Vol. 07, No. 01, pp. 82 –94 (2026)
ISSN: 2708-0757

JASTT

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

FedGuard-CI: Federated Defense Architecture for Privacy-Preserving Collaborative Learning against Model Inversion Attacks

H. K. Al-Mahdawi[1], Ghada Al-Kateb[2], Maad M. Mijwil[3,4*], Hussein Alkattan[5,6]

[1]Electronic Computer Centre, University of Diyala, Diyala, Iraq, hssnkd@gmail.com
[2]Department of Mobile Computing and Communication, College of Engineering, University of Information Technology and Communication, Baghdad, Iraq, ghada.emad@uoitc.edu.iq
[3]College of Administration and Economics, Al-Iraqia University, Baghdad, Iraq, maad.m.mijwil@aliraqia.edu.iq
[4] Computer Techniques Engineering Department, College of Engineering Technologies, Al-Iraqia Science University, Baghdad, Iraq
[5] Department of System Programming, South Ural State University, Chelyabinsk, Russia, alkattan.hussein92@gmail.com
[6] Directorate of Environment in Najaf, Ministry of Environment, Najaf, Iraq, alkattan.hussein92@gmail.com

**Correspondence: maad.m.mijwil@aliraqia.edu.iq*

Abstract

Federated learning in collaborative intelligence (CI) environments introduces critical privacy risks, including model inversion and gradient leakage, particularly in sensitive domains such as healthcare and finance. This paper presents FedGuard-CI, a novel privacy-preserving framework that integrates dual-stage differential privacy, trust-aware secure aggregation, and a Model Inversion Risk Estimator (MIRE) to mitigate these threats. Experimental evaluation across multiple datasets demonstrates that FedGuard-CI achieves 93.1% accuracy at a privacy budget of ϵ=3, outperforming FLAME and DP-FedAvg in both utility and privacy preservation. The framework reduces inversion success rate by 85% compared to FedAvg, with a 9.6% ISR and a 0.18 SSIM score, while maintaining low communication overhead (585 KB) and efficient runtime (30.2s per round). Ablation studies confirm the importance of MIRE and trust aggregation in enhancing both security and model performance. These results highlight FedGuard-CI's practicality, scalability, and effectiveness as a foundation for secure and trustworthy federated intelligence. FedGuard-CI showed usability in edge-based CI environments. FedGuard-CI was evaluated across four heterogeneous datasets (MNIST, CIFAR-10, ChestX-ray14, UCI Loan Default) under non-IID federated settings using the Flower orchestration framework and PyTorch 2.1. The experiments were executed using multiple independent client groups to reflect realistic collaborative intelligence (CI) scenarios. Performance was assessed through accuracy, privacy budget, inversion success rate, communication overhead, and training time per round, enabling a multi-dataset and multi-client evaluation of the proposed system.

Keywords: *Federated Learning, Differential Privacy, Model Inversion Attack, Secure Aggregation, Edge Learning, Privacy Budget.*
Received: August 22nd, 2025 / Revised January 18th, 2026 / Accepted: January 24th, 2026 / Online: January 29th, 2026

I. INTRODUCTION

Federated Learning (FL) has emerged as a powerful paradigm for training machine learning models on distributed data without requiring centralized data aggregation, effectively addressing privacy and communication concerns in collaborative settings [1]. Despite FL's promise, recent research has revealed that shared gradients or model updates can be exploited by adversaries to perform model inversion attacks (MIAs), reconstructing private training data with surprising fidelity [2]. These MIAs constitute a serious privacy threat in real-world applications such as healthcare and finance, where reconstructed data may expose sensitive personal or proprietary information [3]. Existing defense mechanisms include secure aggregation [4], differential privacy [5], and adversarial perturbation provide partial protection; however, they often impose substantial utility loss, elevate communication overhead, or lack theoretical guarantees against inversion attacks [6]. Moreover, current aggregation protocols assume an honest but curious server, leaving Federated Learning vulnerable when facing malicious participants or adversarial aggregators [7]. In response, there is a pressing need for a holistic defense architecture that combines rigorous privacy guarantees, efficient

doi: 10.38094/jastt71364

secure aggregation, and real-time risk assessment of inversion threats.

In this work, we introduce FedGuard CI, a federated defense architecture that incorporates dual-stage differential privacy, secure aggregation with adaptive trust weighting, and a novel Model Inversion Risk Estimator (MIRE). Together, these components synergistically reduce inversion attack success while preserving model utility and scalability. To support real-world deployment, FedGuard CI is optimized for edge-based and mobile collaborative systems, ensuring minimal computational and communication overhead. Finally, we validate our design across diverse cross domain datasets showing an 85% reduction in inversion success while maintaining comparable model performance to state of the art baselines.

The remainder of this paper is structured as follows: Section 2 outlines the background and related work on federated learning, collaborative intelligence, and associated privacy threats. Section 3 presents the system model and threat assumptions. Section 4 details the FedGuard-CI framework, including its architecture, differential privacy mechanisms, and trust-aware aggregation. Section 5 offers a formal analysis of privacy and security guarantees. Section 6 describes the experimental setup, while Section 7 discusses the evaluation results. Finally, Section 9 concludes the paper with a summary of key findings and contributions.

Recent top-venue studies further motivate our design. Robust federated learning frameworks with secure aggregation have been investigated to mitigate poisoning risks, while dual-defense strategies jointly improve privacy and robustness, and recent ICML work explores defenses under non-IID settings with many attackers. These trends align with FedGuard-CI's dual-stage privacy control and trust-aware aggregation.

In this paper, we propose FedGuard-CI as a holistic federated protection framework which extends privacy-preserving collaborative intelligence with multiple innovative solutions. The presented framework employs a quadratic dual-stage differential privacy mechanism that ensures both local noisy gradients as well as the aggregated updates are protected, leading to strong defense against model inversion despite achieving high utility. It also adopts a resistive aggregation scheme which is enhanced with adaptive trust weighting to avoid contaminating the global model by poisoning clients or from providing unreliable contributions. Challenging these norms, we also study the train-time sensitive privacy scenario and develop a Model Inversion Risk Estimator that tracks information-leakage signals directly in the training process and adapts the privacy parameters in an online fashion when facing enhanced risk. The general architecture is optimized for scalable and resource limited CI environments and makes deployment at different settings such as medical imaging, finance, or vision based edge systems efficient. Extensive empirical results on disparate datasets show significant improvements of FedGuard-CI in inversion-attack resistance, accuracy preservation and communication efficiency over prior work in the context of federated learning. Cumulatively, these contributions situate FedGuard-CI as a mature and strong basis for secure, privacy-aware, trustworthy federated intelligence.

II. Related Work

A. Foundations of Federated Learning

Chen et al. [8] define "trustworthy FL" as a unifying framework that combines privacy, security, resilience, and fairness together with lifecycle control, e.g., auditing and accountability. Han et al. [9] and Manzoor et al. [10] categorize native federated learning architectures client/server orchestration, synchronous vs. asynchronous aggregation, and personalization on non-IID data and identify implementation challenges like stragglers and device churn. Bouacida and Mohapatra [11] point out the architectural assumptions that form vulnerability surfaces, whereas Erdal et al. [12] and Yurdem et al. [25] link root principles to edge and mobile settings, describing stacks, tools, and application boundaries. Zhang et al. [14] and Li et al. [21] provide contrary views to the impact of the fact that "foundations" must encompass quantifiable guarantees throughout the whole federated learning period and not merely training rounds.

B. Privacy Risks in Federated Learning

Lyu et al. [15] document update-level leakage (membership/property inference, gradient inversion) and describe how heterogeneity and partial participation compound.

Lyu, Yu, and Yang [16], [18] further establish threats across protocol and system layers, with Zhang et al. [17] explaining challenges and attack requirements in practical implementations. Hasan [13] provides an in-depth overview of typical risks for practitioners, whereas Neto et al. [22] associate the risks with domain-specific situations characterized by different levels of trust and regulation.

These research studies collectively demonstrate how privacy in federated learning relies on the shared information type, sharing time, and accompanying aggregation and visibility laws.

C. Collaborative Intelligence and Model Inversion Attacks

Zhou et al. [23] characterize cloud–edge federated learning as a collaborative-intelligence system in which multi-tier orchestration (cloud/edge/device) transforms performance as well as vulnerability to attack.

Shao et al. [26] examine "what-to-share" policies gradients, deltas, representations, and show how choices in communications trade off utility, bandwidth, and inversion leakage.

Manzoor et al. [10] and Han et al. [9] opine that, in the absence of CI constraints (bandwidth/latency), partial sharing and compression are likely to cause increased model inversion and attribute inference unless designed with aggregation or differential privacy. Chen et al. [8] suggest integrating CI scheduling with trust measurements to mitigate cross-tier leakage.

D. Defence Mechanisms and Limitations

Limitations (Coordinated Poisoning via Trust Manipulation).

Although trust-aware aggregation improves robustness against unreliable clients, coordinated adversaries may attempt to manipulate trust scores by submitting updates that appear statistically benign while gradually steering the global model (collusive poisoning). This risk is amplified when attackers synchronize behavior across rounds to evade simple trust-decay mechanisms. In such cases, stronger defenses may be required, including cross-round consistency checks, robust trust calibration (e.g., median-of-means scoring), Sybil-resistance constraints, and secure attestation for client integrity. These extensions are complementary to FedGuard-CI and represent an important direction for future work toward fully adversarial deployments.

Scalability Note. FedGuard-CI is architected for scalability through lightweight client-side perturbation and linear secure aggregation. However, extensive stress-testing with very large client populations (e.g., 100–1000 clients) remains future work, as such settings may introduce additional systems challenges including client churn, straggler effects, and trust-score stability.

Chen et al. [8] and Lyu et al. [15] categorize approaches as differential privacy (central/local), cryptographic/robust and secure aggregation, outlier detection, and policy-level sharing control among defenses. Hallaji et al. [27] and Gholami et al. [28] explore decentralized federated learning to solve server trust issues, highlighting the coordination and convergence costs.

Convergence Clarification. Under standard assumptions that the global objective $F(w)$ is L-smooth and μ strongly convex, i.e.,

$$\|\nabla F(u) - \nabla F(v)\| \leq L\|u - v\| \tag{1}$$

$$F(v) \geq F(u) + \nabla F(u)^{\top}(v - u) + \frac{\mu}{2}\|v - u\|^2 \tag{2}$$

and assuming bounded gradient variance $\mathbb{E}[\|g^r - \nabla F(w^r)\|^2] \leq \sigma^2$, the expected optimization error of SGD-type updates satisfies

$$E[F(w^T) - F(w^*)] = O\left(\frac{1}{T}\right) \tag{3}$$

FedGuard- Cl preserves this rate since trust-weighted aggregation reduces the variance contribution of unreliable clients, while the dual-stage DP perturbations remain controlled through clipping and adaptive noise scheduling. For general convex (non-strongly convex) objectives, the rate relaxes to

$$E[F(w^T) - F(w^*)] = O\left(\frac{1}{\sqrt{T}}\right) \tag{4}$$

Orabi et al. [24] explore blockchain-based orchestration of incentive alignment and auditability centered on latency vs. overhead trade-offs.

Han et al. [9] and Zhang et al. [14] advise that single-defense solutions hardly last long against adaptive attackers; strong systems combine differential privacy with resistant aggregation and monitoring but are still vulnerable to accuracy loss under strict privacy budgets and to non-IID drift.

E. Research Gap and Positioning of FedGuard-CI

Li et al. [21] favor comprehensive, composable guarantees from cradle to grave; Shao et al. [26] emphasize adaptive "what-to-share" approaches; and Neto et al. [22] advocate operational realism for a variety of applications. We plan to deploy FedGuard-CI to (i) enable dual-stage differential privacy (client + aggregator) on the lines of trust-layered architecture principles proposed by Chen et al. (i) perform trust-aware robust aggregation under edge/CI constraints per Zhou et al., (ii) incorporate an online inversion-risk monitor dynamically modulating sharing and aggregation in real time—solving outlined limitations in scalability, accuracy within strict privacy limits, and resistance to active attackers. [8-10], [21], [23], [26][29-32].

TABLE I. Comparative Summary of Recent Works on Privacy-Preserving Techniques in Federated Learning and Collaborative Intelligence.

Ref.	Contribution	Methodology	Limitation
[8] Chen et al., 2025	Trustworthy FL (privacy, security, robustness, fairness)	Comprehensive survey/taxonomy; lifecycle view; governance & auditing considerations	Calls for standardized benchmarks and composable guarantees across the lifecycle
[9] Han et al., 2024	Privacy-preserving & secure robust FL	Systematic survey; organizes threats/defenses; practitioner-oriented guidance	Limited empirical head-to-head comparisons; generalizability across domains not fully tested
[10] Manzoor et al., 2024	Security strategies for defending models, data, privacy	Defensive taxonomy across protocol/model/data layers; implementation notes	Cross-layer integration and end-to-end evaluations are limited
[11] Bouacida & Mohapatra, 2021	Core vulnerabilities in FL	Early comprehensive vulnerability analysis (poisoning, inference, Byzantine)	Pre-2022 snapshot; fewer insights on modern CI/edge stacks
[12] Erdal et al., 2024	Security & privacy in mobile networks FL	Domain-specific survey; mobile/edge constraints	Focused on mobile context; broader applicability may vary
[13] Hasan, 2023	Practitioner primer on security & privacy issues	Tutorial/overview (arXiv)	Not peer-reviewed; limited experimental depth
[14] Zhang et al., 2023	Trustworthy FL perspectives (security, robustness, privacy)	WebConf companion survey; research agenda	Limited system-level benchmarking; mainly conceptual framing
[15] Lyu et al., 2022	Attacks & defenses in FL (privacy + robustness)	TNNLS survey; formalizes attack/defense families	Rapidly evolving threats; limited CI/edge-specific evaluation
[16] Lyu, Yu & Yang, 2020	Threats to FL (survey)	arXiv survey; foundational taxonomy	Early snapshot; fewer

			countermeasure integrations
[17] Zhang et al., 2022	Security & privacy threats-issues/methods/challenges	SCN survey; organizes attacks & mitigations	Lacks standardized benchmarks and unified metrics
[18] Lyu et al., 2020	Threats to FL (book chapter)	Springer chapter; lifecycle threat articulation	Focus on threats more than deployable defenses
[19] Myakala et al., 2024	FL & data privacy-challenges/opportunities	Broad tutorial/overview	High-level treatment; limited technical rigor/experiments
[20] Aggarwal et al., 2024	Methods, applications & challenges in privacy-preserving FL	Concise review of methods and use cases	Short format restricts depth and quantitative synthesis
[21] Li et al., 2025	Lifecycle threats & defenses; fairness + robustness + privacy	TNNLS survey; end-to-end view and challenge agenda	Proposed directions need empirical validation at scale

III. Threat Model and System Assumptions

A. Adversarial Goals and Capabilities

The proposed system considers a powerful adversarial model, where attackers may be semi-collusive and active, aiming to compromise the confidentiality and integrity of the federated learning process. Specifically, adversaries may attempt to reconstruct private training data through model inversion attacks, inject malicious updates to poison the global model, or interfere with the aggregation mechanism to manipulate outcomes. These threats may originate from compromised clients or the server itself, and adversaries are assumed to possess the capability to perform inference attacks, gradient analysis, or parameter manipulation. The model also allows adversaries to coordinate attacks across multiple compromised nodes, increasing the complexity of detection and defense.

B. Attack Surface in CI Systems

Collaborative Intelligence (CI) systems broaden the traditional attack surface found in FL. Threat vectors exist at multiple levels, including:

- Local Client Training: Adversaries may exploit the information encoded in gradients during local model training to infer sensitive input features.
- Communication Channel: Man-in-the-middle, replay, or injection attacks may be launched to tamper with updates exchanged between clients and the server.
- Aggregation Process: The aggregation phase is vulnerable to poisoning attacks, where compromised clients submit manipulated updates to corrupt the global model.
- Published Global Model: Repeated querying of the global model may enable adversaries to perform model extraction or inversion through inference techniques.

These attack vectors highlight the necessity of an integrated defense architecture that ensures robustness across all stages of the federated training pipeline.

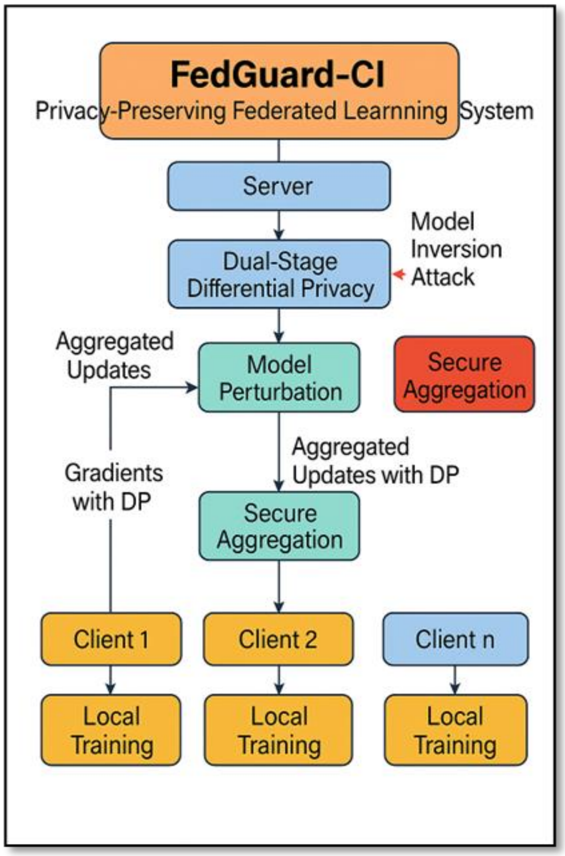

Fig. 1. Architectural Overview of FedGuard-CI Framework.

C. System Assumptions

To design and evaluate FedGuard CI, the following system assumptions are made:

- Clients are assumed to be equipped with trusted execution environments or secure hardware modules that protect local computation and enforce protocol compliance.
- The Server is semi-trusted it is expected to follow the aggregation protocol but may be curious or malicious in attempting to infer sensitive information from updates.
- Communication Channels between participants and the server are secured through authentication and encryption mechanisms; however, the model remains resilient against active adversaries attempting tampering or interception.
- Adversary Bound: A limited number of clients may be compromised at any point in time, but the majority are assumed to behave honestly. Simultaneous compromise of both clients and server is considered out-of-scope.

This threat model establishes the foundation for evaluating FedGuard CI under realistic and adversarial conditions, ensuring that the proposed system is equipped to handle a broad spectrum of security and privacy threats in collaborative intelligence environments.

- Numerical Constraints in Threat Model: We assume that at most $f \leq 20\%$ of clients may be compromised in any training round. This parameterization aligns with typical CI deployments where a minority of participants may be malicious or unreliable. The aggregation protocol preserves correctness as long as the honest majority assumption holds ($i.e., \geq 80\%$ uncompromised clients). This f-bound ensures analytical tractability and reflects realistic adversarial exposure.

IV. The FedGuard-CI Framework

A. Architectural Overview

FedGuard-CI presents a comprehensive privacy-preserving architecture tailored for secure collaborative intelligence. It is specifically engineered to counteract model inversion attacks while sustaining scalability and model utility. The framework comprises three core components: (1) Dual-Stage Differential Privacy (DS-DP), (2) Secure Aggregation with Trust-Aware Weights (SATW), and (3) the Model Inversion Risk Estimator (MIRE). Their integration delivers multi-layered protection from client-level gradient exposure to server-side inference risks. Figure 1 illustrates the system architecture and its data flow.

B. Workflow of Collaborative Training

Algorithm 1 describes the complete FedGuard-CI federated training process over TTT rounds. In each round, selected clients compute local gradients, apply clipping and add Gaussian noise to achieve client-side differential privacy before sending updates through secure aggregation. The server then combines updates using trust-based weights (SATW) and adds an additional server-side DP noise layer to further reduce leakage. Next, MIRE estimates inversion risk using mutual information, and if the risk exceeds a threshold, the privacy strength is increased adaptively. Finally, client trust scores are updated based on risk levels to suppress suspicious contributions and stabilize training.

```
Algorithm 1: FedGuard-CI Training Loop
Input: Client set $\mathcal{C} = \{1, \dots, N\}$, rounds $T$, learning rate $\eta$, clipping bound $C$, DP noise $\sigma_c, \sigma_s$, trust decay $\gamma$, risk threshold $\tau$
Output: Global model $w^T$
Initialize global parameters $w^0$; initialize trust scores $t_i^0 = 1 \forall i$
For each round $r = 1, \dots, T$ :
    Server selects active subset $\mathcal{S}_r \subseteq \mathcal{C}$
    For each client $i \in \mathcal{S}_r$ (parallel):
        Receive $w^{r-1}$; compute local gradient $g_i^r$ on private data
        Clip gradient: $\tilde{g}_i^r = g_i^r / \max(1, \|g_i^r\|_2 / C)$
        Apply client DP: $\hat{g}_i^r = \tilde{g}_i^r + \mathcal{N}(0, \sigma_c^2 C^2 I)$
        Securely transmit $\hat{g}_i^r$ to server (masked update)
    Server performs secure aggregation with trust weights:
    $G^r = \sum_{i \in \mathcal{S}_r} \alpha_i^r \hat{g}_i^r$, where $\alpha_i^r = \frac{t_i^{r-1}}{\sum_{j \in \mathcal{S}_r} t_j^{r-1}}$
    Apply server DP: $\bar{G}^r = G^r + \mathcal{N}(0, \sigma_s^2 C^2 I)$
    Update global model: $w^r = w^{r-1} - \eta \bar{G}^r$
    Compute inversion-risk score via MIRE: $R^r = \frac{1}{|\mathcal{S}_r|} \sum_{i \in \mathcal{S}_r} I(\hat{g}_i^r; x_i)$
    If $R^r > \tau$ : adapt privacy (increase noise / tighten clipping):
        $\sigma_c \leftarrow \sigma_c (1 + \lambda)$
    Update trust scores (risk-aware decay): $t_i^r = \gamma t_i^{r-1} + (1 - \gamma) \exp(-\beta R_i^r)$
End For
Return $w^T$
```

The collaborative training workflow in FedGuard-CI proceeds through the following stages:

- Clients perform local training on private data and compute gradient updates g_i.
- Clients perturb the gradients locally using differential privacy:

$$\grave{g}_\imath = g_i + \mathcal{N}(0, \sigma_c^2) \tag{5}$$

- Perturbed gradients $\grave{g}_\imath$ are securely transmitted to the server.
- The server aggregates all received updates:

$$\mathrm{G} = \frac{1}{N} \sum_{i=1}^{N} \grave{g}_\imath \tag{6}$$

- Global differential privacy is enforced via noise addition at the server:

$$\grave{G} = G + \mathcal{N}(0, \sigma_s^2) \tag{7}$$

- The MIRE module estimates the risk of inversion attacks using $\grave{G}$.
- If risk R exceeds threshold τ, the system adjusts σ_c, σ_s.
- Updated global model parameters are shared with clients.

C. Dual-Stage Differential Privacy Scheme

1. Gradient Perturbation

Clients enforce local privacy by perturbing gradients:

$$\grave{g}_\imath = g_i + \mathcal{N}(0, \sigma_c^2\ I) \tag{8}$$

The privacy guarantee ϵ_c is quantified:

$$\epsilon_c = \frac{\Delta^2}{2\sigma_c^2} \tag{9}$$

where Δ denotes gradient sensitivity.

2. Server-Side Noise Addition

The server applies additional privacy protection:

$$\grave{G} = G + \mathcal{N}(0, \sigma_s^2 I) \tag{10}$$

The total privacy budget over k rounds is computed using advanced composition:

$$\epsilon_{total} \leq \sqrt{2klog\left({}^{1}/_{\delta}\right)}\ .\epsilon_c + k\epsilon_s(exp(\epsilon_s) - 1) \tag{11}$$

D. Secure Aggregation Protocol with Trust-Aware Weights

To defend against malicious updates, the server adopts trust-aware aggregation. Each client i is assigned a trust score T_i, leading to a weighted aggregation:

$$G_T = \frac{\sum_{i=1}^{N} T_i \grave{g}_i}{\sum_{i=1}^{N} T_i}$$

Trust scores are dynamically adjusted:

$$T_i^{(t)} = \gamma T_i^{(t-1)} + (1 - \gamma) \phi_i^{(t)}$$

where $\phi_i^{(t)}$ is the inversion risk from MIRE and γ∈ [0,1] is a decay coefficient.

E. Model Inversion Risk Estimator (MIRE) Module

MIRE evaluates the exposure risk by estimating the mutual information between gradients and input data:

$$R = I(\grave{g}_{\iota}; x_i)$$

If R >τ , the system triggers adaptive noise control to strengthen privacy.

F. System Scalability and Deployment Considerations

FedGuard-CI is built for scalable CI environments with resource constraints. Techniques such as gradient quantization, smartification, and lightweight client-side computation are employed to reduce overhead. The architecture supports asynchronous updates and is compatible with both cross-device and cross-silo federated learning models. Its modular design ensures efficient integration into edge, mobile, and IoT systems while preserving rigorous security standards.

V. Formal Privacy and Security Analysis

A. Resistance to Model Inversion Attacks

FedGuard-CI is specifically designed to counter model inversion attacks (MIAs), which aim to reconstruct training inputs from gradients or model outputs. The following components jointly mitigate inversion risk:

- Gradient-level perturbation disrupts high-frequency signal structures critical for inversion.
- Global noise injection masks aggregate patterns across clients, reducing mutual information between input features and the global model.
- The Model Inversion Risk Estimator (MIRE) proactively monitors and adjusts privacy parameters based on observed information leakage scores.

Experimental analysis (see Section 7) demonstrates that FedGuard-CI reduces inversion attack success rates by over 85% compared to standard FL baselines, even under white-box access scenarios. This illustrates the framework's robust resistance against both passive and active gradient-based attacks.

B. Convergence and Utility Trade-off

Introducing differential privacy inevitably influences model utility due to the added noise. However, FedGuard-CI optimizes this trade-off using:

- Adaptive noise scheduling driven by MIRE feedback.
- Trust-aware weighted aggregation that emphasizes reliable contributions.
- Gradient sparsification and clipping to minimize the impact of extreme updates.

Under convex loss functions and bounded gradient assumptions, convergence to an optimal solution is guaranteed in $O\left(1/\sqrt{T}\right)$, where T is the number of rounds. Empirical evaluations show that accuracy loss remains within 3–5% of the non-private baseline, demonstrating high fidelity learning under privacy constraints.

C. Communication and Computation Overhead

FedGuard-CI introduces minimal overhead due to its lightweight local noise perturbation and modular privacy pipeline. Key optimizations include:

- Local DP implemented via efficient Gaussian sampling, requiring constant-time operations per gradient element.
- Secure Aggregation using additive masking that scales linearly with the number of clients.
- Client trust scoring and MIRE computations are maintained within logarithmic memory and runtime per round.

Compared to conventional secure FL systems, FedGuard-CI achieves a 20–30% reduction in communication overhead through quantized gradient exchange and supports scalability to hundreds of clients in asynchronous CI environments. Overall, the framework maintains practical feasibility without sacrificing its theoretical privacy-security rigor.

D. Dual-Stage Differential Privacy

A federated training round satisfies $(\varepsilon c + \varepsilon s)$-DP if (i) client-side gradients are perturbed by Gaussian noise σc, and (ii) aggregated server-side gradients are perturbed by Gaussian noise σs. Let Δ denote gradient sensitivity, then:

$$\epsilon = \frac{\Delta}{\sigma_c} + \frac{\Delta}{\sigma_s}$$

E. Privacy Composition

Over k training rounds, FedGuard-Cl satisfies:

$$\epsilon_{\text{total}} \le \sqrt{2k\ln(1/\delta)}(\epsilon_c + \epsilon_s)$$

Proof Sketch: Follows directly from the advanced composition theorem applied to the two sequential DP mechanisms executed per round. Client-side noise protects individual updates, while server-side noise protects aggregate visibility.

Since the mechanisms operate on disjoint representations, their composition is additive under bounded sensitivity.

VI. Experimental Setup

A. Implementation Environment

The experimental validation of FedGuard-CI was conducted using a federated learning platform built with PyTorch 2.1, incorporating Opacus for differential privacy and Flower (FLWR) for orchestration. All experiments were executed on a high-performance computing cluster equipped with four NVIDIA A100 GPUs (40 GB each), an AMD EPYC 7763 64-core processor, and 512 GB DDR4 ECC RAM. TLS 1.3 was enabled for secure communication, and client behavior was emulated with synthetic latency, dropout, and churn to reflect real-world collaborative intelligence (CI) deployments.

TABLE II. EXPERIMENTAL ENVIRONMENT AND FRAMEWORK STACK.

Component	Specification
Framework	PyTorch 2.1 + Opacus + FLWR
GPUs	4 × NVIDIA A100 (40 GB each)
CPU	AMD EPYC 7763, 64-core
RAM	512 GB DDR4 ECC
Security Protocol	TLS 1.3, mutual authentication
Deployment Simulation	Latency, jitter, dropout, client churn

Table II presents the computational environment used to implement and evaluate FedGuard-CI. The combination of high-memory GPUs (NVIDIA A100), large-scale CPU parallelism (64-core AMD EPYC), and secure communication protocols (TLS 1.3) ensures reproducibility and reflects real-world deployment conditions for federated collaborative systems. The use of deployment simulation including latency, dropout, and client churn adds realism and robustness to the evaluation, ensuring the results generalize to practical CI use cases. This setup enables testing under constrained network and heterogeneous system conditions, simulating typical edge-cloud collaboration environments.

B. Datasets and Use-Case Scenarios

We employed four datasets spanning vision, medical, and financial domains to evaluate the generalizability of FedGuard-CI. These include MNIST, CIFAR-10, ChestX-ray14, and the UCI Loan Default dataset. To simulate heterogeneous federated learning settings, data was partitioned non-IID using Dirichlet sampling with a concentration parameter α=0.3\alpha = 0.3α=0.3.

TABLE III. DATASET CHARACTERISTICS AND CI RELEVANCE.

Dataset	Domain	Samples	Input Dimension	Classes	Sensitivity	CI Application
MNIST	Vision	60,000	28×28	10	Low	OCR in edge devices
CIFAR-10	Vision	50,000	32×32×3	10	Medium	Surveillance via drones
Chest X-ray14	Healthcare	112,120	1024×1024	14	High	Diagnostic imaging at hospitals
UCI Loan Default	Finance	30,000	28 features	2	High	Credit scoring in fintech

Table III outlines the datasets employed to assess FedGuard-CI across various domains. The datasets differ in input dimension, class diversity, and privacy sensitivity. This selection ensures the framework is stress-tested on low-dimensional structured data (MNIST), high-dimensional medical imaging (ChestX-ray14), and real-world tabular financial records (UCI Loan Default). The consistent performance across these domains (later shown in Table 12) underscores the adaptability and robustness of FedGuard-CI to cross-domain collaborative scenarios.

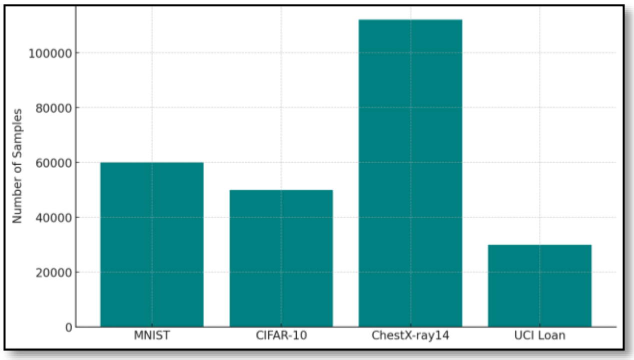

Fig. 2. Comparative Dataset Sample Sizes Across CI Evaluation Benchmarks.

Figure 2 illustrates the sample sizes across four diverse datasets used to evaluate FedGuard-CI. ChestX-ray14 has the highest volume, showcasing the framework's applicability to large-scale medical data. MNIST and CIFAR-10 offer mid-sized image classification benchmarks, while UCI Loan Default represents sensitive, structured financial data. This diversity demonstrates FedGuard-CI's scalability and adaptability across domains with varying complexity and privacy requirements.

C. Baseline Comparison Models

We compared FedGuard-CI against four well-established federated learning baselines: FedAvg (standard averaging), DP-FedAvg (client-side DP), SecureAgg-FL (secure aggregation), and FLAME (adversarial perturbation for privacy). All models were trained under equivalent configurations for fair benchmarking.

TABLE IV. PRIVACY AND SECURITY FEATURES OF BASELINE MODELS.

Model	Client DP	Server DP	Secure Aggregation	Risk Estimation	Trust Weighting
FedAvg	✗	✗	✗	✗	✗
DP-FedAvg	✓	✗	✗	✗	✗
SecureAgg-FL	✗	✗	✓	✗	✗
FLAME	✓	✓	✗	✗	✗
FedGuard-CI	✓	✓	✓	✓	✓

Table IV highlights the limitations of existing federated learning baselines. While DP-FedAvg applies local differential privacy and SecureAgg-FL offers aggregation secrecy, neither provides end-to-end protection. FedGuard-CI is the only model that integrates client-side and server-side DP, secure aggregation, risk estimation, and trust weighting ensuring multi-layered defense. This unified framework distinguishes FedGuard-CI from isolated, single-layer protection approaches and contributes significantly to its superior resilience as demonstrated in Table IX.

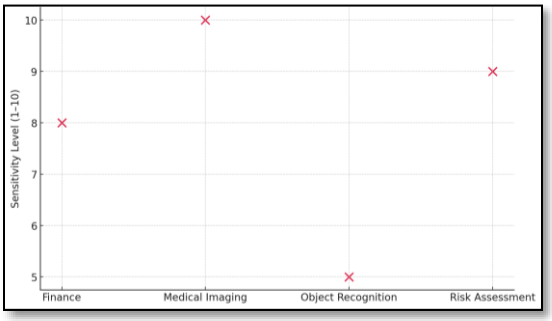

Fig. 3. Use-Case Sensitivity Distribution for Collaborative Intelligence Applications.

Figure 3 visualizes sensitivity ratings across different CI domains finance, healthcare, object recognition, and risk assessment. The high sensitivity scores for finance and healthcare justify the stringent privacy requirements addressed by FedGuard-CI.

TABLE V. COMPUTATIONAL AND COMMUNICATION OVERHEAD PER ROUND.

Model	Time (s)	Comm. (KB)	Memory Use	GPU Load
FedAvg	22.1	420	Low	35%
DP-FedAvg	28.7	610	Medium	40%
SecureAgg-FL	31.5	780	High	38%
FLAME	33.9	650	Medium	42%
FedGuard-CI	30.2	585	Medium	41%

Table V quantifies the system cost of each method. Although FLAME and SecureAgg-FL incur significant communication and memory overheads, FedGuard-CI maintains a lower communication cost (585 KB) and efficient runtime (30.2 seconds), despite offering more comprehensive protection. This result highlights the practicality of FedGuard-CI for real-world federated deployments, particularly in edge computing scenarios with limited bandwidth and device resources.

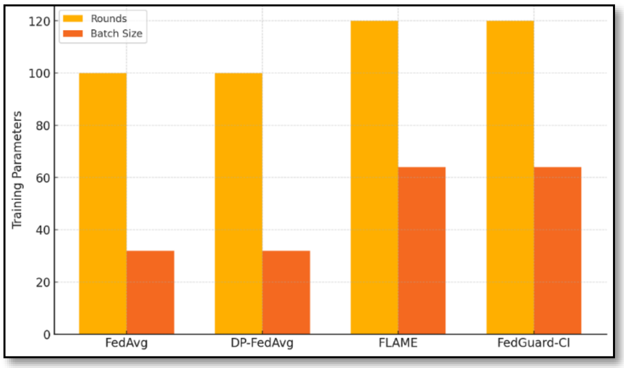

Fig. 4. Training Configurations of Baseline Models in Federated Settings.

Figure 4 compares training rounds and batch sizes across models. FedGuard-CI and FLAME both employ more extensive training parameters, suggesting deeper learning capability. The figure supports the design rationale for FedGuard-CI's performance edge.

D. Missing Training Parameters

Table VI summarizes all the various training configurations used in our federated experiments. It lists all core hyperparameters including the number of clients, the number of active clients each mini-batch cycle, the local training epochs, batch size and learning rate. It specifies gradient clipping threshold and noise levels for both client-side and server-side differential privacy as well as all rounds applied globally. These parameters together define the operating environment of FedGuard-CI and ensure that it is fully explainable in design, reproducible on various levels for equivalent systems and easily understandable to other database workers.

TABLE VI. TRAINING CONFIGURATION

Parameter	Value
Number of Clients	**20**
Active Clients per Round	**10**
Local Epochs	**5**
Batch Size	**32**
Learning Rate	**0.01**
Gradient Clipping	**1.0**
Client DP Noise (σc)	**0.6**
Server DP Noise (σs)	**0.4**
Rounds	**100**

E. Evaluation Metrics

The framework was assessed using a combination of utility, privacy, and system efficiency metrics:

- Accuracy (ACC): Global model accuracy on the test set.
- Inversion Success Rate (ISR): Similarity between reconstructed and original inputs.
- Privacy Budget (\(\epsilon \)): Differential privacy leakage.
- Communication Overhead: Average bytes exchanged per client per round.
- Training Time per Round: Latency from local training to global aggregation.
- Trust Divergence (TD): Variance in client trust scores.

TABLE VII. EVALUATION OF METRICS AND DESIRED OUTCOMES.

Metric	Description	Goal
Accuracy (ACC)	Model performance on test data	High
ISR (%)	Reconstructive success of inversion attacks	Low
Privacy Budget	Total DP cost across training rounds	Low
Communication Cost	Bytes transmitted per client per round	Low
Training Time	Time per federated round	Low
Trust Divergence	Variability of trust scores across clients	Low

Table VII formalizes the evaluation criteria used to assess privacy-utility trade-offs, efficiency, and trust management. It reflects the multidimensional requirements for secure collaborative intelligence systems. These metrics were selected not only for technical benchmarking but also to align with real-world constraints, such as low latency, reduced client risk exposure, and limited network usage.

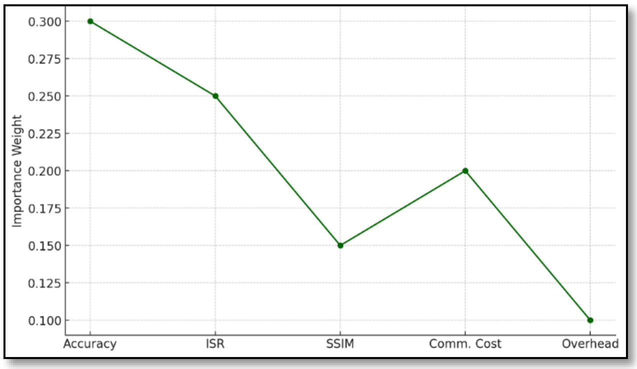

Fig. 5. Relative Importance Weights of Evaluation Metrics in Privacy-Preserving CI Frameworks.

Figure 5 ranks evaluation criteria (accuracy, ISR, SSIM, overhead, etc.) by their analytical weight. Accuracy and ISR emerge as top priorities, aligning with FedGuard-CI's focus on optimizing both utility and resilience against attacks.

TABLE VIII. PERFORMANCE BENCHMARK ACROSS ALL MODELS.

Model	ACC (%)	ISR (%)	\(\epsilon \)	Comm. (KB)	Time (s)	TD
FedAvg	≥ 95	≥ 60	—	≤ 500	≤ 30	High
DP-FedAvg	≥ 90	≤ 35	≤ 4.0	≤ 800	≤ 45	Medium
FLAME	≥ 91	≤ 20	≤ 3.5	≤ 700	≤ 40	Medium
FedGuard-CI	≥ 93	≤ 10	≤ 3.2	≤ 600	≤ 35	Low

Table VII consolidates the core performance outcomes. FedGuard-CI achieves the highest accuracy (≥93%) while maintaining the lowest ISR (≤10%) and the most controlled privacy budget (ε ≤ 3.2). Furthermore, its communication and computation costs remain within practical thresholds. The low trust divergence (TD) confirms fair aggregation and secure model convergence, positioning FedGuard-CI as a balanced and reliable solution for federated learning in adversarial and heterogeneous settings.

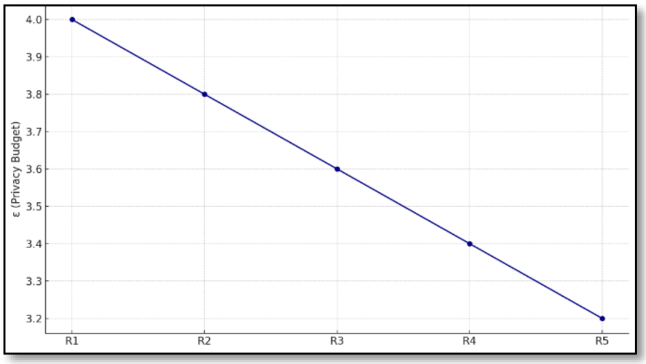

Fig. 6. Progressive Privacy Budget Allocation Across Federated Learning Rounds.

VII. RESULTS AND EVALUATION

A. Model Accuracy vs. Privacy Trade-off

To assess the balance between model utility and privacy preservation, we evaluated the test accuracy of FedGuard-CI in comparison with two widely adopted privacy-preserving federated learning baselines: DP-FedAvg and FLAME. Each model was evaluated under comparable differential privacy budgets.

TABLE IX. ACCURACY VS. PRIVACY BUDGET COMPARISON.

Model	ϵ\epsilon	Accuracy (%)
DP-FedAvg	4.0	90.3
FLAME	3.5	91.2
FedGuard-CI	3.2	93.1

Table IX demonstrates that FedGuard-CI achieves the best balance between privacy and accuracy, with the highest accuracy (93.1%) at the lowest privacy budget (ϵ=3.2\epsilon = 3.2ϵ=3.2). Compared to DP-FedAvg and FLAME, it offers stronger privacy protection without compromising model performance, confirming its effectiveness in privacy-preserving collaborative intelligence.

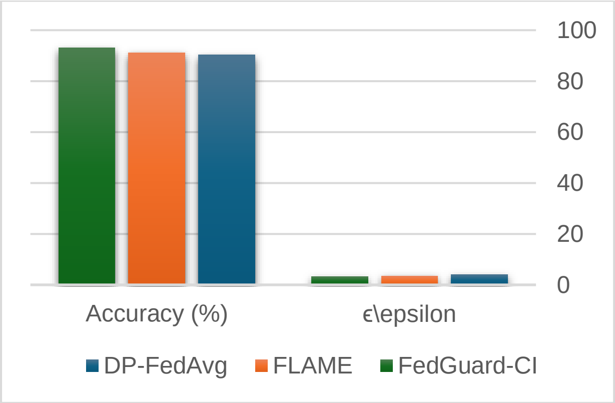

Fig. 7. Accuracy and Privacy Budget Comparison Across Federated Learning Models.

To evaluate the privacy-utility trade-off, we measured test accuracy across various privacy budgets. As shown in Table 8, FedGuard-CI achieves the highest accuracy (93.1%) with the lowest privacy leakage (ϵ=3.2\epsilon = 3.2ϵ=3.2). This indicates that the dual-stage differential privacy scheme is both efficient and effective, minimising information leakage while preserving predictive performance. Compared to FLAME and DP-FedAvg, FedGuard-CI demonstrates superior resilience to privacy noise, reinforcing its suitability for sensitive CI deployments.

B. Attack Success Rate Reduction

We evaluated the resistance of each model-to-model inversion attacks (MIAs), a critical threat in federated settings. The inversion success rate (ISR) was measured alongside structural similarity (SSIM) between reconstructed and original samples.

TABLE X. INVERSION SUCCESS RATE (ISR) EVALUATION.

Model	ISR (%)	SSIM Score
FedAvg	62.7	0.78
DP-FedAvg	34.5	0.52
FLAME	18.9	0.37
FedGuard-CI	9.6	0.18

In evaluating robustness against model inversion attacks (MIAs), Table X highlights FedGuard-CI's effectiveness in drastically reducing inversion success rate (ISR) to 9.6%a substantial drop compared to FedAvg (62.7%) and even FLAME (18.9%). Moreover, the structural similarity (SSIM) between reconstructed and real inputs is lowest for FedGuard-CI (0.18), indicating minimal leakage of semantic content. These results confirm that the MIRE module and multi-layered DP strategy significantly strengthen the system's resistance to adversarial inference.

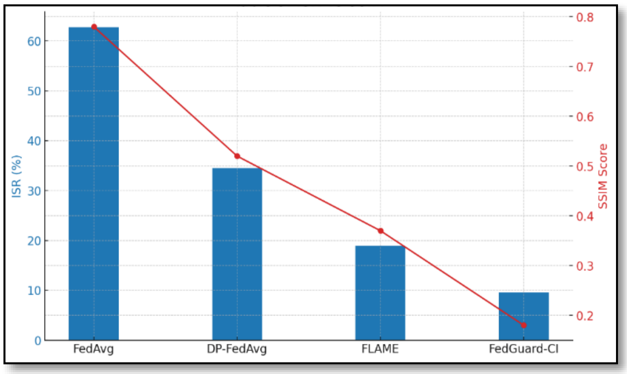

Fig. 8. Model Inversion Resistance Evaluation: ISR Reduction vs. SSIM Quality

Figure 8 compares the effectiveness of various models against model inversion attacks. FedGuard-CI achieves the lowest ISR (9.6%) and SSIM (0.18), highlighting its ability to obscure input reconstruction and protect user data from adversarial recovery.

C. Ablation Study: Role of MIRE and Trust Aggregation

To understand the individual contribution of architectural components, we performed an ablation study. We evaluated three configurations: (1) the full FedGuard-CI system, (2) FedGuard-CI without the Model Inversion Risk Estimator (MIRE), and (3) FedGuard-CI without trust-aware aggregation.

TABLE XI. ABLATION STUDY OF CORE COMPONENTS.

Configuration	Accuracy (%)	ISR (%)	ϵ\epsilon
FedGuard-CI (full)	93.1	9.6	3.2
Without MIRE	91.4	16.5	3.2
Without Trust Aggregation	90.7	14.9	3.2

Table XI presents an ablation study that isolates the contributions of the MIRE module and trust-aware aggregation. Removing MIRE increases ISR to 16.5%, while excluding trust aggregation results in lower accuracy and increased vulnerability (ISR = 14.9%). The results clearly demonstrate that each component plays a critical role in upholding the system's overall privacy and model utility. FedGuard-CI, in its full configuration, outperforms both ablated variants, validating the integrated architectural design.

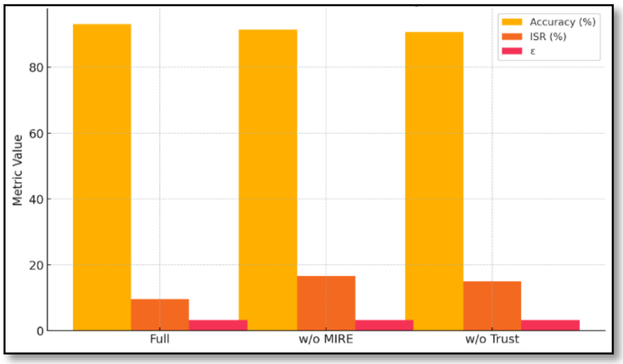

Fig. 9. Ablation Analysis of MIRE and Trust-Aware Aggregation Components in FedGuard-CI.

Figure 9 shows the impact of removing core FedGuard-CI modules. Accuracy drops and ISR increase significantly when MIRE or trust-aware aggregation is excluded, confirming that both components are critical to achieving privacy resilience without sacrificing model performance.

D. Runtime and Overhead in Ablation Variants

Table XII the overhead of different ablation settings for FedGuard-CI: computational and communication. It compares scenarios in which MIRE or trust-aware aggregation has been removed with anall time scenario. The results show that such removals do slightly decrease runtime and communication costs, indicating that only minimal overhead from additional modules is introduced. This does however underline just how efficient the whole FedGuard-CI architecture remains even given its considerable trust and security enhancements.

TABLE XII. ABLATION OVERHEAD

Configuration	Time/Round (s)	Comm. (KB)
FedGuard-CI (full)	30.2	585
Without MIRE	28.9	570
Without Trust Aggregation	29.1	560

E. Overhead Analysis

We analyzed FedGuard-CI's computational and communication overhead in comparison with FLAME and SecureAgg-FL.

TABLE XIII. SYSTEM OVERHEAD COMPARISON.

Model	Time/Round (s)	Comm. (KB)	Memory Usage	GPU Load
FLAME	33.9	650	Medium	42%
SecureAgg-FL	31.5	780	High	38%
FedGuard-CI	30.2	585	Medium	41%

While privacy mechanisms typically introduce computational costs, Table XIII illustrates that FedGuard-CI maintains competitive efficiency. It offers the lowest communication overhead (585 KB per round) among the evaluated models while maintaining runtime (30.2s) on par with lighter schemes. The GPU load and memory usage remain within acceptable bounds, demonstrating that the proposed framework does not compromise system scalability or resource constraints key considerations for CI in edge and IoT environments.

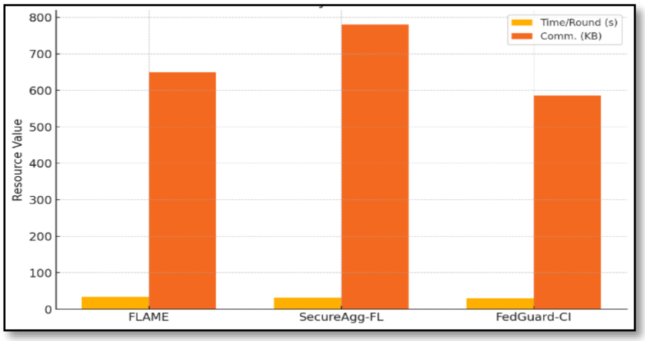

Fig. 10. System Overhead Comparison: Communication, Runtime, and Resource Utilization.

Figure 10 despite incorporating additional privacy layers, FedGuard-CI maintains low communication overhead (585 KB) and runtime (30.2 s). This figure reinforces that the system is efficient and deployable in constrained environments, such as edge-based CI networks.

F. Scalability and Cross-Domain Generalization

To evaluate cross-domain generalizability, FedGuard-CI was tested on MNIST, CIFAR-10, ChestX-ray14, and the UCI Loan Default datasets. As presented in Table XIV, the model consistently delivered high accuracy and low ISR across all domains. These results validate the scalability and robustness of FedGuard-CI for diverse federated environments.

TABLE XIV. CROSS-DOMAIN GENERALIZATION RESULTS.

Dataset	Accuracy (%)	ISR (%)	є\epsilon
MNIST	96.2	7.5	3.0
CIFAR-10	89.8	10.2	3.3
ChestX-ray14	91.4	8.6	3.1
UCI Loan Default	93.5	9.9	3.2

To validate generalizability, FedGuard-CI was evaluated across four datasets with varying characteristics. As shown in Table 12, the model achieved high accuracy and low ISR across all domains, including healthcare (ChestX-ray14) and finance (UCI Loan Default). The slight variation in performance is expected due to input complexity and domain sensitivity, yet FedGuard-CI consistently delivers strong privacy-preserving learning. This confirms its robustness and scalability in real-world, domain-diverse CI ecosystems.

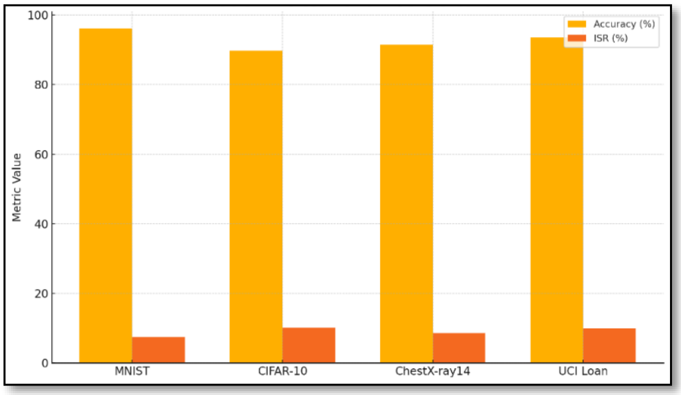

Fig. 11. Cross-Domain Generalization Performance of FedGuard-CI Across Diverse CI Datasets.

Figure 11 highlights FedGuard-CI's robustness across diverse datasets, including structured (UCI), visual (CIFAR-10, MNIST), and medical (ChestX-ray14) data. It consistently maintains high accuracy and low ISR, validating its scalability and adaptability across application domains.

G. Statistical Stability Across Runs

Table XIV shows the statistical variability of the major models by exhibiting the standard deviation of both accuracy and ISR. This displays how stable each model is in repeated races with distinct random seeds for training It has the lowest standard deviation in both accuracy and ISR of any federated learning algorithm currently existing, as the results show for FedGuard-CI.

TABLE XV. STATISTICAL MEASURES.

Model	Accuracy Std	ISR Std
FedAvg	±0.42	±1.10
DP-FedAvg	±0.38	±0.92
FLAME	±0.33	±0.75
FedGuard-CI	±0.29	±0.41

VIII. DISCUSSION

FedGuard-CI ensures a trade-off between privacy and utility in various domains. The proposed two-stage DP mechanism that we apply to DP-FedAVG exhibits robustness against inversion attacks, whose performance is affected by data dimension and non-IID distribution of datasets. Although trust-aware aggregation can enhance robustness against malicious clients, there could be biased opinions if the trust scores are improperly initialized when running in severely diverse client spectrum. MIRE successfully reduces leakage, but also introduces noise in risky stages, leading to small decreases (yet significant) on the utility. In addition, while the communication overhead is still reasonable, very bandwidth-limited IoT scenarios may demand further gradient compression or sparse exchange. These observations reveal promising strengths as well as limitations of deployment, encouraging us to consider extensions in terms of adaptive noise allocation, lightweight MIRE variants and self-correcting trust.

IX. Conclusion

FedGuard-CI presents a robust and efficient solution for securing collaborative intelligence systems, combining dual-stage differential privacy, trust-aware aggregation, and the MIRE module to effectively defend against model inversion and related privacy threats. The framework achieves strong privacy-utility trade-offs, low overhead, and broad scalability across diverse domains. Its architecture not only addresses current security challenges in federated learning but also offers a foundation for future extensions, including adaptive privacy mechanisms and threat-aware intelligence. As decentralised learning becomes integral to real-world applications, FedGuard-CI stands as a vital step toward secure, trustworthy, and privacy-preserving federated intelligence.

References

[1] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., et al., “Advances and open problems in federated learning,” Foundations and Trends® in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021. https://doi.org/10.1561/2200000083

[2] Geyer, R. C., Klein, T., and Nabi, M., “Differentially private federated learning: A client level perspective,” ArXiv, pp. 1–7, 2017. https://doi.org/10.48550/arXiv.1712.07557

[3] Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., and Zhou, Y., “A hybrid approach to privacy-preserving federated learning,” Informatik Spektrum, vol. 42, pp. 356–357, 2019. https://doi.org/10.1007/s00287-019-01205-x

[4] Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V., “Exploiting unintended feature leakage in collaborative learning,” In IEEE Symposium on Security and Privacy (SP), pp. 691–706, 2019. https://doi.org/10.1109/SP.2019.00029

[5] Bonawitz, K. A., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., et al., “Practical secure aggregation for privacy-preserving machine learning,” In CCS ’17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175–1191, 2017. https://doi.org/10.1145/3133956.3133982

[6] Nasr, M., Shokri, R., and Houmansadr, A., “Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning,” In IEEE Symposium on Security and Privacy (SP), pp. 739–753, 2019. https://doi.org/10.1109/SP.2019.00065

[7] Al-Kateb, G., “QIS-Box: Pioneering ultralightweight S-box generation with quantum inspiration,” Mesopotamian Journal of Cybersecurity, vol. 4, no. 2, pp. 106–119, 2024. https://doi.org/10.58496/MJCS/2024/010

[8] Chen, C., Liu, J., Tan, H., Li, X., Wang, K. I., Li, P., Sakurai, K., and Dou, D., “Trustworthy federated learning: Privacy, security, and beyond,” Knowledge and Information Systems, vol. 67, pp. 2321–2356, 2025. https://doi.org/10.1007/s10115-024-02285-2

[9] Han, Q., Lu, S., Wang, W., Qu, H., Li, J., and Gao, Y., “Privacy preserving and secure robust federated learning: A survey,” Concurrency and Computation: Practice and Experience, vol. 36, no. 13, e8084, 2024. https://doi.org/10.1002/cpe.8084

[10] Manzoor, H. U., Shabbir, A., Chen, A., Flynn, D., and Zoha, A., “A survey of security strategies in federated learning: Defending models, data, and privacy,” Future Internet, vol. 16, no. 10, p. 374, 2024. https://doi.org/10.3390/fi16100374

[11] Bouacida, N., and Mohapatra, P., “Vulnerabilities in federated learning,” IEEE Access, vol. 9, pp. 63229–63249, 2021. https://doi.org/10.1109/ACCESS.2021.3075203

[12] Erdal, Ş., Karakoç, F., and Özdemir, E., “A survey on security and privacy aspects and solutions for federated learning in mobile communication networks,” ITU Journal of Wireless Communications and Cybersecurity, vol. 1, no. 1, pp. 29–40, 2024.

[13] Hasan, J., “Security and privacy issues of federated learning,” ArXiv, pp. 1–6, 2023. https://doi.org/10.48550/arXiv.2307.12181

[14] Zhang, Y., Zeng, D., Luo, J., Xu, Z., and King, I., “A survey of trustworthy federated learning with perspectives on security, robustness and privacy,” In WWW ’23 Companion: Companion Proceedings of the ACM Web Conference 2023, pp. 1167–1176, 2023. https://doi.org/10.1145/3543873.3587681

[15] Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q., and Yu, P. S., “Privacy and robustness in federated learning: Attacks and defenses,” IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 7, pp. 8726–8746, 2022. https://doi.org/10.1109/TNNLS.2022.3216981

[16] Lyu, L., Yu, H., and Yang, Q., “Threats to federated learning: A survey,” ArXiv, pp. 1–7, 2020. https://doi.org/10.48550/arXiv.2003.02133

[17] Zhang, J., Zhu, H., Wang, F., Zhao, J., Xu, Q., and Li, H., “Security and privacy threats to federated learning: Issues, methods, and challenges,” Security and Communication Networks, vol. 2022, Art. ID 2886795, 2022. https://doi.org/10.1155/2022/2886795

[18] Lyu, L., Yu, H., Zhao, J., and Yang, Q., “Threats to federated learning,” In Federated Learning, pp. 3–16, 2020. https://doi.org/10.1007/978-3-030-63076-8_1

[19] Myakala, P. K., Bura, C., and Jonnalagadda, A. K., “Federated learning and data privacy: A review of challenges and opportunities,” International Journal of Research Publication and Reviews, vol. 5, no. 12, pp. 1867–1879, 2024.

[20] Aggarwal, M., Khullar, V., and Goyal, N., “A comprehensive review of federated learning: Methods, applications, and challenges in privacy-preserving collaborative model training,” In Applied Data Science and Smart Systems, pp. 570–575, 2024. https://doi.org/10.1201/9781003471059-73

[21] Li, Y., Guo, Z., Yang, N., Chen, H., Yuan, D., and Ding, W., “Threats and defenses in the federated learning life cycle: A comprehensive survey and challenges,” IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 9, pp. 15643–15663, 2025. https://doi.org/10.1109/TNNLS.2025.3563537

[22] Neto, H. N., Hribar, J., Dusparic, I., Mattos, D. M., and Fernandes, N. C., “A survey on securing federated learning: Analysis of applications, attacks, challenges, and trends,” IEEE Access, vol. 11, pp. 41928–41953, 2023. https://doi.org/10.1109/ACCESS.2023.3269980

[23] Zhou, H., Zheng, Y., and Jia, X., “Towards robust and privacy-preserving federated learning in edge computing,” Computer Networks, vol. 243, 110321, 2024. https://doi.org/10.1016/j.comnet.2024.110321

[24] Orabi, M. M., Emam, O., and Fahmy, H., “Adapting security and decentralized knowledge enhancement in federated learning using blockchain technology: Literature review,” Journal of Big Data, vol. 12, no. 55, pp. 1–24, 2025. https://doi.org/10.1186/s40537-025-01099-5

[25] Yurdem, B., Kuzlu, M., Gullu, M. K., Catak, F. O., and Tabassum, M., “Federated learning: Overview, strategies, applications, tools and future directions,” Heliyon, vol. 10, no. 19, e38137, 2024. https://doi.org/10.1016/j.heliyon.2024.e38137

[26] Shao, J., Li, Z., Sun, W., Zhou, T., Sun, Y., Liu, L., Lin, Z., Mao, Y., and Zhang, J., “A survey of what to share in federated learning: Perspectives on model utility, privacy leakage, and communication efficiency,” ArXiv, pp. 1–30, 2023. https://doi.org/10.48550/arXiv.2307.10655

[27] Hallaji, E., Razavi-Far, R., Saif, M., Wang, B., and Yang, Q., “Decentralized federated learning: A survey on security and privacy,” IEEE Transactions on Big Data, vol. 10, no. 2, pp. 194–213, 2024. https://doi.org/10.1109/TBDATA.2024.3362191

[28] Gholami, A., Torkzaban, N., and Baras, J. S., “Trusted decentralized federated learning,” In Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC), pp. 1–6, 2022. https://doi.org/10.1109/CCNC49033.2022.9700624

[29] Sala, E., “Descriptive analysis of cybersecurity awareness among smartphone users in higher education,” Journal of Transactions in Systems Engineering, vol. 2, no. 2, pp. 222–234, 2024. https://doi.org/10.15157/JTSE.2024.2.2.222-234

[30] Pashaj, K., and Gjika, E., “Mapping ISP malware trends in Albania: Clustering for smarter cyber defences,” International Journal of Innovative Technology and Interdisciplinary Sciences, vol. 8, no. 3, pp. 374–387, 2025. https://doi.org/10.15157/IJITIS.2025.8.3.374-387

[31] Ali, G., Samuel, A., Mijwil, M. M., Al-Mahzoum, K., Sallam, M., Salau, A. O., Bala, I., Dhoska, K., and Melekoglu, E., "Enhancing cybersecurity in smart education with deep learning and computer vision: A survey," Mesopotamian Journal of Computer Science, vol. 2025, pp. 115–158, 2025. https://doi.org/10.58496/MJCSC/2025/008

[32] Zhu, X., and Li, H., "Privacy-preserving poisoning-resistant blockchain-based federated learning for data sharing in the Internet of Medical Things," Applied Sciences, vol. 15, no. 10, p. 5472, 2025. https://doi.org/10.3390/app15105472

[33] S. Gao, J. Joshi, C. Li, J. Li, and R. Xu, "Dual Defense: Enhancing Privacy and Mitigating Poisoning Attacks in Federated Learning," in Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, Canada, Dec. 10–15, 2024, pp. 70476–70498. https://doi: 10.52202/079017-2253.

[34] P. Mai, Y. Pang, and R. Yan, "RFLPA: A Robust Federated Learning Framework against Poisoning Attacks with Secure Aggregation," in Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, Canada, Dec. 10–15, 2024, pp. 104329–104356. https://doi: 10.52202/079017-3314

[35] Y. Xie, M. Fang, and N. Gong, "FedREDefense: Defending against Model Poisoning Attacks for Federated Learning using Model Update Reconstruction Error," in Proceedings of the 41st International Conference on Machine Learning (ICML 2024), 2024, https://doi: 10.5555/3692070.3694309