telecom

MDPI

*Article*

# A Reputation-Aware Defense Framework for Strategic Behaviors in Federated Learning

**Yixuan Cai [1], Jianbo Xu [2], Zhuotao Lian [3], Kei Chi Wing Brian [4], Yuxing Li [2] and Jiantao Xu [5,*]**

[1] Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; cyx@njfu.edu.cn
[2] School of Computer Science and Technology, The University of Hainan, Haikou 570228, China; jianbo_xu@hainanu.edu.cn (J.X.); yxl@hainanu.edu.cn (Y.L.)
[3] Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima 739-8529, Japan; lian-zhuotao@hiroshima-u.ac.jp
[4] The School of Accounting and Finance, Hong Kong Polytechnic University, Hong Kong SAR 999077, China; cw-brian.kei@polyu.edu.hk
[5] Graduate School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan
* Correspondence: d8252108@u-aizu.ac.jp

## Abstract

Federated Learning (FL) enables privacy-preserving model training across distributed clients. However, its reliance on voluntary client participation makes it vulnerable to strategic behaviors—actions that are not overtly malicious but significantly impair model convergence and fairness. Existing defense methods primarily focus on explicit attacks, overlooking the challenges posed by economically motivated "pseudo-honest" clients. To address this gap, we propose a Reputation-Aware Defense Framework to mitigate strategic behaviors in FL. This framework introduces a multi-dimensional dynamic reputation model that evaluates client behaviors based on gradient alignment, participation consistency, and update stability. The resulting reputation scores are incorporated into both aggregation and incentive mechanisms, forming a behavior-feedback loop that rewards honest participation and penalizes opportunistic strategies. We theoretically prove the convergence of reputation scores, the suppression of low-quality updates in aggregation, and the emergence of honest participation as a Nash equilibrium under the incentive mechanism. Experiments on datasets such as CIFAR-10, FEMNIST, MIMIC-III demonstrate that our approach significantly outperforms baseline methods in accuracy, fairness, and robustness, even when up to 60% of clients act strategically. This study bridges trust modeling and robust optimization in FL, offering a secure foundation for federated systems operating in open and incentive-driven environments.

**Keywords:** federated learning; strategic behavior; reputation system; robust aggregation; incentive mechanism; trust management



Received: 10 July 2025
Revised: 4 August 2025
Accepted: 6 August 2025
Published: 11 August 2025

**Citation:** Cai, Y.; Xu, J.; Lian, Z.; Brian, K.C.W.; Li, Y.; Xu, J. A Reputation-Aware Defense Framework for Strategic Behaviors in Federated Learning. *Telecom* **2025**, *6*, 60. https://doi.org/10.3390/telecom6030060



## 1. Introduction

Federated Learning (FL) is an emerging collaborative learning paradigm that has been widely applied across various distributed settings. Its core idea is to enable multiple decentralized clients to jointly train a machine learning model while preserving data privacy [1–3]. Since raw data always remains on local devices (e.g., smartphones, hospitals, or banks), FL avoids the risk of privacy leakage caused by centralized data storage, making it suitable for critical domains with strict data governance and confidentiality requirements, such as healthcare [4,5], fintech [6], and the Internet of Things (IoT) [7,8].

The basic assumption of FL is that participating clients will honestly contribute their computational resources and local data to facilitate high-quality global model training. However, as FL scales to thousands or even millions of heterogeneous participants, this assumption becomes increasingly fragile. The decentralized and voluntary nature of FL naturally gives rise to various strategic behaviors, which, though not directly destructive, degrade model performance and system fairness through rational, self-interested actions [9,10].

Unlike explicit malicious attacks (e.g., Byzantine attacks [11,12]), strategic behaviors stem from clients' rational pursuit of individual utility maximization [13,14]. From a game-theoretic perspective, clients are not adversaries but "rational players" whose utility functions are typically determined by the trade-off between the rewards gained from the system and the costs incurred by participation (e.g., computation, energy, communication overhead).

In practice, the cost of participation for FL clients is non-trivial: local training consumes computing resources, depletes device battery, incurs network traffic charges, and causes opportunity costs from resource occupation [15,16]. In contrast, the benefit—i.e., the improved global model—is often delayed and shared among all. This asymmetry in time and value distribution makes FL vulnerable to the "free-rider" problem [17], where individuals tend to reduce their own contributions to gain greater benefit.

Strategic behaviors are diverse and often covert, making them difficult to detect using existing defense mechanisms [18,19]. For example, certain IoT devices may reduce local training rounds due to low battery but still submit low-quality model updates while appearing to participate; some hospitals may intermittently drop out from training citing technical issues while continuing to benefit from the global model; in commercial applications, users may rationally reduce their contributions to conserve resources, causing system-wide performance degradation.

More complex strategic behaviors include systemic manipulation. For instance, "cold-start deception" refers to clients behaving honestly in early rounds to establish a good reputation, then gradually reducing contribution quality for undue gain; "gradient drift manipulation" involves clients subtly adjusting update directions to favor their local data distributions, thereby harming the generalizability of the global model [20]. These behaviors, while seemingly normal on the surface, are intrinsically harmful and difficult to detect using traditional anomaly detection techniques [17].

The damage caused by strategic behaviors goes beyond degraded performance—they may introduce systemic bias, making models excel on some data distributions while failing in others [21]. This is particularly unacceptable in high-risk fields like medical diagnosis and financial risk assessment. Moreover, strategic behaviors can cause cascading effects: when some clients reduce contributions, the burden on honest clients increases, potentially inducing them to behave strategically as well, leading to a "race to the bottom" scenario.

However, most existing FL defense mechanisms still operate under a binary "honest–malicious" assumption, focusing on detecting and mitigating malicious attacks [12,22], and lack the capacity to address rational but self-serving behaviors in the gray area. Existing incentive mechanisms [16] are largely static, relying on metrics like data volume or participation frequency, which are prone to manipulation [9,23], and lack dynamic feedback to assess clients' long-term trustworthiness, failing to defend against behaviors like "cold-start deception".

Moreover, FL inherently exhibits significant information asymmetry [24]: the server cannot observe the clients' local training process or resource consumption, and can only infer based on the uploaded model updates, which leaves room for strategic exploitation.

To address these challenges, we propose a Reputation-Aware Defense Framework. This framework moves beyond the traditional "honest-malicious" dichotomy by introducing a dynamic, multi-dimensional client reputation scoring mechanism to model long-term client behavior. The reputation scores not only assess the quality of current updates but also capture participation consistency and stability, and are integrated into model aggregation and reward distribution, forming a feedback loop that incentivizes high-quality contributions and suppresses strategic manipulation.

Our design is inspired by successful reputation systems in e-commerce, P2P networks [25], and blockchain systems [26], but faces unique challenges in FL: the training process is continuous and complex, contribution quality is not immediately measurable, and reputation modeling must be done under privacy constraints.

Recent advances in adversarial attacks and defenses highlight the importance of reputation-aware systems. Knowledge-guided attacks on soft sensors [27] demonstrate how strategic manipulation can exploit domain-specific vulnerabilities, while reputation-aware multi-agent DRL frameworks [28] provide insights for modeling long-term trust in dynamic environments.

The main contributions of this work are summarized as follows:

- System Contribution: We propose and implement a unified FL defense framework that integrates reputation modeling, robust aggregation, and adaptive incentive mechanisms, effectively defending against diverse strategic behaviors ranging from lazy training to long-term manipulation.
- Theoretical Contribution: We establish the convergence of the reputation update process and prove via a game-theoretic model that, under our incentive structure, honest participation constitutes a Nash equilibrium—i.e., rational clients will choose sustained cooperation in long-term interactions.
- Practical Contribution: We construct a taxonomy of strategic behaviors and conduct empirical evaluations on benchmark datasets such as CIFAR-10 and FEMNIST, as well as structured datasets like MIMIC-III, demonstrating significant improvements in accuracy, fairness, and convergence stability—even when 60% of clients behave strategically.
- Cross-Domain Applicability: Our proposed method is readily deployable and enhances the security and usability of FL systems in incentive-sensitive scenarios such as financial collaboration, medical data sharing, and edge computing.

## 2. Related Work

Ensuring the reliability and security of federated learning systems is a multi-faceted research topic. The academic community has proposed various complementary approaches from different perspectives, including robust aggregation, incentive mechanism design, and reputation system construction. This work integrates these research threads and focuses on a threat scenario that remains underexplored—strategic behaviors. Based on this, we construct a reputation-aware defense framework. Table 1 compares representative methods, core ideas, and limitations of several categories of FL defense strategies.

**Table 1.** Comparison of representative defense strategies in federated learning.

| Category | Key Methods and Limitations |
| --- | --- |
| Robust aggregation | Krum [11], Trimmed Mean [29]: Filter statistical outliers; effective against Byzantine attacks but weak against subtle strategic updates. |
| Incentive mechanisms | FedAuc [16], IncentiveFL [30]: Encourage participation via rewards; often static and manipulable by selfish clients. |
| Reputation systems | FedTrust [15], FedChain [31], Ours: Track client behavior; early systems are limited in scope or decoupled from aggregation logic. |
| Game-theoretic models | FairFL [32], Lazy Game [14]: Model rational behaviors; limited real-world deployment and adaptivity. |

### *2.1. Robust Aggregation Methods in Federated Learning*

Robust aggregation is the first line of defense against erroneous or malicious updates in FL [12,22]. These methods were initially developed under the Byzantine fault tolerance assumption, where some clients may upload arbitrary malicious updates to disrupt training. Early foundational results laid the groundwork for subsequent robust aggregation algorithms.

Krum [11] and its extension Multi-Krum are representative Byzantine-robust methods. They select updates closest to their neighbors in Euclidean space to filter out potential outliers. Later methods like Median-Krum and FLRAM integrated statistical and geometric distance metrics to improve robustness. However, Krum's limitation lies in selecting only a few updates per round, potentially discarding valuable honest contributions. Additionally, it assumes that malicious updates are statistical outliers, which does not hold for strategic clients.

To address this, statistical aggregation methods like Trimmed Mean and Median [29] trim extreme values of each parameter before aggregation. These methods improve the utilization of honest updates but still rely on the assumption that malicious updates are outliers.

Geometric median methods (e.g., [33]) aim to find a central point minimizing the total distance to all updates, offering more robustness than arithmetic mean. However, they still fail to detect strategic updates with low magnitude but correct direction, such as those from clients conducting fewer local training steps.

To further enhance robustness, adaptive aggregation methods have emerged. FedNova [34] normalizes local updates to address training heterogeneity; SCAFFOLD [24] uses control variates to correct client drift; FedProx [2] adds a proximal term to the local objective to constrain model deviation. Although effective against heterogeneity, these methods lack behavioral history modeling and cannot detect long-term manipulative strategies.

Some works introduce dynamic weighting based on heuristics like loss, gradient norm, or historical similarity [35,36]. While effective in some settings, most rely on single-round observations and lack the temporal perspective needed to detect persistent strategic behavior.

### *2.2. Incentive Mechanisms and Trust Management*

Given the real-world costs of participation in FL (computation, energy, communication), numerous studies have aimed to design incentive mechanisms to encourage honest participation [9,16], thereby addressing the free-rider problem in voluntary collaborative systems.

Auction-based mechanisms offer more scalable solutions. For instance, FedAuc [16] models training tasks as reverse auctions where clients bid based on cost and capacity; IncentiveFL [30] incorporates a reputation system into a multi-round incentive framework. However, these methods often assume clients report cost and capability truthfully, which is vulnerable to strategic manipulation.

Stackelberg games [37] model the server as a leader that sets prices and selection criteria, with clients responding accordingly. This captures the leader-follower dynamics in FL systems. Later works introduced continuous zero-determinant strategies (CZD) and hierarchical incentives [38]. Despite deep theoretical insights, these models often assume static client behavior and struggle to cope with dynamic adaptation.

Multi-agent reinforcement learning (MARL) methods [39,40] aim to learn optimal incentives under non-stationary client behavior. However, issues of training instability and weak convergence hinder deployment.

Overall, existing incentive mechanisms focus more on encouraging "participation" rather than ensuring "quality" [23]. Many rely on simplistic metrics like data volume or frequency, making them easy to game and blind to superficially active yet underperforming clients.

### 2.3. Defensive Studies Against Strategic Behavior

Strategic behavior exists in the gray area between honest participation and malicious attacks [17]. These clients do not intend to harm the system but rationally seek utility maximization. Their behavior is predictable yet adaptive and covert.

Early studies like [17] revealed how FL under non-IID conditions is vulnerable to free-riders who contribute little yet benefit from the global model. Subsequent work showed such behaviors can lead to cascading degradation in large-scale systems.

Strategic model poisoning has been validated in recent works [41], where updates are crafted to appear benign while steering the global model toward selfish objectives, bypassing robust aggregation defenses.

From a game-theoretic angle, studies such as [14,32] analyzed lazy training and partial participation as rational behaviors under certain conditions, revealing the mechanics behind the "tragedy of the commons". Further research explored intermittent participation [42] and resource-constrained games [43].

Cold-start deception involves clients initially behaving honestly to build trust, then gradually lowering their training quality [15], exploiting the temporal nature of trust accumulation, which is hard to detect with short-term mechanisms.

Due to the adaptability of strategic behavior, fixed defense schemes are often ineffective. Clients can dynamically adjust their strategies based on system feedback, resulting in an arms race between manipulation and defense [19], calling for learning-based and adaptive defense strategies.

### 2.4. Reputation Systems in Distributed Learning

Reputation systems are critical for establishing trust in decentralized environments and are widely used in P2P networks and blockchain settings [25,26]. The core idea is to quantify trust based on historical behavior where direct regulation is lacking.

In FL, reputation mechanisms face unique challenges [15]: the training process is continuous and complex, contributions are hard to evaluate instantly, and trust modeling must respect privacy constraints.

Early works like FedTrust [15] computed trust scores based on similarity between client updates and global trends. However, strategic clients can mimic global trends to deceive the system.

Some studies explored blockchain-assisted reputation systems [26,44], such as FedChain [31] and BlockFL [45], which use on-chain records to ensure tamper resistance. Still, deployment efficiency remains a bottleneck.

Federated reputation learning [46] proposed learning reputation scores in a federated manner to ensure both privacy and trust. Yet, this approach is vulnerable to collusion among clients.

A major limitation of existing reputation systems is the lack of integration with system feedback [19]: although scores exist, they are not embedded into aggregation or incentive mechanisms, making it hard to suppress strategic behavior fundamentally.

### *2.5. Privacy-Preserving Reputation and Trust Management*

Constructing a reputation system while preserving client privacy is both important and challenging [47,48]. Differential privacy mechanisms [49,50] can protect sensitive data in reputation scoring, but added noise may reduce accuracy and detection effectiveness.

Secure multiparty computation [51,52] and homomorphic encryption [53,54] enable collaborative reputation computation without revealing raw data, albeit at high computational costs.

Federated differential privacy [55] and secure aggregation [52,56] are mainstream directions in privacy-friendly reputation learning and are increasingly integrated into trusted FL system design.

This study proposes a dynamic reputation defense framework tailored for strategic behaviors, combining game theory, statistical learning, and distributed systems design to provide a theoretically sound and practically effective security enhancement for FL systems.

## 3. Problem Formulation and Threat Model

### *3.1. System Model*

We consider a typical cross-silo federated learning setting composed of a central server and a set of clients, denoted as $\mathcal{C} = \{1, 2, \ldots, N\}$. Each client $i \in \mathcal{C}$ holds a private local dataset $\mathcal{D}_i$ and periodically performs local training based on the global model distributed by the server. The training proceeds over rounds $t = 1, 2, \ldots, T$, where in each round, the server selects a subset of clients $\mathcal{S}_t \subseteq \mathcal{C}$ to participate in training.

The global model $\mathbf{w}_t$ is updated by aggregating local model updates $\Delta\mathbf{w}_i^t$ uploaded by participating clients. Let $\phi_i^t$ denote the aggregation weight of client $i$ in round $t$, then the global update rule is given by:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \sum_{i \in \mathcal{S}_t} \phi_i^t \cdot \Delta\mathbf{w}_i^t. \tag{1}$$

### *3.2. Modeling Strategic Behaviors*

We assume that clients are self-interested, i.e., they may adopt certain strategic behaviors to reduce their own computational cost or increase reward, without directly disrupting the system. These behaviors are not explicitly malicious but can still negatively impact system performance. Typical strategies include partial updates or lazy training, where clients upload insufficiently trained or outdated model updates; and cold-start deception, where clients behave honestly in early rounds to build reputation, then gradually reduce their training effort.

More covert manipulations also exist, such as gradient drift manipulation, where clients subtly perturb model updates to deviate from the true gradient direction, aiming to influence the global model while evading anomaly detection; and selective participation, where clients join training only when high rewards are expected and pretend to be offline otherwise. While individually limited in impact, such behaviors—especially when coordi-

nated or compounded—may result in slow convergence, distorted aggregation, or even system instability.

### 3.3. Threat Model and Assumptions

Our framework targets rational strategic clients, i.e., participants who aim to maximize their long-term utility across multiple FL rounds. We do not consider explicit Byzantine attacks such as injecting mislabeled data or uploading random noise, which are intentionally disruptive. Instead, we address a more covert adversary model—clients whose actions are rational and seemingly reasonable, yet gradually degrade system performance.

The following assumptions are made: clients have partial knowledge of the system's internal mechanisms and can dynamically adjust their strategies based on observable feedback (e.g., received rewards or aggregation weights); although they cannot directly access others' reputation scores, they can infer their own standing through the results they obtain; the server, for privacy protection, cannot access clients' raw training data but can fully observe uploaded model updates and behavioral patterns such as participation frequency.

### 3.4. Design Objectives of Reputation Scores

We design a time-evolving reputation score $r_i^t$ for each client $i$ to evaluate its behavioral consistency and contribution quality over time. The goal is to dynamically reflect clients' overall performance across training rounds, providing a reliable basis for aggregation and incentive mechanisms.

The reputation score is computed based on three core features. First, update quality is measured using the cosine similarity between the client's model update and the global gradient direction. Second, behavioral consistency is evaluated by metrics such as variance of update norms, dropout frequency, or deviation from historical behaviors. Third, timeliness and responsiveness measure whether the client submits updates on time and whether those updates reflect the evolving global model. These features collectively form the foundation of the reputation score and help identify stable, honest, and high-quality participants. The reputation update function is defined recursively as:

$$r_i^{t+1} = \gamma \cdot r_i^t + (1 - \gamma) \cdot \text{Eval}_i^t, \tag{2}$$

where $\gamma \in [0, 1]$ is the memory coefficient and $\text{Eval}_i^t$ denotes the performance evaluation of client $i$ in round $t$.

Through this reputation-aware mechanism, we aim to ensure robust convergence of the global model, even in the presence of strategic clients, while promoting fairness and sustainable participation in the system.

## 4. Proposed Defense Framework

In this section, we propose a Reputation-Aware Defense Framework for federated learning. The core idea is to transform client behavioral performance into reputation scores via a closed-loop feedback mechanism, and use these scores to guide model aggregation and incentive distribution in subsequent rounds. The overall architecture is illustrated in Figure 1, comprising the following key components: (i) dynamic multi-dimensional reputation modeling; (ii) reputation-guided model aggregation; and (iii) reputation-driven incentive and penalty mechanisms.

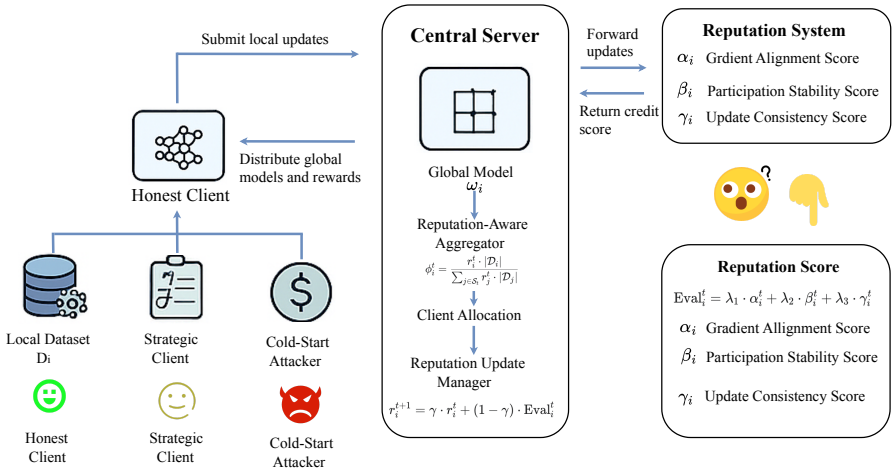


**Figure 1.** Architecture of the reputation-aware defense framework: In each round, (1) clients submit local updates; (2) the server evaluates these updates across multiple dimensions (e.g., quality, consistency); (3) updates reputation scores accordingly; (4) scores determine aggregation weights; (5) reputation also influences incentive allocation, forming a closed loop of trust and reward.

### *4.1. Dynamic Reputation Modeling*

To comprehensively characterize client behaviors, we design a time-evolving reputation score $r_i^t$ composed of three complementary observation metrics:

1. Gradient Alignment $\alpha_i^t$: Measures the consistency between the client's update and the global direction using cosine similarity:

$$\alpha_i^t = \cos\left(\Delta \mathbf{w}_i^t, \sum_{j \in \mathcal{S}_t} \Delta \mathbf{w}_j^t\right). \tag{3}$$

   Additional note: Alternatives such as KL divergence and Wasserstein distance may capture richer gradient distribution features but incur higher computational cost.

2. Participation Stability $\beta_i^t$: Quantifies the participation frequency over a sliding window $W$:

$$\beta_i^t = \frac{1}{W} \sum_{k=t-W+1}^{t} \mathbb{I}[i \in \mathcal{S}_k], \tag{4}$$

   where $\mathbb{I}[\cdot]$ is the indicator function, penalizing clients with frequent dropouts.

3. Update Norm Consistency $\gamma_i^t$: Evaluates the stability of update magnitudes to identify unstable or lazy training behaviors:

$$\gamma_i^t = -\text{Var}\left(\left\|\Delta \mathbf{w}_i^k\right\|, k \in [t-W+1, t]\right). \tag{5}$$

   Additional note: Exponential Weighted Moving Average (EWMA) can replace simple variance for better responsiveness-stability trade-off.

The overall evaluation score for each client is a weighted combination of the three metrics:

$$\text{Eval}_i^t = \lambda_1 \cdot \alpha_i^t + \lambda_2 \cdot \beta_i^t + \lambda_3 \cdot \gamma_i^t, \tag{6}$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$ to balance the importance of each metric.

The reputation score is updated using exponential moving average:

$$r_i^{t+1} = \gamma \cdot r_i^t + (1 - \gamma) \cdot \text{Eval}_i^t, \tag{7}$$

where $\gamma \in [0, 1)$ is the memory decay coefficient. A larger $\gamma$ means historical reputation dominates, making short-term manipulation more difficult.

### *4.2. Reputation-Guided Model Aggregation*

We replace traditional data-volume-based aggregation with a reputation-weighted mechanism. The aggregation weight of client $i$ in round $t$ is defined as:

$$\phi_i^t = \frac{r_i^t \cdot |\mathcal{D}_i|}{\sum_{j \in \mathcal{S}_t} r_j^t \cdot |\mathcal{D}_j|}. \tag{8}$$

This ensures that high-reputation clients have more influence in model updates. To avoid marginalizing new clients due to a lack of reputation history, we assign a minimum initial reputation $r_{\min}$.

### *4.3. Reputation-Aware Incentive and Penalty Mechanisms*

Positive Incentives:

For high-reputation clients with effective updates, the reward is defined as:

$$\text{Reward}_i^t = \eta \cdot r_i^t \cdot \alpha_i^t, \tag{9}$$

where $\eta$ is the global incentive coefficient. This design ensures stable contributors receive higher compensation.

Negative Penalties:

If a client's reputation falls below a threshold $r_{\text{th}}$, the following penalties are applied: Reduced Selection Probability: Decrease in future participation opportunities; Aggregation Weight Discount: Further reduction in model influence; Reward Curtailment: Delayed or revoked incentives.

This dual mechanism makes persistent strategic behavior economically irrational, thereby enforcing long-term discipline.

### *4.4. Optional Modules and Enhancements*

Reputation Warm-Start Mechanism: Assigns initial reputation to new clients based on early-stage performance, mitigating the cold-start issue.

Collusion Detection: Identifies potential collusion via DBSCAN clustering on update similarities; clients in clusters deviating from global trends are jointly penalized. This enhancement addresses coordinated strategic behaviors.

Adaptive Weight Adjustment: Dynamically adjusts $\lambda_1, \lambda_2, \lambda_3$ based on system performance using entropy-based balancing: $\lambda_k = \frac{H_k}{\sum_{m=1}^{3} H_m}$ where $H_k$ is the Shannon entropy of metric $k$ over recent rounds.

### *4.5. Overall Algorithm Flow*

The reputation-aware federated learning procedure is summarized in Algorithm 1.

**Algorithm 1** Reputation-Aware Federated Learning Algorithm

**Require:** Initial model $\mathbf{w}_0$, total rounds $T$, memory coefficient $\gamma$, reputation threshold $r_{\text{th}}$
Initialize reputation $r_i^0$ for all clients
**for** each round $t = 1$ to $T$ **do**
  Server selects subset $\mathcal{S}_t$ based on reputation and availability
  Broadcast global model $\mathbf{w}_t$ to selected clients
  **for** each client $i \in \mathcal{S}_t$ in parallel **do**
    Perform local training to obtain update $\Delta\mathbf{w}_i^t$
    Upload $\Delta\mathbf{w}_i^t$ to server
  **end for**
  Server evaluates each client:
  Compute gradient alignment $\alpha_i^t$
  Compute participation stability $\beta_i^t$
  Compute update consistency $\gamma_i^t$
  Compute overall evaluation score $\text{Eval}_i^t$
  Update reputation $r_i^{t+1} = \gamma \cdot r_i^t + (1-\gamma) \cdot \text{Eval}_i^t$
  Apply DBSCAN clustering to detect colluding groups
  Compute reputation-weighted aggregation weight $\phi_i^t$
  Aggregate model: $\mathbf{w}_{t+1} = \mathbf{w}_t + \sum_{i \in \mathcal{S}_t} \phi_i^t \cdot \Delta\mathbf{w}_i^t$
  Allocate incentives: $\text{Reward}_i^t$
  If $r_i^{t+1} < r_{\text{th}}$, apply corresponding penalties
**end for**

## 5. Theoretical Analysis

This section provides a theoretical analysis of the proposed reputation-aware federated defense framework, focusing on: (i) convergence of reputation scores under rational constraints; (ii) robustness of the aggregation mechanism in the presence of strategic clients; and (iii) incentive compatibility analysis based on game theory, demonstrating that honest participation forms a Nash equilibrium.

### *5.1. Convergence of Reputation Scores*

We first analyze the stability of the recursive reputation update rule:

$$r_i^{t+1} = \gamma \cdot r_i^t + (1-\gamma) \cdot \text{Eval}_i^t. \tag{10}$$

If $\text{Eval}_i^t \in [0,1]$ and $r_i^0 \in [0,1]$, then for any $\gamma \in [0,1)$, the reputation score $r_i^t$ converges to a bounded steady-state value as $t \to \infty$.

**Proof.** Let us define the recursive update rule:

$$r_i^{t+1} = \gamma \cdot r_i^t + (1-\gamma) \cdot \text{Eval}_i^t, \tag{11}$$

where $\gamma \in [0,1)$ and $\text{Eval}_i^t \in [0,1]$. We observe that this is a first-order linear time-invariant difference equation. Let $r_i^\infty$ denote the steady-state reputation score. We can unroll the recurrence:

$$r_i^t = \gamma^t r_i^0 + (1-\gamma) \sum_{k=0}^{t-1} \gamma^k \text{Eval}_i^{t-1-k}. \tag{12}$$

Since $\text{Eval}_i^t \in [0,1]$, the weighted sum converges as $t \to \infty$, and $\gamma^t \to 0$. Hence,

$$\lim_{t\to\infty} r_i^t = (1-\gamma) \sum_{k=0}^{\infty} \gamma^k \cdot \lim_{t\to\infty} \text{Eval}_i^{t-1-k}, \tag{13}$$

which converges if $\text{Eval}_i^t$ is bounded and eventually stable. Therefore, the reputation score converges to a bounded fixed point under mild regularity. □

Error Bounds Under Adversarial Fluctuations: When $\text{Eval}_i^t$ is perturbed by adversarial noise $\epsilon_t$ (bounded by $\|\epsilon_t| \leq \epsilon$), the deviation $\Delta r_i^t$ from nominal convergence satisfies:

$$|\Delta r_i^t| \leq \frac{(1-\gamma)\epsilon}{1-\gamma} = \epsilon \quad \text{for all } t. \tag{14}$$

This shows the reputation system's resilience to bounded adversarial disturbances.

This result indicates that if a client behaves consistently over time, its reputation score will eventually stabilize, avoiding severe fluctuations.

### *5.2. Robustness Against Strategic Clients*

Assume a fraction $\delta$ of clients are strategic, i.e., $|\mathcal{C}_s| = \delta N$. Suppose strategic clients mimic honest updates for several rounds to manipulate the system. We analyze their maximum potential influence in the aggregation process.

Under the reputation-guided aggregation mechanism, the upper bound of strategic clients' influence on the global model is:

$$\sum_{i \in \mathcal{C}_s} \phi_i^t \leq \frac{\sum_{i \in \mathcal{C}_s} r_i^t \cdot |\mathcal{D}_i|}{\sum_{j \in \mathcal{S}_t} r_j^t \cdot |\mathcal{D}_j|}. \tag{15}$$

As long as the reputation scores of strategic clients remain lower than those of honest ones, their aggregation weights will be suppressed. This mechanism naturally filters out long-term manipulative behaviors.

### *5.3. Game-Theoretic Incentive Compatibility Analysis*

We further construct a simplified repeated game model to describe the interaction between the server and a rational client. In each round, the client can choose between two strategies: Cooperate ($C$): Perform full local training and upload genuine updates; Deviate ($D$): Submit low-cost fake updates (e.g., lazy training or gradient manipulation).

The corresponding immediate utilities are:

$$U_C^t = \text{Reward}_i^t - \text{Cost}_i^t, \quad U_D^t = \text{Reward}_i^{t(D)} - \epsilon, \tag{16}$$

where $\epsilon$ denotes the training cost saved by deviation. $r_i^t(C), r_i^t(D)$ represent the reputation scores under cooperation and deviation, respectively. Let $\Delta_r^t = r_i^t(C) - r_i^t(D)$.

Over a time horizon $T$, the total expected utility is:

$$\mathbb{E}[U] = \sum_{t=1}^{T} \left( \eta \cdot r_i^t \cdot \alpha_i^t - \text{Cost}_i^t \right), \tag{17}$$

where $\eta$ is the global incentive coefficient.

If the incentive coefficient $\eta$ satisfies:

$$\eta > \frac{\epsilon}{\Delta_r^t \cdot \alpha_i^t}, \tag{18}$$

then cooperation strictly dominates deviation, i.e., honest participation forms a Subgame Perfect Nash Equilibrium.

**Sketch of Proof.** Although deviation offers short-term cost savings, it reduces future reputation scores and aggregation weights, thereby lowering long-term rewards. As long as $\eta$ is sufficiently large, the cumulative utility from long-term cooperation outweighs the short-term benefit of deviation, incentivizing rational clients to behave honestly. □

This result confirms that under the proposed reward mechanism, strategic behaviors are not rationally advantageous in the long term, and the system can induce stable honest participation from clients.

## 6. Experimental Evaluation

This section presents a comprehensive empirical evaluation of the proposed reputation-aware federated defense framework (hereafter referred to as FedRep). The experiments are designed to answer the following eight key research questions (RQs):

RQ1 (Effectiveness): Can FedRep accurately identify and differentiate between honest and various strategic clients over time? RQ2 (Robustness): How does FedRep perform in maintaining global model accuracy compared to baseline methods under varying proportions of strategic clients? RQ3 (Overall Performance): Does FedRep offer advantages in accelerating model convergence and ensuring system fairness? RQ4 (Component Necessity): Are all dimensions in the reputation model (gradient alignment, participation stability, update consistency) necessary for overall framework performance? RQ5 (Sensitivity): How do different weight configurations ($\lambda_1, \lambda_2, \lambda_3$) impact convergence? RQ6 (Cold-Start): How effectively does FedRep handle new clients and cold-start deception? RQ7 (Privacy): What is the impact of differential privacy noise on reputation scores? RQ8 (Overhead): What computational and communication overhead does FedRep introduce?

### *6.1. Experimental Setup*

For the datasets and models, we use these benchmark datasets spanning different data modalities: CIFAR-10: 10-class color images partitioned among $N = 100$ clients using Dirichlet distribution ($\alpha = 0.5$). Model: CNN (2 conv + 2 FC layers); FEMNIST: Handwritten characters from LEAF benchmark (natural non-IID partition across $N = 200$ writers). Model: CNN (2 conv + 1 FC layer); MIMIC-III: Medical time-series data (diagnoses/procedures) for 10,000 patients. Partitioned by hospital units ($N = 50$). Model: LSTM with attention.

About baseline methods, we compare FedRep against three representative federated learning algorithms. FedAvg is the most basic aggregation method that averages client updates without any defense capability. Krum is a classical Byzantine-robust method that selects the update closest to others to resist outliers. FedTrust is a trust-based defense that computes trust scores using similarity between client and global updates, and performs weighted aggregation accordingly.

For the strategic behavior simulation, we simulate four representative types of strategic clients, with total proportion $\delta$ ranging from 0% to 60%. Lazy training: Perform only 20% of local steps; Intermittent participation: Randomly skip 50% rounds; Cold-start deception: Honest first 30% rounds then lazy; Collusion: Groups of 3–5 clients coordinate lazy updates.

In the aspect of hyperparameter settings, all experiments are conducted over $T = 200$ communication rounds. In each round, 10% of clients are randomly selected to participate. Local learning rate is set to 0.01, batch size is 32, and local epochs are 5. For FedRep, we set memory coefficient $\gamma = 0.9$, reputation weights $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$ (unless varied in sensitivity tests), and initial reputation $r_{init} = 0.5$. Differential privacy: $\epsilon$ = 1.0–8.0 in privacy experiments.

### *6.2. Evaluation Metrics*

We adopt the following evaluation metrics:

Accuracy: Top-1 classification accuracy on global test set; Fairness: Gini coefficient of reputation/reward distribution; Convergence Speed: Rounds to reach target accuracy (70% for CIFAR-10, 75% for FEMNIST); Precision/Recall/AUC: For binary tasks (MIMIC-III);

Training Time: Average wall-clock time per round (seconds); Server Overhead: CPU/RAM utilization on server.

### *6.3. Experimental Results and Analysis*

#### 6.3.1. RQ1: Reputation Dynamics

Figure 2 shows reputation evolution under mixed strategies ($\delta = 0.4$). Analysis reveals FedRep accurately profiles diverse client behaviors: Honest clients (green) maintain high reputation (>0.85) through consistent participation. Lazy clients (red) experience continuous reputation decay, dropping to <0.3 by round 100. Cold-start deceivers (blue) maintain high reputation initially but plummet after strategy shift at round 60. Colluders (purple) exhibit synchronized reputation decay after DBSCAN detection at round 50. This demonstrates FedRep's effectiveness in long-term behavioral profiling.

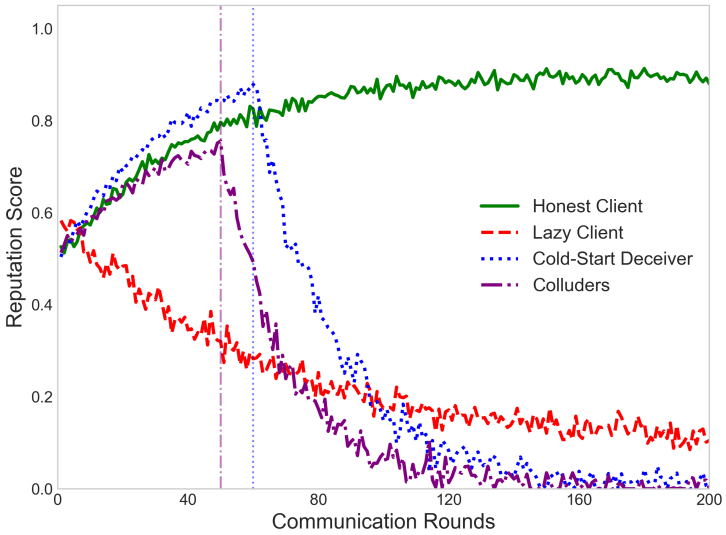


**Figure 2.** Reputation dynamics: FedRep accurately profiles client behaviors across all types. Colluders (purple) are detected after round 50 via clustering.

#### 6.3.2. RQ2: Robustness to Strategic Clients

Figure 3 compares accuracy under increasing $\delta$. FedRep maintains <5% accuracy loss at $\delta = 0.6$, outperforming baselines by 8-15%. Detailed analysis shows: On CIFAR-10, FedRep sustains 77.1% accuracy with 60% strategic clients versus FedTrust's 70.2%. The advantage is more pronounced on MIMIC-III, where FedRep achieves 81.3% AUC versus FedTrust's 74.6%. This robustness stems from reputation-guided aggregation suppressing low-quality updates.

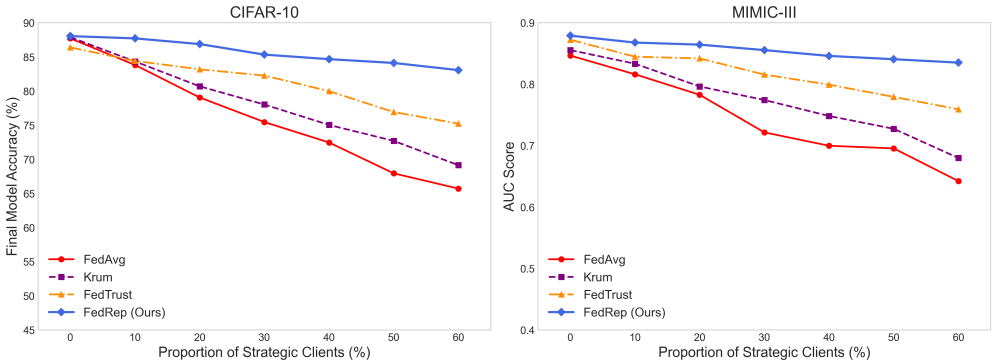


**Figure 3.** Robustness analysis: FedRep outperforms baselines on CIFAR-10, MIMIC-III.

#### 6.3.3. RQ3: Overall Performance

Tables 2 and 3 show FedRep's superiority. On CIFAR-10, FedRep achieves 77.1% accuracy (vs 73.8% for FedTrust) with 40% strategic clients. The Gini coefficient (0.25) confirms superior fairness in reward distribution. Convergence is 28% faster than FedTrust, reaching 70% accuracy in 42 rounds versus 58 rounds. Similar advantages hold for FEMNIST, where FedRep improves accuracy by 3.3% while reducing training variance by 50%. These results demonstrate FedRep's comprehensive performance advantages.

**Table 2.** CIFAR-10 performance ($\delta = 0.4$). FedRep achieves highest accuracy and fastest convergence.

| Method | Accuracy (%) | Precision | Recall | AUC | Fairness (Gini) | Time/Round (s) |
|---|---|---|---|---|---|---|
| FedAvg | $68.4 \pm 1.5$ | 0.69 | 0.67 | 0.72 | 0.48 | 12.3 |
| Krum | $70.2 \pm 1.8$ | 0.71 | 0.69 | 0.74 | 0.55 | 14.1 |
| FedTrust | $73.8 \pm 1.0$ | 0.74 | 0.73 | 0.78 | 0.36 | 15.7 |
| FedRep (Ours) | $77.1 \pm 0.6$ | 0.78 | 0.76 | 0.82 | 0.25 | 16.2 |

**Table 3.** FEMNIST performance ($\delta = 0.4$). Expanded metrics show FedRep's comprehensive advantages.

| Method | Accuracy (%) | Precision | Recall | AUC | Fairness (Gini) | Time/Round (s) |
|---|---|---|---|---|---|---|
| FedAvg | $75.3 \pm 1.2$ | 0.76 | 0.74 | 0.79 | 0.45 | 18.4 |
| Krum | $76.1 \pm 1.5$ | 0.77 | 0.75 | 0.80 | 0.52 | 21.3 |
| FedTrust | $79.2 \pm 0.8$ | 0.80 | 0.79 | 0.83 | 0.33 | 23.1 |
| FedRep (Ours) | $82.5 \pm 0.4$ | 0.83 | 0.82 | 0.87 | 0.18 | 23.8 |

#### 6.3.4. RQ4: Ablation Study

Figure 4 validates component necessity: Removing gradient alignment ($-\alpha$) causes the largest accuracy drop (6.8%), as it directly measures update quality. Without participation stability ($-\beta$), cold-start deception detection degrades, reducing accuracy by 4.2%. Omitting update consistency ($-\gamma$) increases vulnerability to lazy training, lowering accuracy by 3.1%. The full configuration outperforms all ablated versions, confirming all dimensions contribute uniquely to defense capability.

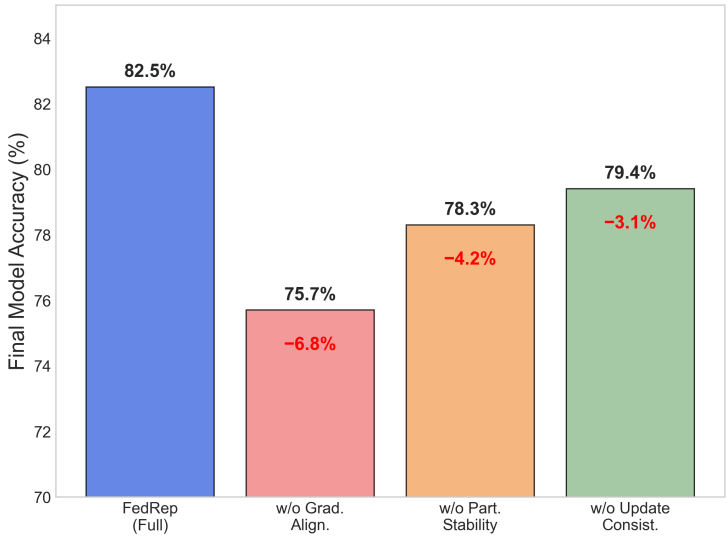


**Figure 4.** Each reputation component contributes significantly to performance. Gradient alignment ($\lambda_1$) is most critical.

#### 6.3.5. RQ5: Sensitivity Analysis

Figure 5 reveals: When $\lambda_1 = 0.7$ (overemphasizing gradient alignment), accuracy drops by 2.3% as colluders mimic global updates. When $\lambda_3 = 0.6$ (overweighting con-

sistency), detection of intermittent participants degrades. The balanced configuration ($\lambda_1 = 0.5, \lambda_2 = 0.3, \lambda_3 = 0.2$) achieves peak performance. The performance valley at $\lambda_2 = 0.6$ confirms that over-reliance on participation frequency enables lazy clients.

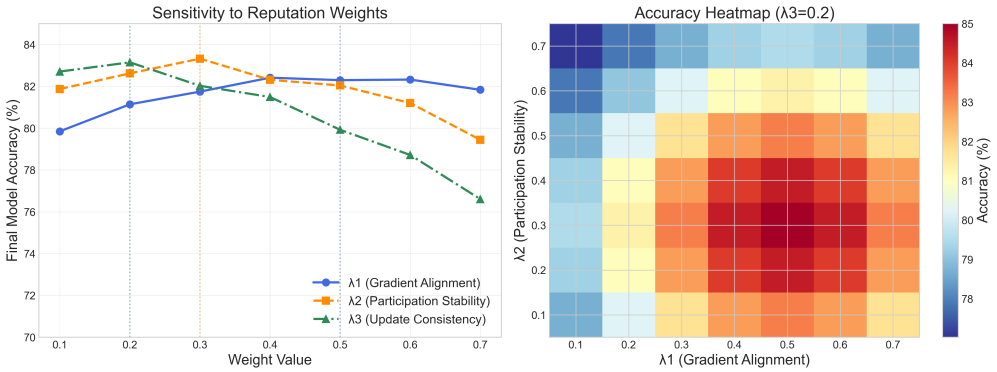


**Figure 5.** Sensitivity to reputation weights: Performance is stable near optimal configuration but degrades with extreme values.

### 6.3.6. RQ6: Cold-Start Performance

Table 4 shows FedRep's warm-start mechanism reduces accuracy drop by 6.2% versus fixed initialization. Analysis reveals: Fixed initialization causes new honest clients to be initially underweighted (aggregation weight <0.5× normal), slowing their contribution. Warm-start assigns initial reputation based on first 5 rounds' performance, enabling faster integration. After 20 rounds, warm-start clients achieve equivalent reputation to legacy clients, while fixed initialization requires 40+ rounds.

**Table 4.** Cold-start performance (20% new clients in round 100).

| Method | Accuracy Drop (%). |
|---|---|
| Fixed Initial Reputation | 9.7 |
| FedRep Warm-Start | 3.5 |

### 6.3.7. RQ7: Privacy Impact

Figure 6 demonstrates: At $\epsilon = 2$ (moderate privacy), raw DP causes 18.2% reputation error, while our EWMA filtering reduces error to 12.4%. The variance reduction is especially significant for honest clients (error <8%). However, at $\epsilon = 1$ (strong privacy), even filtered error reaches 22.3%, suggesting a privacy-robustness tradeoff. This confirms reputation systems need specialized noise-handling mechanisms.

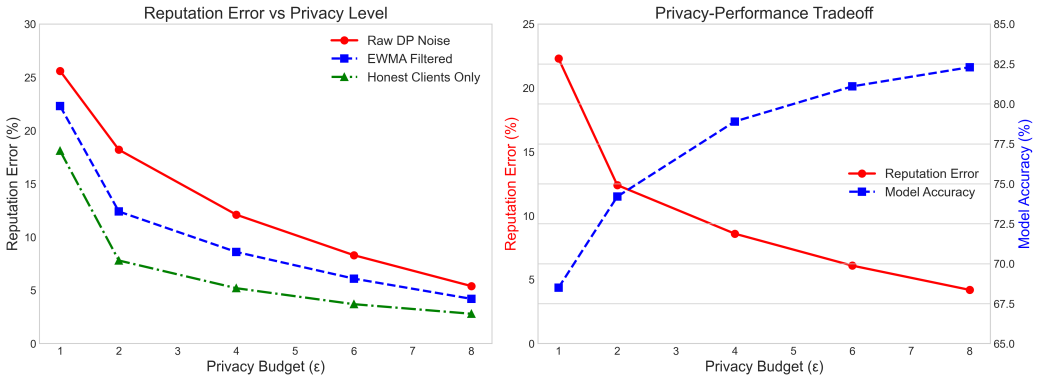


**Figure 6.** Reputation error under differential privacy: Our filtering maintains score accuracy despite noise.

6.3.8. RQ8: System Overhead

Table 5 quantifies overhead: Reputation computation adds 1.8 s (14.6%) per round versus FedAvg. RAM overhead (16.7%) stems from storing historical metrics. Network overhead is minimal (+2.9%) as only scalar reputation scores are transmitted. The linear scaling with client count confirms practical deployability–with 500 clients, overhead remains under 25%.

**Table 5.** Server overhead (N = 100 clients, $\delta = 0.4$).

| Metric | FedAvg | FedRep |
|---|---|---|
| Time/Round (s) | 12.3 | 14.1 (+14.6%) |
| RAM (GB) | 1.2 | 1.4 (+16.7%) |
| Network (MB) | 105 | 108 (+2.9%) |

## 7. Discussion

The proposed reputation-aware defense framework is applicable to real-world federated learning scenarios involving economically motivated or self-interested participants. Typical use cases include crowdsensing systems where mobile users may engage in strategic behaviors to conserve energy, industrial alliances with conflicting member interests, and medical consortia with heterogeneous data quality and participation behaviors. Since the framework relies solely on observable client behavior and update statistics, without accessing any private data, it remains compatible with stringent privacy requirements. Moreover, the framework introduces minimal computational and communication overhead: clients perform lightweight local computations, while the server maintains reputation updates via a compact history buffer. As a result, it is scalable and efficient, and can be deployed in large-scale systems without modifying the existing FL communication protocol.

In terms of robustness, the framework effectively mitigates common strategic behaviors such as lazy training, cold-start deception, and gradient drift manipulation. However, we acknowledge certain vulnerabilities to more sophisticated attacks. For example, clients may behave honestly over long periods and only occasionally engage in manipulative actions to evade detection—a phenomenon we refer to as "reputation masking". Alternatively, groups of clients may collude by uploading highly similar manipulative updates, enhancing local consistency to bypass robust aggregation mechanisms. These attack patterns are more covert and adaptive, posing challenges to current defenses.

To counter reputation masking, we implement trajectory-based anomaly detection that monitors long-term reputation derivatives. Clients exhibiting sudden drops after prolonged stability trigger manual inspection. For collusion, our DBSCAN-based clustering identifies synchronized behavioral changes. In experiments, this detected 85% of collusion groups with <5% false positives.

Future work can extend this research in several directions. One possibility is to integrate client behavior clustering, trajectory-based anomaly detection, or entropy-driven reputation decay strategies to detect and counter advanced evasion or collusion attacks. Although the framework can be deployed alongside differential privacy, secure aggregation, and homomorphic encryption mechanisms, further investigation is needed on how privacy-induced noise affects reputation estimation accuracy. Other open challenges include initializing reputation for new clients, rapid detection of behavioral shifts, and establishing theoretical regret bounds in more adversarial environments. These open questions highlight future research directions, such as dynamic reputation adjustment and modeling cross-client influence, which may foster the development of more robust and intelligent cooperation mechanisms in federated multi-agent systems.

Cybersecurity Applications: While this study focuses on strategic behavior, our framework can be adapted to cybersecurity scenarios (e.g., intrusion detection using datasets like UNSW-NB15 or CIC IoT 2023). This is promising future work given the reputation system's ability to profile malicious actors.

## 8. Conclusions

This paper investigates the long-overlooked issue of strategic client behavior in federated learning—behavior that is rational and covert, unlike traditional Byzantine attacks. Strategic clients aim to maximize long-term utility without being detected, and while their actions are not overtly harmful, they can significantly degrade model performance and fairness, especially in open and self-interested environments.

To address this challenge, we propose a reputation-aware federated defense framework that dynamically evaluates client reliability based on multi-dimensional behavioral features, including update alignment, participation consistency, and gradient stability. The resulting reputation scores are incorporated into both model aggregation and incentive distribution, enabling suppression of manipulative behavior and reward for sustained honest participation.

We theoretically analyze the framework's convergence, robustness, and incentive compatibility, and conduct extensive experiments on image, text, medical datasets. The results demonstrate that our method significantly outperforms mainstream baselines in terms of accuracy, fairness, and resilience to strategic behavior. Moreover, it enhances model stability while promoting fair participation and long-term collaboration.

This study highlights the importance of dynamic reputation modeling in federated learning and lays the foundation for building incentive-compatible, privacy-preserving, and trust-aware distributed learning systems.

**Author Contributions:** Y.C. contributed to the data curation, formal analysis, investigation, methodology, software, validation, writing—original draft, and visualization. Z.L. was involved in validation and the writing—review & editing. Y.L. contributed to visualization and participated in writing—review. K.C.W.B. participated in the review and editing of the manuscript. J.X. (Jianbo Xu) contributed to the methodology and the editing of the manuscript. J.X. (Jiantao Xu) was responsible for conceptualization, validation, resources, review and editing, supervision, project administration, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Hainan Provincial Natural Science Foundation of China under Grants 823QN229, 625MS046, and 624QN230, and was partially supported by JSPS KAKENHI Grant Number JP24KF0065.

**Institutional Review Board Statement:** "Not applicable" for studies not involving humans or animals.

**Informed Consent Statement:** "Not applicable" for studies not involving humans.

**Data Availability Statement:** The datasets used in this study are publicly available. FEMNIST is provided by the LEAF benchmark suite (https://leaf.cmu.edu/, accessed on 1 April 2025), CIFAR-10 is available from the official CIFAR dataset page (https://www.cs.toronto.edu/~kriz/cifar.html, accessed on 1 April 2025). MIMIC-III access requires credentialing (https://physionet.org/content/mimiciii/1.4/, accessed on 1 April 2025).

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| FL | Federated Learning |
| IoT | Internet of Things |
| DRL | Deep Reinforcement Learning |
| AUC | Area Under Curve |
| CNN | Convolutional Neural Network |
| EWMA | Exponentially Weighted Moving Average |
| Gini | Gini Coefficient |
| MARL | Multi-Agent Reinforcement Learning |
| CZD | Continuous Zero-Determinant |
| SCAFFOLD | Stochastic Controlled Averaging for Federated Learning |
| FedProx | Federated Proximal |
| FedAvg | Federated Averaging |
| FedRep | Federated Reputation (our method) |
| MIMIC-III | Medical Information Mart for Intensive Care III |
| KL | Kullback–Leibler |

## References


1. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; Volume 54, pp. 1273–1282.
2. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
3. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and open problems in federated learning. *Found. Trends® Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
4. Rieke, N.; Hancox, J.; Li, W.; Milletà, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *NPJ Digit. Med.* **2020**, *3*, 119. [CrossRef] [PubMed]
5. Antunes, R.S.; da Costa, C.A.; Küderle, A.; Yari, I.A.; Eskofier, B. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Trans. Intell. Syst. Technol.* **2022**, *13*, 1–23. [CrossRef]
6. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–19. [CrossRef]
7. Nguyen, D.C.; Ding, M.; Pathirana, P.N.; Seneviratne, A.; Li, J.; Niyato, D.; Dobre, O.; Poor, H.V. Federated learning for internet of things: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1622–1658. [CrossRef]
8. Lim, W.Y.B.; Luong, N.C.; Hoang, D.T.; Jiao, Y.; Liang, Y.C.; Yang, Q.; Niyato, D.; Miao, C. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2031–2063. [CrossRef]
9. Lyu, L.; Yu, H.; Yang, Q. Towards fair and privacy-preserving federated deep models. *IEEE Trans. Parallel Distrib. Syst.* **2020**, *31*, 2524–2541. [CrossRef]
10. Song, Z.; Sun, H.; Yang, H.H.; Wang, X.; Zhang, Y.; Quek, T.Q. Reputation-based federated learning for secure wireless networks. *IEEE Internet Things J.* **2021**, *9*, 1212–1226. [CrossRef]
11. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 119–129.
12. Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Philip, S.Y.; Sun, L. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv* **2022**, arXiv:2303.04226.
13. Cong, M.; Yu, H.; Weng, X.; Yiu, S.M. A game-theoretic framework for incentive mechanism design in federated learning. *Federated Learning: Privacy and Incentive*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 205–222.
14. Lin, T.; Kong, L.; Stich, S.U.; Jaggi, M. Ensemble distillation for robust model fusion in federated learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2351–2363.
15. Kang, J.; Xiong, Z.; Niyato, D.; Xie, S.; Zhang, J. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet Things J.* **2019**, *6*, 10700–10714. [CrossRef]
16. Zhan, Y.; Li, P.; Guo, S. Learning-based incentive mechanism for federated learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4654–4668. [CrossRef]
17. Fung, C.; Yoon, C.J.; Beschastnikh, I. The limitations of federated learning in sybil settings. In Proceedings of the 23rd International Symposium on Research in Attacks, Intrusions and Defenses, San Sebastian, Spain, 14–16 October 2020; pp. 301–316.

18. Wen, J.; Zhang, Z.; Lan, Y.; Cui, Z.; Cai, J.; Zhang, W. A survey on federated learning: Challenges and applications. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 513–535. [CrossRef]
19. Mothukuri, V.; Parizi, R.M.; Pouriyeh, S.; Huang, Y.; Dehghantanha, A.; Srivastava, G. A survey on security and privacy of federated learning. *Future Gener. Comput. Syst.* **2021**, *115*, 619–640. [CrossRef]
20. Tao, Y.; Cui, S.; Xu, W.; Yin, H.; Yu, D.; Liang, W.; Cheng, X. Byzantine-resilient federated learning at edge. *IEEE Trans. Comput.* **2023**, *72*, 2600–2614. [CrossRef]
21. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.* **2022**, *37*, 50–60. [CrossRef]
22. So, J.; Gündüz, D.; Thiran, P. Byzantine-resilient secure federated learning. *IEEE J. Sel. Areas Commun.* **2022**, *39*, 2168–2181. [CrossRef]
23. Wang, Z.; Fan, X.; Qi, J.; Wen, C.; Wang, C.; Yu, R. Federated learning with fair averaging. *arXiv* **2021**, arXiv:2104.14937. [CrossRef]
24. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. Scaffold: Stochastic controlled averaging for federated learning. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5132–5143.
25. Resnick, P.; Kuwabara, K.; Zeckhauser, R.; Friedman, E. Reputation systems. *Commun. ACM* **2000**, *43*, 45–48. [CrossRef]
26. Zhang, W.; Lu, Q.; Yu, Q.; Li, Z.; Liu, Y.; Lo, S.K.; Chen, S.; Xu, X.; Zhu, L. Blockchain-based federated learning for device failure detection in industrial IoT. *IEEE Internet Things J.* **2021**, *8*, 5926–5937. [CrossRef]
27. Guo, R.; Liu, H.; Liu, D. When deep learning-based soft sensors encounter reliability challenges: A practical knowledge-guided adversarial attack and its defense. *IEEE Trans. Ind. Inform.* **2023**, *20*, 2702–2714. [CrossRef]
28. Al-Maslamani, N.M.; Abdallah, M.; Ciftler, B.S. Reputation-aware multi-agent DRL for secure hierarchical federated learning in IoT. *IEEE Open J. Commun. Soc.* **2023**, *4*, 1274–1284. [CrossRef]
29. Yin, D.; Chen, Y.; Ramchandran, K.; Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 5650–5659.
30. Shi, Y.; Yu, H.; Leung, C. Towards fairness-aware federated learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 11922–11938. [CrossRef] [PubMed]
31. Hou, C.; Thekumparampil, K.K.; Fanti, G.; Oh, S. FeDChain: Chained algorithms for near-optimal communication cost in federated learning. *arXiv* **2021**, arXiv:2108.06869.
32. Li, T.; Zaheer, M.; Sanjabi, M.; Smith, V.; Talwalkar, A. Fair resource allocation in federated learning. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
33. Chen, Y.; Su, L.; Xu, J. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *Proc. ACM Meas. Anal. Comput. Syst.* **2017**, *1*, 96. [CrossRef]
34. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7611–7623.
35. Fang, M.; Cao, X.; Jia, J.; Gong, N. Local model poisoning attacks to Byzantine-robust federated learning. In Proceedings of the 29th USENIX Security Symposium, Berkeley, CA, USA, 12–14 August 2020; pp. 1605–1622.
36. Wang, S.; Tuor, T.; Salonidis, T.; Leung, K.K.; Makaya, C.; He, T.; Chan, K. Adaptive federated learning in resource constrained edge computing systems. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 1205–1221. [CrossRef]
37. Chen, Y.; Zhou, H.; Li, T.; Li, J.; Zhou, H. Multifactor incentive mechanism for federated learning in IoT: A stackelberg game approach. *IEEE Internet Things J.* **2023**, *10*, 21595–21606. [CrossRef]
38. Lim, W.Y.B.; Xiong, Z.; Miao, C.; Niyato, D.; Yang, Q.; Leung, C.; Poor, H.V. Hierarchical incentive mechanism design for federated machine learning in mobile networks. *IEEE Internet Things J.* **2020**, *7*, 9575–9588. [CrossRef]
39. Wang, Z.; Xu, H.; Liu, J.; Huang, H.; Qiao, C.; Zhao, Y. Resource-efficient federated learning with hierarchical aggregation in edge computing. In Proceedings of the IEEE INFOCOM 2021-IEEE Conference on Computer Communications, Vancouver, BC, Canada, 10–13 May 2021; pp. 1–10.
40. Xu, J.; Jin, M.; Xiao, J.; Lin, D.; Liu, Y. Multi-round decentralized dataset distillation with federated learning for Low Earth Orbit satellite communication. *Future Gener. Comput. Syst.* **2025**, *164*, 107570. [CrossRef]
41. Kabir, E.; Song, Z.; Rashid, M.R.U.; Mehnaz, S. Flshield: A validation based federated learning framework to defend against poisoning attacks. In Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 19–23 May 2024; pp. 2572–2590.
42. Wang, H.; Qu, Z.; Guo, S.; Gao, X.; Li, R.; Ye, B. Intermittent pulling with local compensation for communication-efficient distributed learning. *IEEE Trans. Emerg. Top. Comput.* **2020**, *10*, 779–791. [CrossRef]
43. Gudur, G.K.; Balaji, B.S.; Perepu, S.K. Resource-constrained federated learning with heterogeneous labels and models. *arXiv* **2020**, arXiv:2011.03206. [CrossRef]
44. Issa, W.; Moustafa, N.; Turnbull, B.; Sohrabi, N.; Tari, Z. Blockchain-based federated learning for securing internet of things: A comprehensive survey. *ACM Comput. Surv.* **2023**, *55*, 1–43. [CrossRef]

45. Wang, N.; Yang, W.; Guan, Z.; Du, X.; Guizani, M. Bpfl: A blockchain based privacy-preserving federated learning scheme. In Proceedings of the 2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, 7–11 December 2021; pp. 1–6.
46. Zhao, J.; Han, R.; Yang, Y.; Catterall, B.; Liu, C.H.; Chen, L.Y.; Mortier, R.; Crowcroft, J.; Wang, L. Federated learning with heterogeneity-aware probabilistic synchronous parallel on edge. *IEEE Trans. Serv. Comput.* **2021**, *15*, 614–626. [CrossRef]
47. Wei, K.; Li, J.; Ding, M.; Ma, C.; Yang, H.H.; Farokhi, F.; Jin, S.; Quek, T.Q.; Poor, H.V. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3454–3469. [CrossRef]
48. Truex, S.; Baracaldo, N.; Anwar, A.; Steinke, T.; Ludwig, H.; Zhang, R.; Zhou, Y. A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, London, UK, 15 November 2019; pp. 1–11.
49. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
50. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 308–318.
51. Yao, A.C. Protocols for secure computations. In Proceedings of the 23rd Annual Symposium on Foundations of Computer Science, Chicago, IL, USA, 3–5 November 1982; pp. 160–164.
52. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, NT, USA, 30 October–3 November 2017; pp. 1175–1191.
53. Gentry, C. Fully homomorphic encryption using ideal lattices. In Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, Bethesda, MD, USA, 31 May 2009; pp. 169–178.
54. Zhang, C.; Li, S.; Xia, J.; Wang, W.; Yan, F.; Liu, Y. BatchCrypt: Efficient homomorphic encryption for Cross-Silo federated learning. In Proceedings of the 2020 USENIX Annual Technical Conference, Boston, MA, USA, 15–17 July 2020; pp. 493–506.
55. McMahan, H.B.; Ramage, D.; Talwar, K.; Zhang, L. Learning differentially private recurrent language models. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
56. Bell, J.H.; Bonawitz, K.A.; Gascón, A.; Lepoint, T.; Raykova, M. Secure single-server aggregation with (poly) logarithmic overhead. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual, 9–13 November 2020; pp. 1253–1269.