform better under heterogeneous data condition and are far more robust to poisoning. The imbalance can be alleviated by applying weighted aggregation and adaptive client selection techniques [57] to prevent data centralization by preserving fairness. Multimodal privacy-preserving FL frameworks have demonstrated the feasibility of integrating heterogeneous clinical modalities using complementary privacy mechanisms. PRI-FLEX [76] combines ECG signals from PTB-XL with clinical text from MIMIC-IV and evaluates differential privacy, secure aggregation, and their hybrid combination (DP+SA), whereas the secure multimodal FTL framework [79] integrates ECG signals and images using CKKS homomorphic encryption and differential privacy. Efficiency remains a major challenge, especially for medical devices with limited computational resources, so lightweight privacy policies [64] and hardware-accelerated FL [68] have also been considered as possible options. Furthermore, RAB2-DEF [49] shows that clients with poorer-quality data may also be retained to improve fairness. On edge devices, new defenses like HyperShield [46] demonstrate that

hypernetwork-based personalization offers strong poisoning resistance with very low overhead (1.4× CPU, 0.7× RAM) and is therefore appropriate for real ICU deployments. BlockAuthFL [41] shows that hierarchical FL with blockchain recovery can restore model accuracy from about 25% to greater than 80% under severe heterogeneity and sybil attacks. Overall, healthcare-based FL needs to shift from a classical FedAvg approach to personalized algorithms that jointly maximize privacy, robustness, and clinical utility, as outlined by these findings.

6 Future Directions

We conduct a comprehensive study of attacks targeting healthcare FL systems, demonstrating technical limitations and fundamental inadequacies in existing trust models and risk assessment methods. In the case of the work presented here, these gaps have been further exposed recently: ref. [44] found that static direction-based trust breaks down against sign-flipping backdoor attacks, and ref. [41] demonstrated that identity-based trust alone cannot protect against sybil collusion. Further, ref. [77] showed that even honest-but-curious aggregators can collaboratively reconstruct client updates once secret sharing thresholds are surpassed. These research results suggest that trust models which are built on binary assumptions must be developed urgently. In this section we synthesize such knowledge to provide guidance on future security frameworks, on practical development, policy-making, and further developments.

6.1 Transition and Risk Assessment: The Trust Model

There has been much discussion in the literature against prevailing binary trust models (e.g., “server is honest-but-curious,” “clients are benign”) that underpin many FL architectures. Future research should engage with trust and risk from a more nuanced, context-specific, and active perspective. Within an Adaptive Multi-Layer Trust (AMLT) framework, each participant (server, client, or individual data point) has their respective dynamic risk profile. The security parameters (encryption level, noise magnitude, aggregation rule) are also modulated and tuned according to this profile and clinical context (emergency monitoring vs routine monitoring). This transcends traditional protection from isolated attacks toward the implementation of an evolving risk landscape. The future direction of active trust and risk modeling is summarized in Table 15.

Table 15: Evolution from static to dynamic trust and risk models in healthcare FL.

Current Paradigm (Challenged)	Key Limitation	Proposed Future Direction & Synthesis Perspective
Static, Binary Trust: Server is either fully trusted or semi-honest; clients are either honest or malicious.	Attackers take advantage in these gaps: a semi-honest server can reconstruct [59]; malicious clients can impersonate benign behavior by using DP as its mask [43]. Sign-flipping attacks avoid direction-influence trust [44]. Sybil clients with valid identities bypass authentication [41]. Colluding aggregators violate non-collusion assumptions [77].	Refocus to Continuous, Behavioral Trust Scoring: Trust must be a continuous metric, updating per client per round as per behavior (update consistency, contribution quality [57]), provenance and hardware attestation. That said, frameworks such as [67] that dynamically scale security based on client risk are a start. Dual-dimensional trust (direction + sign [44]) and hypernetwork-based behavioral embeddings [46] appear promising directions.

(Continued)

Table 15 (continued)

Current Paradigm (Challenged)	Key Limitation	Proposed Future Direction & Synthesis Perspective
Isolated Risk Assessment: Privacy risk is evaluated separately from model performance and clinical safety.	Attacks like <i>DP-Poison</i> [43] exploit the privacy-robustness trade-off, while adversarial attacks [69] degrade utility but do not violate privacy. The utilities suffer with no loss of respect for privacy in this trade-off. Poisoning attacks can significantly affect ICU outcome prediction while maintaining privacy metrics [46].	Integrated Clinical Risk Scoring: Data from risk models shall combine <i>privacy leakage probability</i> (based on sensitivity analysis [61]), <i>integrity threat level</i> (from anomaly detection [47]), and <i>clinical impact severity</i> (patient acuity [67]). A low-privacy-risk update that seriously degrades diagnostic performance in critical cases represents a high composite risk.
One-Size-Fits-All Security: Using one-size-fits-all encryption (fixed DP ϵ) as the defense parameters: Uniform security settings for all data.	This either overprotects low-sensitivity data (harming utility) or underprotects high-sensitivity data (e.g., rare disease images) [61]. Homomorphic encryption overhead varies widely across FL scenarios (30% to 456%) [62], rendering uniform encryption inefficient.	Adaptive, Granular, Data-Focused, Risk Mechanisms: On the sample level or feature level, security needs to adapt. In the future, systems should self-calibrate protection based on <i>inferred data sensitivity</i> [61] from attack simulations, <i>regulatory class</i> (e.g., genetic data vs. vitals), and <i>processing context</i> (research vs. real-time diagnosis). Multi-modal FL frameworks (see, for example, [79]) demonstrate how different modalities may need different protection levels.
Post-Hoc Defense: Security mechanisms are added onto existing FL pipelines.	This leads to complex systems that cannot deal with unexpected user input (e.g., DP poisoning attacks). Such overhead in defense can be an insurmountable challenge to IoMT [63]. Furthermore, those defenses that need auxiliary validation data violate privacy principles [44].	Co-Designed, Adjustable Security Primitives: The security (privacy, integrity) of future FL frameworks should be treated as a first-class and tunable parameter. Hardware-software co-design, such as in [68], paves the way. Systems must permit the running change of a “security dial” (ranging from “max efficiency” to “max privacy”) in line with the dynamic risk score. Lightweight defenses such as client dynamic weighting [71] and hypernetwork-based detection [46] affirm that security can be incorporated without excessive overhead.

Illustrated in Fig. 5, the flowchart organises defense into four key dimensions, namely computational privacy (homomorphic encryption, secure multi-party computation, hardware-accelerated encryption), statistical privacy (differential privacy, hybrid DP, sensitivity-aware DP), robust aggregation (Krum, trimmed mean, median, blockchain verification, adaptive weighting) and anomaly detection (KL divergence auditing, shadow model detection, interpretability-based filtering). The drawbacks of each pillar are detailed: high computational cost for encryption; DP noise can be exploited (e.g., DP-Poison); robust aggregation can be circumvented through adaptive poisoning; anomaly detection often requires public validation sets or shadow models. The diagram therefore concludes that no single mechanism fits all cases, prompting the need to design adaptive, multi-layer defense frameworks—a key message of the paper.

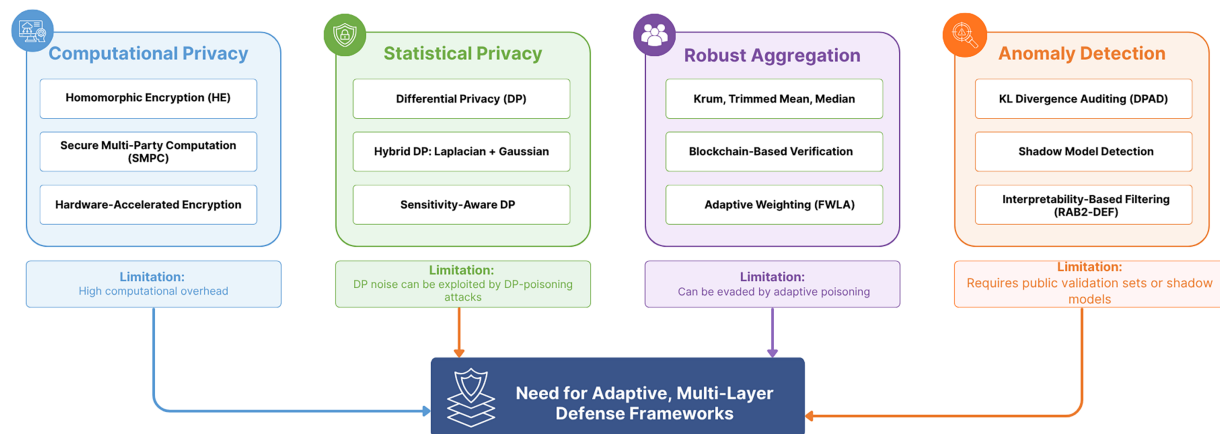


Figure 5: Defense mechanisms, limitations and adaptive countermeasures.

Accordingly, in this paper, a dynamic trust model is offered in Fig. 6 to replace static binary assumptions (honest and malicious). There are three linked stages of it; each one is connected to the other. Continuous risk identification involves client behavior monitoring (update consistency, contribution quality), data sensitivity scoring (per-sample attack simulation), and clinical context (e.g., patient NEWS2 score, emergency flag). Dynamic trust scores generate a client trust score (ongoing, per-round), a data-point risk profile (to inform privacy budget distribution), and a composite risk index that incorporates the probability of privacy leakage, the integrity threat level, and the severity of clinical impact. Adaptive security primitives such as privacy dial (DP epsilon or encryption level adjustment), robust aggregation (weighted by trust scores) and audit/explainability mechanisms. A feedback loop from security primitives to risk assessment enables real-time adaptation. This figure can add credence to the future directions (Section 6) that support context-driven, evolving trust models. Table 16 provides recommendations for practices, standards, and research directions.

Step forward will require a simultaneous and integrated, multidisciplinary response of technological innovation, regulatory evolution, and clinical application. It's time to design and implement federated learning systems with a security-first mindset; ensure privacy and robustness as requirements—not options. Policymakers must revise policy frameworks and implement guidelines to safeguard institutions that deploy verifiably secure FL systems in the rapidly evolving era of distributed AI, with clear guidance and safe harbors. Researchers must bridge theoretical guarantees and real-world clinical constraints to build practical systems that are transparent and trustworthy, and to build consensus among patients, clinicians, and regulators. Last but not least, not just to minimize isolated attack vectors, but also to engineer federated learning ecosystems that actually work, that are solid and accountable and that are responsible for protecting sensitive types of medical data.

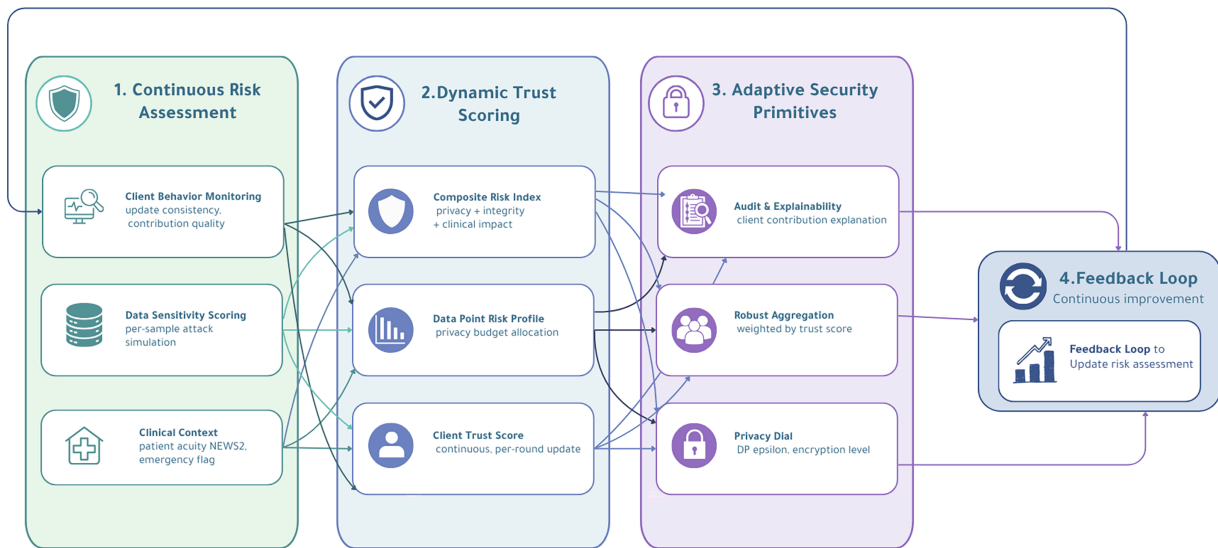


Figure 6: Adaptive multi-layer trust model for healthcare federated learning.

Table 16: Recommendations for practice, policy, and research agendas.

Domain	Key Implications & Recommendations	Specific Future Research Avenues
Practice (System Developers & Hospitals)	- Adopt Defense-in-Depth: No single mechanism (DP, HE) is sufficient. Deploy layered defenses (e.g., DP + secure aggregation + anomaly detection [73]). Dual-dimensional detection (direction + sign) [44] and unsupervised embedding clustering [46] should complement cryptographic protections. - Implement Continuous Auditing: FL pipelines must log and monitor update patterns for anomalies [49], enabling detection of sophisticated, adaptive attacks. Blockchain-based immutable logs [41] can provide tamper-evident audit trails. - Conduct Pre-Deployment Red-Teaming: Hospitals must simulate realistic attack scenarios (including adaptive ones) on their specific data distributions before clinical deployment. Testing should include gradient inversion [62], poisoning [46], and sybil attacks [41].	- Develop lightweight, composable defense toolkits suitable for edge/IoMT devices (e.g., hypernetwork-based detectors [46], dynamic weighting [71]). - Create standardized FL security audit protocols and benchmarks for healthcare, including realistic non-IID partitions and multi-attack scenarios. - Research secure and fair incentive mechanisms to encourage participation while disincentivizing malicious behavior [55]. Blockchain-based reputation systems [41] are a promising direction.

(Continued)

Table 16 (continued)

Domain	Key Implications & Recommendations	Specific Future Research Avenues
Policy (Regulators & Standards Bodies)	- Update Compliance Guidance: Regulations (HIPAA, GDPR, EU AI Act) must explicitly address FL. Guidance should state that <i>model updates are not de-identified data</i> and require evidence of protection against reconstruction and inference attacks. Homomorphic encryption [62,79] and differential privacy [78] should be recognized as acceptable safeguards. - Mandate Transparency & Explainability: Policies should require that FL operators can explain client contributions and detect unfair bias, as in [49], to meet AI Act requirements. Frameworks like PRIFLEX [76] that measure privacy-aware explainability set a benchmark. - Foster Secure Data Ecosystems: Encourage the development of trusted, hardware-backed federations and secure data markets with built-in privacy guarantees. Blockchain-enabled FL [41] can provide decentralized trust anchors.	- Develop quantitative metrics for “acceptable privacy risk” in FL, moving beyond qualitative principles. Sensitivity-aware DP [61] offers a model for empirical risk quantification. - Study legal frameworks for liability distribution in federated systems when a model failure or breach occurs. This includes attribution of poisoning attacks [77] and accountability for malicious aggregators. - Create certification standards for “Healthcare-Grade FL” platforms that mandate specific defenses (e.g., secret sharing [77], sign consensus [44], authenticated aggregation [78]).
Research (Academia & Industry Labs)	- Prioritize Realism & Clinical Relevance: Move beyond IID/CIFAR-10 to complex, non-IID, imbalanced medical data (e.g., MIMIC-III, PTB-XL, Kvasir [41]). Integrate clinician-in-the-loop evaluation to assess the meaningfulness of attack/defense. - Embrace Holistic Evaluation: Papers must report the full triad: privacy/security efficacy, model utility (accuracy, AUC), and system overhead (time, energy, communication). For example, [62] reports encryption overhead (30%–456%), [46] reports CPU/RAM on edge hardware, and [79] reports both accuracy drop (0.24%) and training time increase (36×). - Focus on Underrepresented Modalities: Shift significant focus from imaging to genomics, EHRs, and longitudinal sensor data. Multi-modal frameworks [79] and ICU time-series [46] are steps in the right direction.	- Investigate foundational models and federated fine-tuning in healthcare, with associated new attack surfaces. Transfer learning with frozen layers [79] can reduce cryptographic overhead but may introduce new vulnerabilities. - Explore post-quantum cryptography and homomorphic encryption for long-term security in FL. CKKS-based schemes [62,79] already provide post-quantum guarantees under RLWE. - Develop theories of convergence and robustness for FL under the AMLT framework with adaptive, personalized defenses. Formal convergence proofs for encrypted FL [79] and Byzantine-robust aggregation [44] are emerging but need extension to multi-attack scenarios. - Design privacy-preserving sign consensus mechanisms [44] compatible with secure aggregation (e.g., using threshold secret sharing [77]) to avoid exposing plaintext gradient signs.

The mind map in Fig. 7 summarizes the major gaps and suggests future research following the systematic review. The first branch deals with realism & clinical validity, which calls for non-IID medical data, clinician-in-the-loop evaluation, and diagnostic-meaningful metrics beyond PSNR/accuracy. Genomic data, structured EHR, and longitudinal sensor data are underrepresented modalities, moving beyond the

current bias toward medical imaging. Adaptive & multi-vector threats require research on attack-aware adversaries, combined attacks (poisoning + inference + evasion), and defenses tested against compound threats. Lightweight & practical defenses focus on feasibility for IoMT edge devices, hardware-software co-design, and tunable security-efficiency trade-offs. Regulatory & compliance frameworks need explicit guidance on FL model updates, certification standards for “Healthcare-Grade FL”, and liability distribution. Finally, holistic evaluation metrics should standardize privacy-utility-efficiency reporting and promote reproducibility and code sharing.

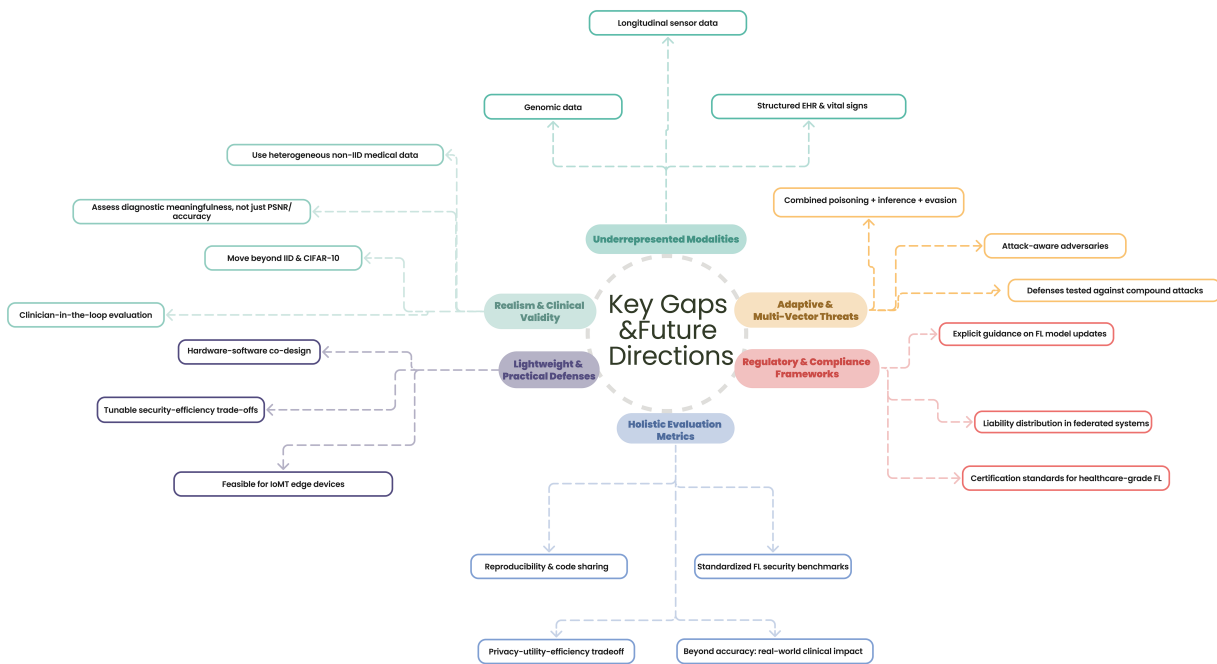


Figure 7: A mindmap of research gaps and future directions.

7 Conclusion

Our systematic review synthesized 45 peer-reviewed papers on privacy and security threats in healthcare FL. The evidence shows that decentralized training does not, in itself, make patient information private: gradients and model updates may enable reconstruction and membership inference; poisoning, backdoors, adversarial examples, sybil behavior, and other approaches are known to threaten confidentiality and clinical integrity. The central cross-study finding is not that one defense works best; rather, each family offers a different layer and fails when used outside that layer. Differential privacy limits disclosure but can reduce utility and obscure small malicious updates; cryptographic aggregation hides update contents but does not establish their trustworthiness; robust aggregation struggles to distinguish adaptive attacks from legitimate clinical heterogeneity; and blockchain adds provenance without demonstrating that an update is benign.

The study-level risk-of-bias assessment identified greater uncertainty in outcome measurement and study conduct/data-flow reporting than in threat-model specification and implementation fidelity. It was not appropriate to conduct formal publication bias tests because the included studies did not provide sufficiently comparable effect measures and sampling variances. The fact that good findings are now very much the dominant outcome suggests that publication bias is likely and selective. Nevertheless, the main qualitative findings remained unchanged across sensitivity analyses that excluded studies with a high risk of bias and restricted the synthesis to studies with a low-risk threat-model specification or low-risk outcome measurement.

Healthcare FL should be designed as a continuously assessing security system rather than a privacy-preserving communication pattern. A reliable deployment should be based on cryptographic confidentiality, clear privacy accounting, authenticated participation, robust and subgroup-aware anomaly detection, recovery and audit, and context-sensitive trust updates. Future studies should test compound adaptive adversaries on non-IID clinical data, report privacy and utility efficiency results together, present clinically relevant and subgroup-specific consequences, and provide sufficient implementation details to enable independent reproduction. These steps are necessary for FL to support collaborative clinical intelligence while still providing credible protection for patients and institutions.

Acknowledgement: The authors extend their appreciation to Umm Al-Qura University, Saudi Arabia, for funding this research work through grant number: 26UQU44680053GSSR02

Funding Statement: This research work was funded by Umm Al-Qura University, Saudi Arabia under grant number: 26UQU44680053GSSR02.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Nada Hasan Alfahmi, Manal AlGhamdi, Muhammad Arif; methodology, Muhammad Arif, Nada Hasan Alfahmi; software, Nada Hasan Alfahmi; validation, Nada Hasan Alfahmi, Manal AlGhamdi, Muhammad Arif; formal analysis, Muhammad Arif; investigation, Nada Hasan Alfahmi, Manal AlGhamdi, Muhammad Arif; resources, Nada Hasan Alfahmi; data curation, Nada Hasan Alfahmi, Muhammad Arif; writing—original draft preparation, Nada Hasan Alfahmi; writing—review and editing, Manal AlGhamdi, Muhammad Arif; visualization, Nada Hasan Alfahmi; supervision, Manal AlGhamdi, Muhammad Arif; project administration, Nada Hasan Alfahmi; funding acquisition, Nada Hasan Alfahmi. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The database-specific search strategies, screening rules, study-level risk-of-bias judgments, reporting-bias assessment, and sensitivity analyses are provided in the Supplementary Material.

Ethics Approval: Not applicable. This is a literature review paper.

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

Supplementary Materials: The supplementary material is available online at <https://www.techscience.com/doi/10.32604/cmc.2026.086394/s1>.

References

1. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Decis Mak.* 2021;21(1):125. doi:10.1186/s12911-021-01488-9.
2. Pasham SD. The function of artificial intelligence in healthcare: a systematic literature review. *Int J Acta Inform.* 2023;2(1):32–42.
3. Seh AH, Zarour M, Alenezi M, Sarkar AK, Agrawal A, Kumar R, et al. Healthcare data breaches: insights and implications. *Healthcare.* 2020;8(2):133. doi:10.3390/healthcare8020133.
4. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: *Artificial intelligence and statistics*. New York, NY, USA: PMLR; 2017. p. 1273–82.
5. Zhu L, Liu Z, Han S. Deep leakage from gradients. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; 2019 Dec 8–14; Vancouver, BC, Canada. p. 14774–84.
6. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box Inference attacks against Centralized and federated learning. In: *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*; 2019 May 19–23; San Francisco, CA, USA. New York, NY, USA: IEEE; 2019. p. 739–53.
7. Zhang J, Zhu H, Wang F, Zhao J, Xu Q, Li H. Security and privacy threats to federated learning: issues, methods, and challenges. *Secur Commun Netw.* 2022;2022(1):2886795. doi:10.1109/nana53684.2021.00062.

8. Yin X, Zhu Y, Hu J. A comprehensive survey of privacy-preserving federated learning: a Taxonomy, review, and future directions. *ACM Comput Surv.* 2022;54(6):1–36. doi:10.1145/3460427.
9. Jobic P, Mayoue A, Tucci-Piergiovanni S, Terrier F. Extending the scope of gradient reconstruction attacks in federated averaging. In: *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*; 2024 Jun 24–26; Baiona, Spain. New York, NY, USA: Association for Computing Machinery; 2024. p. 235–46. doi:10.1145/3658664.3659636.
10. Rao B, Zhang J, Wu D, Zhu C, Sun X, Chen B. Privacy Inference attack and defense in Centralized and federated learning: a comprehensive survey. *IEEE Trans Artif Intell.* 2025;6(2):333–53. doi:10.1109/tai.2024.3363670.
11. Valadi V, Åkesson M, Östman J, Toor S, Hellander A. Practical feasibility of gradient inversion attacks in federated learning. *arXiv:2508.19819.* 2025. doi:10.48550/arxiv.2508.19819.
12. Zhao J, Bagchi S, Avestimehr S, Chan K, Chaterji S, Dimitriadis D, et al. The federation strikes back: a survey of federated learning privacy attacks, defenses, applications, and policy landscape. *ACM Comput Surv.* 2025;57(9):1–37. doi:10.1145/3724113.
13. Wei K, Qian J, Jia C, Chen J, Zhang J, Wu Y, et al. Medical image privacy in federated learning: segmentation-reorganization and sparsified gradient matching attacks. *IEEE J Biomed Health Inform.* 2026;30(2):1443–51. doi:10.1109/jbhi.2025.3593631.
14. Kumar A, Purohit V, Bharti V, Singh R, Singh SK. MediSecFed: private and secure medical image classification in the presence of Malicious clients. *IEEE Trans Ind Inform.* 2022;18(8):5648–57. doi:10.1109/tii.2021.3138919.
15. Hallaji E, Razavi-Far R, Saif M. Federated and transfer learning: a survey on adversaries and defense mechanisms. In: *Federated and transfer learning*. Berlin/Heidelberg, Germany: Springer; 2022. p. 29–55.
16. Li Y, Guo Z, Yang N, Chen H, Yuan D, Ding W. Threats and defenses in the federated learning life cycle: a comprehensive survey and challenges. *IEEE Trans Neural Netw Learn Syst.* 2025;36(9):15643–63. doi:10.1109/TNNLS.2025.3563537.
17. Aouedi O, Sacco A, Piamrat K, Marchetto G. Handling privacy-sensitive medical data with federated learning: challenges and future directions. *IEEE J Biomed Health Inform.* 2023;27(2):790–803. doi:10.1109/jbhi.2022.3185673.
18. Oyekan B. Developing privacy-preserving federated learning models for collaborative health data analysis across multiple institutions without compromising data security. *J Knowl Learn Sci Technol.* 2024;3(3):139–64. doi:10.60087/jklst.vol3.n3.p139-164.
19. Ali MS, Ahsan MM, Tasnim L, Afrin S, Biswas K, Hossain MM, et al. Federated learning in healthcare: model misconducts, security, challenges, applications, and future research directions—a systematic review. *arXiv:2405.13832.* 2024.
20. Parampottupadam S, Floca R, Bounias D, Hamm B, Roy S, Sav S, et al. Client security alone fails in federated learning: 2D and 3D attack insights. In: *Meets africa workshop*. Berlin/Heidelberg, Germany: Springer; 2024. p. 235–44.
21. Das BC, Amini MH, Wu Y. Privacy risks analysis and Mitigation in federated learning for medical images. In: *Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; 2023 Dec 5–8; Istanbul, Turkiye. New York, NY, USA: IEEE; 2023. p. 1870–3.
22. Wu L, Liu Z, Pu B, Wei K, Cao H, Yao S. DGGI: deep generative gradient Inversion with diffusion model. *Inf Fusion.* 2025;113(7):102620. doi:10.1016/j.inffus.2024.102620.
23. Shi S, Haque MS, Parida A, Linguraru MG, Hou YT, Anwar SM, et al. Harvesting private medical images in federated learning systems with crafted models. *arXiv:2407.09972.* 2024.
24. Bai L, Hu H, Ye Q, Li H, Wang L, Xu J. Membership Inference attacks and defenses in federated learning: a survey. *ACM Comput Surv.* 2025;57(4):1–35. doi:10.1145/3704633.
25. Wang Z, Huang Y, Song M, Wu L, Xue F, Ren K. Poisoning-assisted property Inference attack against federated learning. *IEEE Trans Dependable Secur Comput.* 2023;20(4):3328–40. doi:10.1109/tdsc.2022.3196646.
26. Kim H, Cho Y, Lee Y, Bae H, Paek Y. Exploring clustered federated learning’s vulnerability against property Inference attack. In: *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*; 2023 Oct 16–18; Hong Kong China. p. 236–49.
27. Xia G, Chen J, Yu C, Ma J. Poisoning attacks in federated learning: a survey. *IEEE Access.* 2023;11:10708–22.

28. Kumar KN, Mohan CK, Cenkeramaddi LR, Awasthi N. Minimal data poisoning attack in federated learning for medical image classification: an attacker perspective. *Artif Intell Med.* 2025;159(1):103024. doi:10.1016/j.artmed.2024.103024.
29. Stelter L, Corbetta V, Beets-Tan R, Silva W. Assessing the impact of federated learning and differential privacy on multi-centre Polyp segmentation. In: *Proceedings of the 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2024 Jul 15–19; Orlando, FL, USA.* New York, NY, USA: IEEE; 2024. p. 1–4.
30. Kalapaaking AP, Khalil I, Yi X. Blockchain-based federated learning with SMPC model verification against poisoning attack for healthcare systems. *IEEE Trans Emerg Top Comput.* 2024;12(1):269–80. doi:10.55041/ijcope.v2i4.327.
31. Yesmin F. MedHE: communication-efficient privacy-preserving federated learning with adaptive gradient sparsification for healthcare. arXiv:2511.09043. 2025. doi:10.48550/arxiv.2511.09043.
32. Pfitzner B, Maurer MM, Winter A, Riepe C, Sauer IM, Water R, et al. Differentially-private federated learning with non-IID data for surgical risk prediction. *Int J Semant Comput.* 2025;19(03):343–67. doi:10.1142/sl793351x25430019.
33. Ali A, Snášel V, Platoš J. Health-FedNet: a privacy-preserving federated learning framework for scalable and secure healthcare analytics. *Results Eng.* 2025;27:106484.
34. Pati S, Kumar S, Varma A, Edwards B, Lu C, Qu L, et al. Privacy preservation for federated learning in health care. *Patterns.* 2024;5(7):100974. doi:10.1016/j.patter.2024.100974.
35. Madathil NT, Dankar FK, Gergely M, Belkacem AN, Alrabaaee S. Revolutionizing healthcare data analytics with federated learning: a comprehensive survey of applications, systems, and future directions. *Comput Struct Biotechnol J.* 2025;28:217–38.
36. Kumar KS, Nelson L, Jibinsingh BR. Systematic review of privacy-preserving federated learning in decentralized healthcare systems. *Frankl Open.* 2025;13(3):100440. doi:10.1016/j.fraope.2025.100440.
37. Neto HNC, Hribar J, Dusparic I, Mattos DMF, Fernandes NC. A survey on securing federated learning: analysis of applications, attacks, challenges, and trends. *IEEE Access.* 2023;11:41928–53. doi:10.1109/access.2023.3269980.
38. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71. doi:10.1016/j.jclinepi.2021.03.001.
39. Anunita KP, Devisri SR, Arumugam C. Federated learning: countering label flipping attacks in retinal OCT. *Procedia Comput Sci.* 2025;254(12):58–67. doi:10.1016/j.procs.2025.02.064.
40. Zahir-Ismail AT, Shukla R. Analyzing the vulnerabilities in split federated learning: assessing the robustness against data poisoning attacks. *Sci Rep.* 2025;15(1):30468. doi:10.1038/s41598-025-15993-8.
41. Khedekar D, Pokarna H, Sahay A, Richhariya B, Mahapatra T, Rajput AS. BlockAuthFL: a robust blockchain-enabled framework for addressing data heterogeneity and sybil attacks in healthcare federated learning. *Smart Health.* 2026;40(4):100649. doi:10.1016/j.smhl.2026.100649.
42. Chen CL, Babakniya S, Paolieri M, Golubchik L. Defending against poisoning backdoor attacks on federated meta-learning. *ACM Trans Intell Syst Technol.* 2022;13(5):1–25. doi:10.1145/3523062.
43. Zheng H, Chen J, Liu T, Cheng Y, Wang Z, Wang Y, et al. DP-poison: poisoning federated learning under the cover of differential privacy. *ACM Trans Priv Secur.* 2025;28(1):1–28. doi:10.1145/3702325.
44. Kang J, Du R, Zhang X. DABD: direction alignment-based backdoor defense with sign consensus in federated learning. *Knowl-Based Syst.* 2026;341(3):115727. doi:10.1016/j.knosys.2026.115727.
45. Soliman A, Amin KM, Semary NAEM, Mousa H. GRASP-FL: a closed-loop defence framework against adaptive backdoor attacks in federated learning. *IEEE Access.* 2026;14:38816–38.
46. Dao BT, Ma VD, Vu TA, Nguyen HT, Tran KP, Truong TH. HyperShield: a personalized hypernetwork-based defense against poisoning attacks in federated learning for healthcare. *Expert Syst Appl.* 2026;299(5):130237. doi:10.1016/j.eswa.2025.130237.
47. Ning Y, Zhang Z, Li H, Xia Y, Li M. A federated weighted learning algorithm against poisoning attacks. *Int J Comput Intell Syst.* 2025;18(1):89. doi:10.21203/rs.3.rs-5680571/v1.

48. Basak S, Chatterjee K. DPAD: data poisoning attack defense mechanism for federated learning-based system. *Comput Electr Eng*. 2025;121(1):109893. doi:10.1016/j.compeleceng.2024.109893.
49. Rodríguez-Barroso N, Luzón MV, Herrera F. RAB2-DEF: dynamic and explainable defense against adversarial attacks in federated learning to fair poor clients. *Mach Intell Res*. 2026;23(1):133–46. doi:10.1007/s11633-025-1557-1.
50. Sun L, Tian J, Muhammad G. FedKC: personalized federated learning with robustness against model poisoning attacks in the Metaverse for consumer health. *IEEE Trans Consum Electron*. 2024;70(3):5644–53. doi:10.1109/tce.2024.3386932.
51. Munusamy S, Jothi KR. Blockchain-enabled federated learning with edge analytics for secure and efficient electronic health records management. *Sci Rep*. 2025;15(1):27524. doi:10.1038/s41598-025-12225-x.
52. Shi J, Wang W, Hu Y, Lin Y. FLDP-DM: defense method against poisoning attacks with local differential privacy in federated learning systems. *IEEE Trans Cogn Commun Netw*. 2026;12(2):342–54. doi:10.1109/tccn.2025.3561316.
53. Jaszcz A, Połap D, Srivastava G. MultiMOORA-guided federated learning with KD Stabilization for health monitoring devices as protection against poisoning attacks. *IEEE Internet Things J*. 2026;13(4):6993–7001. doi:10.1109/jiot.2025.3638263.
54. Li Y, Tan Q, Shin BS. CryptoGAN: privacy-preserving federated generative adversarial networks with Homomorphic encryption in healthcare systems. *IEEE Trans Comput Soc Syst*. 2025;12(6):5330–41. doi:10.1109/tcss.2025.3570990.
55. Wang X, Hu J, Lin H, Liu W, Moon H, Piran MJ. Federated learning-Empowered disease diagnosis mechanism in the internet of medical things: from the privacy-preservation perspective. *IEEE Trans Ind Inform*. 2023;19(7):7905–13. doi:10.1109/tii.2022.3210597.
56. Sumitra, Sumitra J, Shenoy MV. HAFedL: a Hessian-aware adaptive privacy preserving horizontal federated learning scheme for IoT applications. *IEEE Access*. 2024;12:126738–53.
57. Zhang L, Xu J, Vijayakumar P, Sharma PK, Ghosh U. Homomorphic encryption-based privacy-preserving federated learning in IoT-enabled healthcare system. *IEEE Trans Netw Sci Eng*. 2023;10(5):2864–80. doi:10.1109/tNSE.2022.3185327.
58. Chen R, Xia H, Wang K, Xu S, Zhang R. KDRSFL: a knowledge distillation resistance transfer framework for defending model inversion attacks in split federated learning. *Future Gener Comput Syst*. 2025;166:107637.
59. Jeter TR, Nguyen T, Alharbi R, Seo JT, Thai M. Securing federated learning against active reconstruction attacks. *ACM Trans Internet Technol*. 2026;26(1):1–27. doi:10.1145/3762639.
60. Jiang L, Ma L, Yang G. Shadow defense against gradient inversion attack in federated learning. *Med Image Anal*. 2025;105(9):103673. doi:10.1016/j.media.2025.103673.
61. Zheng L, Cao Y, Yoshikawa M, Shen Y, Rashed EA, Taura K, et al. Sensitivity-aware differential privacy for federated medical imaging. *Sensors*. 2025;25(9):2847. doi:10.3390/s25092847.
62. Yuksel BB, Yilmazer Metin A. Federated learning with homomorphic encryption for secure real time ECG anomaly detection: a multi institutional privacy preserving framework. *Biomed Signal Process Control*. 2026;116(4):109557. doi:10.1016/j.bspc.2026.109557.
63. Barnawi A, Chhikara P, Tekchandani R, Kumar N, Alzahrani B. A differentially privacy assisted federated learning scheme to preserve data privacy for IoMT applications. *IEEE Trans Netw Serv Manag*. 2024;21(4):4686–700. doi:10.1109/tnsm.2024.3393969.
64. Wang W, Li X, Qiu X, Zhang X, Brusica V, Zhao J. A privacy preserving framework for federated learning in smart healthcare systems. *Inf Process Manag*. 2023;60(1):103167. doi:10.1016/j.ipm.2022.103167.
65. Stripelis D, Gupta U, Saleem H, Dhinagar N, Ghai T, Anastasiou C, et al. A federated learning architecture for secure and private neuroimaging analysis. *Patterns*. 2024;5(8):101031. doi:10.1016/j.patter.2024.101031.
66. Jiang S, Liu Y, Wang Z, Li Y, Chi H, Du X. Hybrid makes better: hybrid differential privacy medical image classification based on federated learning. *IEEE Trans Dependable Secur Comput*. 2025;22(5):5265–76. doi:10.1109/tdsc.2025.3565196.
67. Benabderrahmane F, Kerkouche E, Bouchemal N. Risk-aware privacy-preserving federated learning for remote patient monitoring: a multi-layer adaptive security framework. *Appl Sci*. 2025;16(1):29. doi:10.3390/app16010029.

68. Lin Y, Wang L, Zhang Z, Liu X. A hardware-accelerated federated learning with Intuitionistic fuzzy security measures for AI adaptive image computing. *IEEE Access*. 2025;13(12):198358–72. doi:10.1109/access.2025.3633060.
69. Darzi E, Dubost F, Sijtsema NM, Ooijen PMA. Exploring adversarial attacks in federated learning for medical imaging. *IEEE Trans Ind Inform*. 2024;20(12):13591–9. doi:10.1109/tii.2024.3423457.
70. Rehman A, Razzak I, Xu G. Federated learning for privacy preservation of healthcare data from smartphone-based side-channel attacks. *IEEE J Biomed Health Inform*. 2023;27(2):684–90. doi:10.1109/jbhi.2022.3171852.
71. Naeem H, Alsirhani A, Alserhani FM. Adversarial-aware federated learning with dynamic client weighting for healthcare record security. *ICT Express*. 2026;8(12):9603. doi:10.1016/j.icte.2026.03.010.
72. Matthew UO, Rosa RL, Saadi M, Rodríguez DZ. Interpretable AI framework for secure and reliable medical image analysis in IoMT systems. *IEEE J Biomed Health Inform*. 2026;30(5):3689–702. doi:10.1109/jbhi.2025.3591737.
73. Kalodanis K, Feretzakis G, Anastasiou A, Rizomiliotis P, Anagnostopoulos D, Koumpourous Y. A privacy-preserving and attack-aware AI approach for high-risk healthcare systems under the EU AI act. *Electronics*. 2025;14(7):1385. doi:10.3390/electronics14071385.
74. Tawfik M, Abu-Ein AA, Noaman HM, Abdelhaliem AH, Fathi IS. FedMedSecure: federated few-shot learning with cross-attention mechanisms and explainable AI for collaborative healthcare cybersecurity. *Sci Rep*. 2025;15(1):40050. doi:10.1038/s41598-025-25107-z.
75. Alharbi AA. Federated transfer learning for attack detection for internet of medical things. *Int J Inf Secur*. 2024;23(1):81–100. doi:10.1007/s10207-023-00805-9.
76. Nassar MO, AL-Mashagba FF. PRIFLEX: a secure federated learning framework for evaluating privacy leakage and defense in cross-modal medical data. *Mesop J CyberSecur*. 2025;5(3):1057–80.
77. Azad MA, Shah SA. Privacy-preserving federated learning system under malicious collaborators and aggregators. *Future Gener Comput Syst*. 2026;183(1):108530. doi:10.1016/j.future.2026.108530.
78. Singh AK, Kumar A, Subbiyan B, L. K, Prasath AR, Athithan S. MIoT-FL: a privacy-preserving federated learning architecture for securing ECG data in heart disease prediction. *Frankl Open*. 2026;15(21):100600. doi:10.1016/j.fraope.2026.100600.
79. Phan QB, Nguyen TT. A secure multi-modal federated transfer learning system for ECG classification. *Neurocomputing*. 2026;681(4):133402. doi:10.1016/j.neucom.2026.133402.
80. Melhem A, Aleroud A, Al-Mamun A, Ibrahim MI. Clustered federated learning for healthcare analytics: a dual blockchain-functional encryption method for fortified aggregation. *IEEE Internet Things J*. 2026;13(8):15959–75.
81. Zhao Y, Ullah F, Mahmood K, Saeed S, Mohammad N, Raza U. XAI-EdgeSFL: explainable edge intelligence with adaptive intrusion-resilient split federated learning for consumer healthcare ecosystems. *IEEE Trans Consum Electron*. 2026;72(1):2185–96.
82. Shi Q, Ren L, He X. Enhancing black-box membership inference attacks in federated learning. *J Inf Secur Appl*. 2026;96:104302. doi:10.1016/j.jisa.2025.104302.