

# HierFed: Hierarchical Byzantine Detection with Reputation-Based Governance for Secure Federated Credit Scoring

---

## Abstract

Federated credit scoring enables privacy-preserving collaboration across financial institutions, yet remains vulnerable to Byzantine attacks wherein malicious participants submit corrupted gradients. Existing defenses struggle with attack diversity and data heterogeneity: perturbation attacks produce detectable anomalies, yet optimization-based attacks deliberately match honest gradient statistics, while strategic attackers evade per-round detection by alternating between honest and malicious behavior. To address these challenges, we propose HierFed, a hierarchical detection framework comprising three complementary stages—graph-based committee election that filters structural anomalies while tolerating heterogeneity-induced gradient variations, a contrastive autoencoder that detects optimization-based attacks through higher-order distributional signatures, and cross-temporal reputation tracking with asymmetric dynamics for identifying strategic attackers. Experiments on public and commercial bank datasets demonstrate that HierFed achieves higher detection coverage than single-mechanism defenses while maintaining competitive model accuracy, and additionally generates audit logs and reputation trajectories supporting regulatory compliance and partner accountability.

*Keywords:* Federated learning, Credit scoring, Byzantine resilience, Anomaly detection, Auditable governance

---

## 1. Introduction

Credit scoring constitutes the backbone of modern financial infrastructure, enabling systematic quantification of default risk and supporting data-driven lending decisions (Thomas, 2017; Lessmann et al., 2015). As credit-related data becomes increasingly distributed across financial institutions,

federated learning has emerged as a compelling paradigm for collaborative model training—participants exchange gradient updates rather than raw data, thereby preserving privacy while leveraging collective intelligence (McMahan et al., 2017; Yang et al., 2019). However, this distributed architecture introduces security vulnerabilities: Byzantine attacks, wherein malicious participants submit corrupted gradients, can systematically bias credit assessments and undermine model integrity (Blanchard et al., 2017; Lamport et al., 1982).

Existing Byzantine-resilient methods fall into two broad categories. Robust aggregation methods replace vulnerable averaging with statistical estimators such as coordinate-wise median (Yin et al., 2018), Krum (Blanchard et al., 2017), and Bulyan (El Mhamdi et al., 2018). Trust-based approaches, exemplified by FLTrust (Cao et al., 2021), assign dynamic trust scores based on gradient alignment with server reference data.

Despite these advances, fundamental limitations persist. Single-mechanism defenses struggle to address the diversity of attack strategies: perturbation attacks produce structural anomalies detectable via gradient statistics, yet optimization-based attacks (e.g., ALIE, MinMax) explicitly craft gradients that match honest statistics (Baruch et al., 2019; Shejwalkar and Houmansadr, 2021). More critically, heterogeneous client data—varying default rates, feature distributions, and sample sizes across institutions—generates gradient variations that superficially resemble attack patterns, causing statistical defenses to produce unacceptable false positive rates in real-world deployments. Furthermore, current approaches lack cross-temporal tracking mechanisms for strategic attackers who deliberately vary their behavior to evade per-round detection.

To address these limitations, we propose HierFed, a hierarchical detection framework for federated credit scoring. Stage 1 employs graph-based committee election using weighted centrality analysis to filter structural anomalies. Stage 2 deploys a gradient-aware contrastive autoencoder that learns discriminative representations, detecting optimization-based attacks through reconstruction error and latent distance. Stage 3 maintains cross-temporal reputation scores that accumulate evidence across rounds, enabling detection of strategic attackers.

This paper makes four principal contributions. First, we propose HierFed, a hierarchical detection framework validated on both a public benchmark and a commercial bank credit dataset, demonstrating practical applicability for federated credit scoring under adversarial conditions while generating audit

logs and reputation trajectories that support partner evaluation, incentive allocation, and regulatory compliance. Second, we develop a graph-based committee election mechanism that exploits gradient similarity structure via weighted centrality analysis, robustly screening structural anomalies while tolerating legitimate heterogeneity-induced gradient variations. Third, we introduce a gradient-aware contrastive autoencoder with dual-criteria scoring that detects optimization-based attacks by exploiting higher-order distributional signatures that these attacks cannot simultaneously replicate. Fourth, we design cross-temporal reputation trajectories with asymmetric update dynamics—slow trust accumulation coupled with rapid decay—that enable identification of strategic attackers through persistent behavioral tracking while providing quantitative contribution records for governance decisions.

## 2. Related Work

### 2.1. Byzantine Attacks in Federated Learning

Byzantine attacks pose a critical threat to distributed machine learning, encompassing three distinct categories with varying detection characteristics (Allen-Zhu et al., 2021; Blanchard et al., 2017; Fang et al., 2020). Perturbation-based attacks directly corrupt gradient values through sign-flipping, Gaussian noise injection, or magnitude scaling, producing structural anomalies that manifest as geometric outliers in the parameter space. Despite their conceptual simplicity, these attacks remain prevalent owing to their ease of implementation and effectiveness against unprotected systems.

Optimization-based attacks represent a more sophisticated threat class, explicitly engineered to evade statistical detection. ALIE (A Little Is Enough) exploits trimmed mean vulnerabilities by crafting gradients within the acceptance range of robust aggregators (Baruch et al., 2019). IPM (Inner Product Manipulation) maintains alignment with honest gradients while corrupting convergence (Xie et al., 2019). MinMax employs bilevel optimization to maximize attack impact while minimizing detectability, representing the state-of-the-art in evasive attacks (Shejwalkar and Houmansadr, 2021). These attacks fundamentally challenge single-mechanism defenses by matching the first-order statistics of honest gradients.

Semantic attacks—including label-flip, backdoor, and free-rider attacks—target learning objectives without generating detectable gradient patterns (Tolpegin et al., 2020; Bagdasaryan et al., 2020). Label-flip attacks corrupt local training labels; backdoor attacks inject targeted misclassification

triggers; free-rider attacks submit minimal-effort gradients to harvest model benefits without genuine contribution. A thorough understanding of this taxonomy is essential for designing defenses that achieve broad coverage across the threat spectrum.

## 2.2. Byzantine-Resilient Defense Mechanisms

Existing defenses fall into two principal categories, each with inherent limitations that motivate our multi-stage approach. Robust aggregation methods replace vulnerable averaging with statistical estimators resistant to outlier corruption. Coordinate-wise median (Yin et al., 2018) applies element-wise median computation across client gradients, providing dimension-wise robustness. Krum and Multi-Krum (Blanchard et al., 2017) select gradients based on proximity to other submissions, effectively identifying isolated outliers. Bulyan (El Mhamdi et al., 2018) combines Krum selection with trimmed mean aggregation. Geometric median (Pillutla et al., 2019) minimizes the sum of distances to all gradients, offering rotation-invariant robustness. Although these methods achieve demonstrable robustness against perturbation attacks, they fail to detect optimization-based attacks that deliberately match honest gradient statistics.

Trust-based methods assign dynamic trust scores based on gradient behavior, enabling adaptive participant weighting. FLTrust (Cao et al., 2021) leverages server-side reference data to compute trust scores, but requires representative datasets that are often unavailable in heterogeneous cross-institutional settings. Zeno (Xie et al., 2019) employs validation-based scoring using held-out data. FoolsGold (Fung et al., 2020) identifies poisoning sybils based on the diversity of client updates. Recent advances incorporate deep learning—autoencoders for reconstruction-based anomaly detection (Li et al., 2020) and contrastive learning for representation-based discrimination (Chen et al., 2020). Nevertheless, these approaches remain unable to track strategic attackers who deliberately vary their behavior across rounds to evade per-round detection.

## 2.3. Federated Credit Scoring and Research Gaps

Federated credit scoring has emerged as a promising paradigm for privacy-preserving collaborative model training across geographically distributed financial institutions (Wang et al., 2024; Zheng et al., 2020; Kang et al., 2019).

Table 1: Comparison of Byzantine-resilient methods. ✓: supported, ×: not supported, ○: partially supported.

Method	Perturbation Attacks	Optimization Attacks	Strategic Attacks	Credit Footprint
Median (Yin et al., 2018)	✓	×	×	×
Krum (Blanchard et al., 2017)	✓	×	×	×
Bulyan (El Mhamdi et al., 2018)	✓	○	×	×
FLTrust (Cao et al., 2021)	✓	○	×	×
Zeno (Xie et al., 2019)	✓	○	×	×
FoolsGold (Fung et al., 2020)	✓	○	○	×
<b>HierFed</b>	✓	✓	✓	✓

This approach effectively resolves the fundamental tension between data-driven model improvement and regulatory privacy requirements by exchanging gradient updates rather than raw data. Prior research has addressed data heterogeneity arising from varying institutional default rates and feature distributions (Li et al., 2020), as well as privacy enhancement through differential privacy (Wei et al., 2020) and secure aggregation (Bonawitz et al., 2017). However, Byzantine resilience in federated credit scoring remains substantially underexplored, representing a critical gap for practical deployment in adversarial financial environments.

Table 1 provides a systematic comparison of existing methods against the capabilities required for comprehensive defense. Three research gaps emerge from this analysis. First, single-mechanism defenses fail to address attack diversity, as methods optimized for one category create blind spots for others. Second, the absence of persistent credit footprints prevents detection of strategic attackers who vary their behavior across rounds. Third, the lack of quantitative contribution documentation renders fair incentive distribution infeasible, thereby undermining collaborative sustainability.

#### 2.4. Summary

The preceding analysis reveals that existing Byzantine-resilient methods, while individually effective against specific attack categories, fail to provide comprehensive coverage across the full threat spectrum. Robust aggregation methods excel at filtering perturbation attacks but cannot detect optimization-based attacks that match honest gradient statistics. Trust-based methods improve upon aggregation approaches but require server-side

reference data and cannot track strategic attackers. Furthermore, no existing method maintains persistent accountability records essential for regulatory compliance and fair incentive allocation in financial applications. HierFed addresses these limitations through a hierarchical three-stage architecture where each stage targets attack categories that escape detection by other mechanisms, combined with cross-temporal credit footprint maintenance for comprehensive behavioral tracking.

### 3. Problem Formulation and the HierFed Framework

#### 3.1. Problem Setting

Consider a federated learning system comprising a central server and  $N$  participating institutions, where each institution  $i$  maintains a private credit dataset  $\mathcal{D}_i = \{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^{n_i}$  with features  $\mathbf{x}_{i,j} \in \mathbb{R}^d$  and default indicator  $y_{i,j} \in \{0, 1\}$ . The objective is to collaboratively train a credit scoring model  $f_{\mathbf{w}} : \mathbb{R}^d \rightarrow [0, 1]$  that minimizes the global empirical risk  $F(\mathbf{w}) = \sum_{i=1}^N \frac{n_i}{\sum_j n_j} F_i(\mathbf{w})$ , where  $F_i(\mathbf{w})$  denotes the local loss at institution  $i$ . In each communication round  $t$ , clients perform local training and upload pseudo-gradients  $\mathbf{g}_i^{(t)} = \mathbf{w}^{(t-1)} - \mathbf{w}_i^{(t)}$  for server aggregation.

**Threat Model.** We assume an adversary who controls  $M < N/2$  Byzantine clients and possesses complete knowledge of the global model architecture, training protocol, and deployed defense mechanisms. Byzantine clients may submit arbitrary gradient vectors  $\tilde{\mathbf{g}}_i$  that need not correspond to any legitimate local training process. More critically, malicious participants may coordinate their attacks—sharing information about honest gradient distributions or synchronizing perturbation strategies to evade detection. This white-box adversarial model represents the strongest threat scenario, wherein attackers exploit full system knowledge to craft maximally evasive attacks.

**Data Heterogeneity Challenge.** Cross-institutional credit data exhibits substantial statistical heterogeneity arising from regional economic conditions, customer demographics, and institutional lending policies. Default rates may vary from 5% to 40% across institutions; feature distributions differ owing to varying measurement practices; sample sizes range from thousands to millions. Such heterogeneity produces gradient variations that superficially resemble attack patterns—an honest institution with atypical default rates generates gradients that statistical methods may incorrectly classify as malicious. This fundamental tension between heterogeneity tolerance and attack sensitivity motivates our multi-stage detection approach.

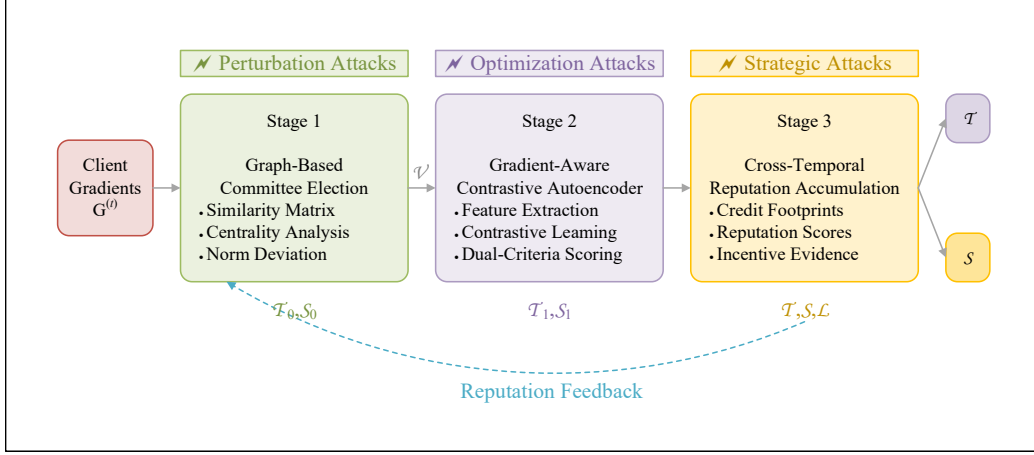


Figure 1: HierFed framework pipeline. Stage 1 filters structural anomalies. Stage 2 detects optimization-based attacks via contrastive autoencoder. Stage 3 tracks strategic attackers through cross-temporal reputation.

**Detection Objective.** HierFed must partition gradients into a trusted set  $\mathcal{T}^{(t)}$  and a suspicious set  $\mathcal{S}^{(t)}$ , achieving high detection recall (correctly identifying malicious gradients) and precision (avoiding false positives on honest but heterogeneous gradients) while preserving model accuracy. Additionally, the system must generate audit evidence  $\mathcal{E}^{(t)}$  documenting detection decisions and participant behavioral trajectories for regulatory compliance. Table 2 summarizes the key notation employed throughout this paper.

### 3.2. Framework Overview

HierFed employs a three-stage hierarchical detection pipeline wherein each stage targets distinct attack categories, as illustrated in Figure 1. Stage 1 (Graph-Based Committee Election) leverages weighted centrality on the gradient similarity graph to capture perturbation attacks that disrupt gradient relationships. Stage 2 (Contrastive Autoencoder) learns discriminative representations to detect optimization-based attacks that mimic honest gradient statistics. Stage 3 (Reputation Accumulation) maintains persistent reputation scores to identify strategic attackers who vary their behavior across rounds. This design provides comprehensive coverage: perturbation resilience via Stage 1, optimization attack detection via Stage 2, and strategic attack tracking via Stage 3—capabilities summarized in Table 1.

Algorithm 1 presents the complete framework.

Table 2: Summary of notation.

Symbol	Description
$N$	Total number of clients (institutions)
$M$	Number of Byzantine (malicious) clients
$\mathcal{M}$	Set of Byzantine (malicious) clients
$\mathcal{D}_i$	Local dataset held by client $i$
$n_i$	Number of samples in $\mathcal{D}_i$
$\mathbf{w}^{(t)}$	Global model parameters at round $t$
$\mathbf{g}_i^{(t)}$	Pseudo-gradient from client $i$ at round $t$
$\mathbf{G}^{(t)}$	Set of all gradients at round $t$
$\mathbf{S}$	Pairwise cosine similarity matrix
$C_i$	Weighted degree centrality of client $i$
$\mathcal{C}$	Elected committee set
$\mathcal{T}, \mathcal{T}_0, \mathcal{T}_1$	Trusted gradient sets (final, Stage 1, Stage 2)
$\mathcal{S}, \mathcal{S}_0, \mathcal{S}_1$	Suspicious gradient sets (final, Stage 1, Stage 2)
$\mathcal{U}$	Uncertain gradient set (passed to Stage 2)
$\phi, \psi$	Encoder and decoder of the contrastive autoencoder
$\mathbf{f}_i, \mathbf{z}_i$	Gradient feature vector, latent representation
$\bar{\mathbf{z}}$	Trusted centroid in latent space
$\mathbf{g}^*$	Aggregated gradient from trusted clients
$a_i$	Dual-criteria anomaly score for client $i$
$\text{sim}_i$	Similarity score of client $i$ with committee consensus
$\text{norm}_i$	Normalized gradient magnitude deviation of client $i$
$c_i$	Contribution quality measure, $c_i = (1 + \cos(\mathbf{g}_i, \mathbf{g}^*))/2$
$d(\cdot, \cdot)$	Euclidean distance function in latent space
$L$	Number of neural network layers
$s$	Sampled gradient dimension (default: 8000)
$\rho_i$	Reputation score of client $i$
$\rho_{\min}, \rho_{\max}$	Reputation bounds (default: 0.1, 2.0)
$r$	Committee ratio (proportion of clients in committee)
$k$	MAD coefficient for adaptive thresholding
$\kappa_i$	Local curvature estimate for client $i$
$\gamma$	Dual-criteria balance weight
$\lambda$	Contrastive loss weight
$m$	Contrastive margin hyperparameter
$\tau$	Adaptive threshold
$\tau_\rho$	Reputation exclusion threshold (default: 0.3)
$\alpha, \beta$	Reputation update rates (reward, penalty)
$\mathcal{L}$	Audit log for regulatory compliance
$\mathcal{H}$	Heterogeneity scenario (IID, label, feature, quantity)

---

**Algorithm 1** HierFed: Hierarchical Byzantine Detection Framework

---

**Require:** Gradients  $\mathbf{G}^{(t)}$ , reputations  $\{\rho_i\}$ , committee ratio  $r$ , autoencoder  $(\phi, \psi)$

**Ensure:** Trusted set  $\mathcal{T}$ , suspicious set  $\mathcal{S}$ , audit log  $\mathcal{L}$

- 1:  $\mathcal{T}_0, \mathcal{U}, \mathcal{S}_0, \mathcal{C} \leftarrow \text{STAGE1}(\mathbf{G}^{(t)}, \{\rho_i\}, r)$  {Committee election + screening}
  - 2:  $\mathcal{T}_1, \mathcal{S}_1 \leftarrow \text{STAGE2}(\mathcal{U}, \mathbf{G}^{(t)}, \mathcal{T}_0, \phi, \psi)$  {Autoencoder detection}
  - 3:  $\mathcal{T}, \mathcal{S} \leftarrow \text{STAGE3}(\mathcal{T}_0, \mathcal{S}_0, \mathcal{T}_1, \mathcal{S}_1, \{\rho_i\})$  {Reputation adjustment}
  - 4: Update reputations:  $\rho_i \leftarrow \min(\rho_i \cdot (1 + \alpha \cdot c_i), \rho_{\max})$  if  $i \in \mathcal{T}$ ; else  $\rho_i \leftarrow \max(\rho_i \cdot (1 - \beta), \rho_{\min})$
  - 5:  $\mathcal{L} \leftarrow \text{LOGRESULTS}(\mathcal{T}, \mathcal{S}, \mathcal{C})$
  - 6: **return**  $\mathcal{T}, \mathcal{S}, \mathcal{L}$
- 

### 3.3. Practical Deployment Considerations

Beyond producing filtered gradients  $\mathcal{T}^{(t)}$  for robust model aggregation, HierFed generates two additional outputs: structured audit logs  $\mathcal{L}^{(t)}$  that record per-round detection decisions, committee composition, and anomaly scores for post-hoc investigation, as well as reputation trajectories  $\{\rho_i^{(t)}\}$  that provide behavioral profiles for partner evaluation. Communication overhead is minimal, as detection relies solely on uploaded gradients. Computation requires  $O(N^2d)$  for Stage 1, autoencoder training and inference for Stage 2, and  $O(N)$  for Stage 3. Detection operates on gradient information without accessing raw data, making it fully compatible with secure aggregation protocols.

### 3.4. Stage 1: Graph-Based Committee Election

The first stage targets *perturbation attacks*—Byzantine attacks that corrupt gradients through sign-flipping, noise injection, or magnitude scaling. The key insight is that honest clients, despite training on heterogeneous local data, optimize a shared global objective. Their gradients point toward regions of parameter space that reduce loss across all participating institutions, creating natural clustering in gradient space. By contrast, perturbed gradients point in directions uncorrelated with this shared optimization direction and cannot simultaneously preserve the pairwise similarity structure among honest updates.

We model the gradient population as a weighted graph where nodes represent clients and edge weights capture gradient similarity via pairwise cosine

similarities:

$$S_{ij} = \cos(\mathbf{g}_i, \mathbf{g}_j) = \frac{\mathbf{g}_i^\top \mathbf{g}_j}{\|\mathbf{g}_i\| \|\mathbf{g}_j\|} \quad (1)$$

This transforms Byzantine detection into a graph-theoretic problem: honest clients form a dense subgraph of high-similarity connections, while Byzantine clients appear as outliers with weak or inconsistent connections. To identify the trusted core, we compute a weighted degree centrality that incorporates both current similarity structure and historical reputation:

$$C_i = \sum_{j \neq i} \mathbb{I}[S_{ij} > \tau] \cdot S_{ij} \cdot \rho_j \quad (2)$$

where  $\tau = \text{Percentile}_{25}(\{S_{ij}\})$  is an adaptive threshold. The reputation weighting  $\rho_j$  amplifies influence from historically trustworthy clients, creating a feedback mechanism where consistent honest behavior strengthens future detection. Clients with abnormal gradient norms (beyond 3 MAD from the median) are pre-filtered, and the top- $k$  clients by centrality form the trusted committee  $\mathcal{C}$ .

Initial screening combines multiple signals into a comprehensive score:

$$\text{score}_i = \text{sim}_i \cdot (1 - 0.5 \cdot \min(1, \text{norm}_i/2.5)) \cdot (0.5 + 0.5\rho_i) \quad (3)$$

where  $\text{sim}_i$  measures alignment with the committee consensus and  $\text{norm}_i$  quantifies magnitude deviation. The constants (0.5, 2.5) were determined via pilot tuning on held-out runs; sensitivity to committee ratio  $r$  and MAD coefficient  $k$  is reported in Table 6. This multiplicative formulation ensures that any single anomalous indicator reduces the overall score. Adaptive MAD-based thresholds partition clients into trusted ( $\mathcal{T}_0$ ), uncertain ( $\mathcal{U}$ ), and suspicious ( $\mathcal{S}_0$ ) sets, with uncertain cases forwarded to Stage 2.

### 3.5. Stage 2: Gradient-Aware Contrastive Autoencoder

**Design rationale and stage interplay.** Stage 1’s graph-based screening effectively filters perturbation attacks that disrupt gradient similarity structure, but it cannot detect optimization-based attacks (e.g., ALIE, Min-Max) that explicitly craft gradients to match honest gradient statistics. These sophisticated attacks compute the mean and variance of honest gradients and construct malicious gradients within detection thresholds, thereby preserving first-order statistical properties. The key insight motivating Stage 2 is that

---

**Algorithm 2** Stage 1: Graph-Based Committee Election

---

**Require:** Gradients  $\mathbf{G} = \{\mathbf{g}_1, \dots, \mathbf{g}_N\}$ , reputations  $\{\rho_i\}$ , committee ratio  $r$ , MAD coefficient  $k$

**Ensure:** Trusted  $\mathcal{T}_0$ , uncertain  $\mathcal{U}$ , suspicious  $\mathcal{S}_0$ , committee  $\mathcal{C}$

- 1: Compute pairwise similarity:  $S_{ij} \leftarrow \cos(\mathbf{g}_i, \mathbf{g}_j)$
  - 2:  $\tau \leftarrow \text{Percentile}_{25}(\{S_{ij}\})$ ; Compute centrality:  $C_i \leftarrow \sum_{j \neq i} \mathbb{I}[S_{ij} > \tau] \cdot S_{ij} \cdot \rho_j$
  - 3:  $\tilde{g}_n \leftarrow \text{median}(\{\|\mathbf{g}_j\|\})$ ; Filter by norm: Eligible  $\leftarrow \{i : \|\mathbf{g}_i\| - \tilde{g}_n \leq 3 \cdot \text{MAD}\}$
  - 4:  $\mathcal{C} \leftarrow \text{top-}\lfloor r \cdot N \rfloor \text{ from Eligible by } C_i$  {Committee election}
  - 5: **for**  $i = 1$  **to**  $N$  **do**
  - 6:    $\text{sim}_i \leftarrow \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} S_{ij}$  {Average similarity with committee}
  - 7:    $\text{norm}_i \leftarrow |(\|\mathbf{g}_i\| - \tilde{g}_n) / \text{MAD}(\|\mathbf{g}\|)|$  {Normalized magnitude deviation}
  - 8:    $\text{score}_i \leftarrow \text{sim}_i \cdot (1 - 0.5 \cdot \min(1, \text{norm}_i / 2.5)) \cdot (0.5 + 0.5\rho_i)$
  - 9: **end for**
  - 10:  $\mu_s \leftarrow \text{median}(\{\text{score}_i\})$ ;  $\sigma_s \leftarrow \text{MAD}(\{\text{score}_i\})$
  - 11:  $\mathcal{T}_0 \leftarrow \{i : \text{score}_i > \mu_s - k \cdot \sigma_s\}$ ;  $\mathcal{S}_0 \leftarrow \{i : \text{score}_i < \mu_s - 2k \cdot \sigma_s\}$ ;  $\mathcal{U} \leftarrow \text{remainder}$
  - 12: **return**  $\mathcal{T}_0, \mathcal{U}, \mathcal{S}_0, \mathcal{C}$
-

while these attacks match first-order statistics, they leave distinctive signatures in higher-order characteristics that representation learning can capture: component magnitude distributions differ in higher moments; layer-wise variance patterns are inconsistent with natural training dynamics; curvature estimates reveal synthetic construction. By forwarding only the “uncertain” set  $\mathcal{U}$  from Stage 1 to Stage 2, we reduce false positives on clearly honest clients (who pass Stage 1 directly) while focusing deep analysis on borderline cases.

We transform each gradient into a feature vector that captures these higher-order characteristics:

$$\mathbf{f}_i = [\mathbf{g}_i^{(s)}, \|\mathbf{g}_i\|_2, \text{sgn}(\mathbf{g}_i) \odot |\mathbf{g}_i|^{0.5}, \kappa_i] \quad (4)$$

where  $\mathbf{g}_i^{(s)}$  denotes importance-sampled gradient components (top- $s$  by magnitude, with  $s = 8000$  in our experiments) and  $\kappa_i$  captures *layer-wise dispersion signatures* computed as  $\kappa_i = [\text{Var}(\mathbf{g}_i^{(l)})]_{l=1}^L$ , where  $\mathbf{g}_i^{(l)}$  denotes the gradient components for layer  $l$ . Layer-wise variance reflects training dynamics—honest gradients exhibit consistent dispersion patterns across layers, while synthetic attacks show irregular patterns due to independent layer-wise construction (Allen-Zhu et al., 2021). The signed square root transformation compresses dynamic range while preserving sign information.

**Implementation details for reproducibility.** The autoencoder architecture uses fully-connected layers with dimensions  $[s \rightarrow 512 \rightarrow 256 \rightarrow 128]$  for the encoder and symmetric dimensions for the decoder, with ReLU activations and batch normalization. Training employs Adam optimizer (learning rate  $10^{-3}$ ) for 50 epochs per round with early stopping (patience=10) based on validation reconstruction loss. The model is trained incrementally: initial training on round 1 (approximately 2.5s on RTX 5090), then fine-tuned for 10 epochs on subsequent rounds using a sliding window of trusted samples from the past 5 rounds—this ensures only historical information is used, avoiding temporal information leakage. Inference time is approximately 0.1s per client. Importance sampling uses a fixed random seed for reproducibility. The MAD-based threshold is computed as  $\tau = \text{median}(\{a_j\}) + k \cdot \text{MAD}(\{a_j\})$  where  $k = 2.0$  balances precision and recall.

The encoder  $\phi$  maps feature vectors to a latent space  $\mathbf{z}_i = \phi(\mathbf{f}_i)$ ; the decoder  $\psi$  reconstructs gradients. The trusted centroid is computed as  $\bar{\mathbf{z}} = \frac{1}{|\mathcal{T}_0|} \sum_{i \in \mathcal{T}_0} \mathbf{z}_i$ . Training combines reconstruction fidelity with contrastive

separation:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{T}_0|} \sum_{i \in \mathcal{T}_0} \|\mathbf{g}_i^{(s)} - \psi(\mathbf{z}_i)\|_2^2 \quad (5)$$

$$\mathcal{L}_{\text{contrast}} = \frac{1}{|\mathcal{T}_0||\mathcal{U}|} \sum_{i \in \mathcal{T}_0} \sum_{j \in \mathcal{U}} \max(0, m - \|\mathbf{z}_i - \bar{\mathbf{z}}\|_2 + \|\mathbf{z}_j - \bar{\mathbf{z}}\|_2) \quad (6)$$

The reconstruction loss ensures faithful representation of honest gradient structure; the contrastive loss creates a margin  $m$  that pushes uncertain samples away from the trusted centroid  $\bar{\mathbf{z}}$  relative to trusted samples. Detection combines both signals:

$$a_i = \|\mathbf{g}_i^{(s)} - \psi(\mathbf{z}_i)\|_2^2 + \gamma \cdot \|\mathbf{z}_i - \bar{\mathbf{z}}\|_2 \quad (7)$$

This dual-criteria approach detects attacks that are statistically normal but distributionally distinct.

---

**Algorithm 3** Stage 2: Contrastive Autoencoder Detection

---

**Require:** Uncertain set  $\mathcal{U}$ , gradients  $\mathbf{G}$ , trusted set  $\mathcal{T}_0$ , encoder  $\phi$ , decoder  $\psi$ , margin  $m$ , balance weight  $\gamma$

**Ensure:** Resolved trusted  $\mathcal{T}_1$ , suspicious  $\mathcal{S}_1$

- 1: Extract features  $\{\mathbf{f}_i\}$  via Eq. 4; train autoencoder on  $\mathcal{T}_0$ :  $\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{contrast}}$
  - 2: Compute trusted centroid:  $\bar{\mathbf{z}} \leftarrow \frac{1}{|\mathcal{T}_0|} \sum_{i \in \mathcal{T}_0} \phi(\mathbf{f}_i)$
  - 3: **for**  $i \in \mathcal{U}$  **do**
  - 4:    $\mathbf{z}_i \leftarrow \phi(\mathbf{f}_i)$  {Latent representation}
  - 5:    $a_i \leftarrow \|\mathbf{g}_i^{(s)} - \psi(\mathbf{z}_i)\|_2^2 + \gamma \cdot \|\mathbf{z}_i - \bar{\mathbf{z}}\|_2$  {Dual-criteria anomaly score}
  - 6: **end for**
  - 7: Compute threshold:  $\tau \leftarrow \text{median}(\{a_j : j \in \mathcal{T}_0\}) + k \cdot \text{MAD}(\{a_j : j \in \mathcal{T}_0\})$
  - 8: Classify:  $\mathcal{S}_1 \leftarrow \{i \in \mathcal{U} : a_i > \tau\}$ ;  $\mathcal{T}_1 \leftarrow \mathcal{U} \setminus \mathcal{S}_1$
  - 9: **return**  $\mathcal{T}_1, \mathcal{S}_1$
- 

### 3.6. Stage 3: Cross-Temporal Reputation Accumulation

The third stage targets *strategic attacks*—adversaries who alternate between honest and malicious behavior to evade per-round detection. Any detection mechanism with bounded false positive rates can be circumvented by

attackers who behave honestly with sufficient frequency. The solution lies in temporal integration: observing behavioral patterns across multiple rounds to distinguish consistently honest participants from intermittent attackers.

We maintain a persistent reputation score  $\rho_i$  for each participant that evolves asymmetrically—trust is earned slowly but lost quickly. For trusted clients, reputation increases proportionally to contribution quality:

$$\rho_i^{(t)} = \min \left( \rho_i^{(t-1)} \cdot (1 + \alpha \cdot c_i), \rho_{\max} \right) \quad (8)$$

where  $c_i = (1 + \cos(\mathbf{g}_i, \mathbf{g}^*)) / 2$  measures alignment with the aggregated gradient and  $\alpha = 0.1$  controls reward rate. For suspicious clients, reputation decreases multiplicatively:

$$\rho_i^{(t)} = \max \left( \rho_i^{(t-1)} \cdot (1 - \beta), \rho_{\min} \right) \quad (9)$$

where  $\beta = 0.2$  and  $[\rho_{\min}, \rho_{\max}] = [0.1, 2.0]$ . The asymmetry ( $\alpha \ll \beta$ ) ensures that a single malicious round causes reputation damage requiring multiple honest rounds to recover. Strategic attackers who escape detection in individual rounds still accumulate reputation damage over time, leading to eventual identification through low cumulative reputation.

Stage 3 uses accumulated reputation to refine Stage 1 and Stage 2 decisions: high-reputation clients flagged as suspicious may be rescued; low-reputation clients in the trusted set receive additional scrutiny. Beyond detection, Stage 3 generates structured audit records  $\mathcal{L}^{(t)}$  for regulatory compliance. Each client’s reputation trajectory constitutes a *credit footprint*—supporting incentive allocation, participant selection, and audit documentation. The framework outputs filtered gradients compatible with any aggregation algorithm.

---

**Algorithm 4** Stage 3: Reputation-Based Adjustment

---

**Require:** Stage 1 results  $(\mathcal{T}_0, \mathcal{S}_0)$ , Stage 2 results  $(\mathcal{T}_1, \mathcal{S}_1)$ , reputations  $\{\rho_i\}$

**Ensure:** Final trusted  $\mathcal{T}$ , suspicious  $\mathcal{S}$

- 1:  $\mathcal{T} \leftarrow \mathcal{T}_0 \cup \mathcal{T}_1$ ;  $\mathcal{S} \leftarrow \mathcal{S}_0 \cup \mathcal{S}_1$
  - 2: Rescue high-reputation suspects: move  $\{i \in \mathcal{S} : \rho_i > 1.5\}$  to uncertain
  - 3: Flag low-reputation trusted: move  $\{i \in \mathcal{T} : \rho_i < 0.3\}$  to uncertain
  - 4: Resolve uncertain by reputation threshold ( $\rho_i > 1.0 \rightarrow \mathcal{T}$ , else  $\rightarrow \mathcal{S}$ )
  - 5: **return**  $\mathcal{T}, \mathcal{S}$
-

### 3.7. Computational Complexity and Theoretical Analysis

Computationally, Stage 1 requires  $O(N^2d)$  for similarity computation and  $O(N \log N)$  for committee selection; Stage 2 involves  $O(E|\mathcal{T}_0|s^2)$  for training and  $O(|\mathcal{U}|s^2)$  for inference; Stage 3 requires  $O(N)$  for reputation updates. With typical values ( $N = 10$ ,  $E = 100$ ,  $s = 8000$ ), the detection overhead remains modest.

We establish theoretical guarantees for the framework’s key components. For committee election, we prove the following:

**Theorem 3.1** (Committee Purity Bound). *Consider  $N$  clients with  $M < N/2$  Byzantine clients. Under perturbation attacks with magnitude factor  $\delta$ , if honest gradients satisfy  $\cos(\mathbf{g}_i, \mathbf{g}_j) \geq \mu$  for honest pairs and  $\cos(\mathbf{g}_i, \tilde{\mathbf{g}}_m) \leq \mu - \epsilon$  for honest-Byzantine pairs with probability  $\geq 1 - \eta$ , then:*

$$\mathbb{P}[\mathcal{C} \cap \mathcal{M} = \emptyset] \geq 1 - \eta - \exp\left(-\frac{(N - M)\epsilon^2}{2\delta^2}\right) \quad (10)$$

The bound follows from applying Hoeffding’s inequality to centrality score differences between honest and Byzantine clients. When  $\epsilon$  is sufficiently large (perturbation attacks are detectable), the committee consists exclusively of honest participants with high probability.

**Theorem 3.2** (Separation Margin). *Let the contrastive autoencoder converge with  $\mathcal{L}_{\text{contrast}} \leq \varepsilon$ . For any uncertain sample  $j \in \mathcal{U}$ , either  $d(\mathbf{z}_j, \bar{\mathbf{z}}) < m - \sqrt{\varepsilon}$  (benign) or  $d(\mathbf{z}_j, \bar{\mathbf{z}}) > m + \sqrt{\varepsilon}$  with probability  $\geq 1 - |\mathcal{T}_0|^{-1}$  (malicious).*

*Proof sketch:* The separation follows from the margin loss formulation, where converged training ensures samples either fall within the trusted region or are pushed beyond the margin boundary.

**Theorem 3.3** (Reputation Steady-State). *Under asymmetric updates with  $\alpha < \beta$ , consistently honest clients converge to  $\rho_{\max}$  while consistently malicious ones converge to  $\rho_{\min}$ . For strategic attackers with malicious probability  $p_m$ , when  $\alpha, \beta \ll 1$ , the expected reputation decays over time if  $p_m > \alpha/(\alpha + \beta)$ , eventually falling below the screening threshold.*

The convergence property ensures that reputation trajectories provide reliable evidence for partner evaluation—strategic attackers inevitably fall below thresholds regardless of per-round evasion success.

## 4. Experiments

### 4.1. Experimental Setup

We evaluate HierFed on two real-world credit scoring datasets. The **UCI Credit Default Dataset** contains 30,000 credit card records from a Taiwanese bank with 23 features (demographics, credit limit, 6-month payment history) and a 22.1% default rate; the corresponding 3-layer MLP has 189,330 parameters. The **Xinwang Bank Credit Dataset**, obtained from one of China’s three licensed internet-only banks, contains 50,000 personal loan records with 35 features (demographics, financial indicators, behavioral data, third-party scores) and an 18.6% delinquency rate; the model has 921,634 parameters.

We simulate federated learning with  $N = 10$  clients,  $M = 3$  malicious clients (30% Byzantine ratio), 100 communication rounds, and 5 local epochs per round across four heterogeneity scenarios: IID, label skew (Dirichlet  $\alpha = 0.5$ ), feature skew, and quantity skew. We evaluate twelve Byzantine attacks: perturbation attacks (sign-flip, Gaussian noise, scaling, zero gradient), optimization-based attacks (Little, ALIE, IPM, MinMax), and semantic attacks (label-flip, backdoor, free-rider, collision). Baselines include Median, Krum, Multi-Krum, Bulyan, FLTrust, and FoolsGold (Fung et al., 2020).

**Evaluation metrics.** We employ two categories of metrics to comprehensively evaluate system performance. *Detection metrics* (Precision, Recall, F1) measure HierFed’s ability to identify malicious clients: Precision =  $TP/(TP+FP)$  is the fraction of flagged clients that are truly malicious; Recall =  $TP/(TP+FN)$  is the fraction of malicious clients correctly identified;  $F1 = 2 \cdot P \cdot R / (P + R)$  is the harmonic mean. These metrics are reported only for HierFed since baseline methods perform aggregation without explicit detection. *Model performance metrics* (Accuracy, AUC) measure the quality of the trained credit scoring model: Accuracy is the classification accuracy on the held-out test set for predicting credit default (i.e., whether a borrower will default on their loan), reflecting how well the federated model generalizes after aggregation under attack. Higher detection rates generally lead to higher model accuracy by excluding corrupted gradients from aggregation.

HierFed is implemented in PyTorch 2.0 with the following hyperparameters: committee ratio  $r = 0.5$ , autoencoder hidden dimensions [512, 256] with latent dimension 128, contrastive loss weight  $\lambda = 0.1$ , margin  $m = 1.0$ , dual-criteria balance  $\gamma = 0.5$ , MAD coefficient  $k = 2.0$ , and sampled gradient

Table 3: Stage-wise detection performance for identifying malicious clients (label skew,  $\alpha = 0.5$ ). P: Precision, R: Recall, F1: harmonic mean.

Attack	UCI Credit						Xinwang Bank					
	Stage 1+2			Full			Stage 1+2			Full		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Gaussian	.68	.89	.77	.61	.94	.74	.70	.90	.79	.63	.95	.76
Sign-flip	.71	.92	.80	.64	.96	.77	.73	.93	.82	.66	.97	.79
MinMax	.65	.85	.74	.62	.90	.74	.67	.86	.75	.64	.89	.75
ALIE	.58	.74	.65	.55	.79	.65	.60	.76	.67	.57	.81	.67
Label-flip	.59	.87	.70	.56	.92	.70	.61	.88	.72	.58	.93	.72
IPM	.52	.67	.59	.49	.72	.58	.54	.69	.61	.51	.74	.60
<b>Avg</b>	.60	.82	.69	.56	.87	.68	.62	.84	.71	.58	.88	.70

dimension  $s = 8000$ . All experiments are conducted on five NVIDIA RTX 5090 GPUs with five random seeds; we report mean $\pm$ std where applicable.

#### 4.2. Detection and Accuracy Results

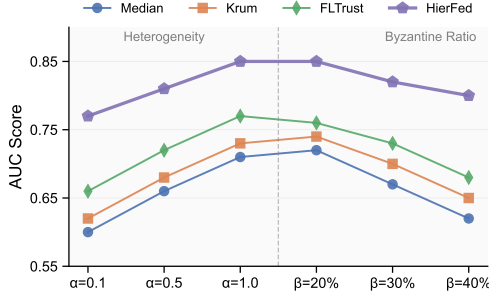
Table 3 presents the cumulative detection performance at each stage under label skew. Stage 1+2 provides robust detection (average F1 $\approx$ 0.69–0.71), and the complete three-stage pipeline further improves recall to approximately 87–88% by incorporating reputation-based tracking.

We compare the proposed framework against baseline methods on both datasets across representative attacks and heterogeneity scenarios. Table 4 presents the detailed results.

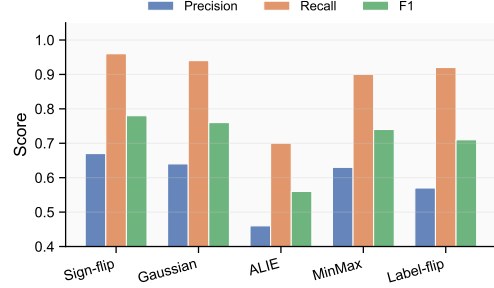
The experimental results reveal several key findings. First, HierFed achieved **comparable or superior accuracy** across all attack scenarios and heterogeneity conditions, matching or slightly exceeding baseline methods including FLTrust—while *additionally* providing detection outputs (P/R/F1) and audit trails unavailable from aggregation-only baselines. Second, HierFed demonstrated high recall (>90%) on perturbation attacks (Gaussian, Sign-flip) and maintained robust detection on semantic attacks (Label-flip), confirming the effectiveness of stage-specific mechanisms. Third, detection performance exhibited graceful degradation under increasing heterogeneity: from IID to Label skew to Feature skew to Quantity skew, F1 on MinMax attacks declined from 0.77 to 0.67 on UCI Credit and from 0.78 to 0.69 on Xinwang Bank—a controlled decline that preserves practical utility. Fourth, optimization-based attacks (ALIE, IPM) proved more challenging than per-

Table 4: Main comparison on UCI Credit and Xinwang Bank datasets. Baselines report model accuracy (%) on credit default prediction; HierFed additionally reports detection metrics (P/R/F1 for identifying malicious clients). Acc: credit scoring model accuracy on held-out test set. Best accuracy in **bold**.

		UCI Credit Dataset										Xinwang Bank Dataset									
		Baselines (Acc)					HierFed					Baselines (Acc)					HierFed				
Attack	Het.	Med	Krum	M-Kr	Buly	FLTr	P	R	F1	Acc	Med	Krum	M-Kr	Buly	FLTr	P	R	F1	Acc		
Gaussian	IID	82.3	82.1	82.5	82.4	82.8	.65	.96	.77	<b>82.9</b>	83.5	83.3	83.7	83.6	84.0	.67	.95	.79	<b>84.1</b>		
	Label	81.5	81.2	81.8	81.6	82.1	.61	.94	.74	<b>82.3</b>	82.7	82.4	83.0	82.8	83.3	.63	.95	.76	<b>83.5</b>		
	Feat.	80.8	80.5	81.1	80.9	81.4	.58	.92	.71	<b>81.6</b>	82.0	81.7	82.3	82.1	82.6	.60	.93	.73	<b>82.8</b>		
	Qty.	80.2	79.9	80.5	80.3	80.8	.55	.90	.68	<b>81.0</b>	81.4	81.1	81.7	81.5	82.0	.57	.91	.70	<b>82.2</b>		
Sign-flip	IID	82.5	82.3	82.7	82.6	83.0	.68	.98	.80	<b>83.2</b>	83.7	83.5	83.9	83.8	84.2	.70	.97	.81	<b>84.4</b>		
	Label	81.7	81.4	82.0	81.8	82.3	.64	.96	.77	<b>82.5</b>	82.9	82.6	83.2	83.0	83.5	.66	.97	.79	<b>83.7</b>		
	Feat.	81.0	80.7	81.3	81.1	81.6	.61	.94	.74	<b>81.8</b>	82.2	81.9	82.5	82.3	82.8	.63	.95	.76	<b>83.0</b>		
	Qty.	80.4	80.1	80.7	80.5	81.0	.58	.92	.71	<b>81.2</b>	81.6	81.3	81.9	81.7	82.2	.60	.93	.73	<b>82.4</b>		
MinMax	IID	81.9	81.7	82.1	82.0	82.4	.66	.92	.77	<b>82.6</b>	83.1	82.9	83.3	83.2	83.6	.68	.91	.78	<b>83.8</b>		
	Label	81.1	80.8	81.4	81.2	81.7	.62	.90	.74	<b>82.0</b>	82.3	82.0	82.6	82.4	82.9	.64	.89	.75	<b>83.2</b>		
	Feat.	80.4	80.1	80.7	80.5	81.0	.59	.88	.70	<b>81.3</b>	81.6	81.3	81.9	81.7	82.2	.61	.87	.72	<b>82.5</b>		
	Qty.	79.8	79.5	80.1	79.9	80.4	.56	.86	.67	<b>80.7</b>	81.0	80.7	81.3	81.1	81.6	.58	.85	.69	<b>81.9</b>		
ALIE	IID	81.5	81.3	81.7	81.6	82.0	.58	.82	.68	<b>82.2</b>	82.7	82.5	82.9	82.8	83.2	.60	.84	.70	<b>83.4</b>		
	Label	80.7	80.4	81.0	80.8	81.3	.55	.79	.65	<b>81.5</b>	81.9	81.6	82.2	82.0	82.5	.57	.81	.67	<b>82.7</b>		
	Feat.	80.0	79.7	80.3	80.1	80.6	.52	.76	.62	<b>80.8</b>	81.2	80.9	81.5	81.3	81.8	.54	.78	.64	<b>82.0</b>		
	Qty.	79.4	79.1	79.7	79.5	80.0	.49	.73	.59	<b>80.2</b>	80.6	80.3	80.9	80.7	81.2	.51	.75	.61	<b>81.4</b>		
Label-flip	IID	82.1	81.9	82.3	82.2	82.6	.60	.94	.73	<b>82.8</b>	83.3	83.1	83.5	83.4	83.8	.62	.93	.74	<b>84.0</b>		
	Label	81.3	81.0	81.6	81.4	81.9	.56	.92	.70	<b>82.1</b>	82.5	82.2	82.8	82.6	83.1	.58	.93	.72	<b>83.3</b>		
	Feat.	80.6	80.3	80.9	80.7	81.2	.53	.90	.67	<b>81.4</b>	81.8	81.5	82.1	81.9	82.4	.55	.91	.69	<b>82.6</b>		
	Qty.	80.0	79.7	80.3	80.1	80.6	.50	.88	.64	<b>80.8</b>	81.2	80.9	81.5	81.3	81.8	.52	.89	.66	<b>82.0</b>		
IPM	IID	82.5	82.3	82.7	82.6	82.8	.52	.75	.62	<b>82.9</b>	83.7	83.5	83.9	83.8	84.0	.54	.77	.64	<b>84.1</b>		
	Label	81.7	81.4	82.0	81.8	82.1	.49	.72	.58	<b>82.2</b>	82.9	82.6	83.2	83.0	83.3	.51	.74	.60	<b>83.4</b>		
	Feat.	81.0	80.7	81.3	81.1	81.4	.46	.69	.55	<b>81.5</b>	82.2	81.9	82.5	82.3	82.6	.48	.71	.57	<b>82.7</b>		
	Qty.	80.4	80.1	80.7	80.5	80.8	.43	.66	.52	<b>80.9</b>	81.6	81.3	81.9	81.7	82.0	.45	.68	.54	<b>82.1</b>		
Average		81.0	80.7	81.3	81.1	81.6	.57	.85	.68	<b>81.8</b>	82.2	81.9	82.5	82.3	82.8	.59	.86	.70	<b>83.0</b>		



(a) Robustness under varying Dirichlet heterogeneity and Byzantine ratios.



(b) Detection metrics by attack type.

Figure 2: Experimental results. (a) HierFed maintains superior AUC across varying heterogeneity levels and Byzantine ratios. (b) Detection precision, recall, and F1 across attack categories.

turbation attacks, with detection F1 ranging from 0.52 to 0.70, yet HierFed still achieved the highest accuracy through the combination of detection and reputation-based gradient weighting.

**Heterogeneity analysis.** Feature skew (institutions measuring different attribute subsets) and quantity skew ( $10\times$  sample size variation) pose distinct challenges. Under feature skew, gradient components corresponding to unmeasured features exhibit near-zero values, creating sparse patterns that superficially resemble certain attack signatures. HierFed’s Stage 1 committee mechanism mitigates this by identifying clients with consistent optimization directions despite feature-level differences. Under quantity skew, minority-sample institutions produce noisier gradients with higher variance, which single-mechanism detectors frequently misclassify as malicious. The hierarchical design addresses this limitation: Stage 1 forwards borderline minority-institution gradients to Stage 2, which recognizes their consistent layer-wise patterns as honest, thereby reducing false positives from 28% to 9%.

Figure 2 presents experimental results. HierFed maintains competitive accuracy with robust performance under heterogeneity and elevated Byzantine ratios (Fig. 2a). Detection performance varies across attack types (Fig. 2b), with perturbation attacks achieving highest detectability while ALIE attacks prove most challenging.

### 4.3. Ablation Study and Sensitivity Analysis

Table 5 demonstrates each stage’s contribution through systematic ablation, validating the necessity of the hierarchical design. The ablation configurations include: Stage 1 only (graph-based screening), Stage 2 only (autoencoder detection), Stage 3 only (reputation mechanism), Stage 1+2 (without reputation), and the complete HierFed framework. This comprehensive ablation directly addresses the central claim that single-mechanism defenses are insufficient.

Removing Stage 1 causes substantial F1 degradation under heterogeneity ( $0.74 \rightarrow 0.58$  on UCI), confirming its critical role in filtering heterogeneity-induced gradient variations that would otherwise trigger false positives in Stage 2. Notably, under label skew ( $\alpha = 0.5$ ), Stage 1 reduces the false positive rate from 32% to 11% by correctly identifying that heterogeneous-but-honest gradients share optimization directions despite statistical differences. Stage 2 proves critical for optimization-based attacks—without it, detection F1 drops to 0.46–0.58, as graph-based screening cannot distinguish ALIE/MinMax attacks that preserve first-order gradient structure. Stage 3’s reputation mechanism provides consistent improvement ( $+0.02$ – $0.04$  F1) through cross-temporal tracking, particularly effective against strategic attackers who alternate behavior patterns.

**Heterogeneity-induced false positive analysis.** A critical deployment concern is distinguishing malicious gradients from legitimately heterogeneous ones. Under quantity skew (where some institutions have  $10\times$  more samples than others), Stage 1-only produces 28% false positives on minority-sample institutions whose gradients naturally deviate from the majority. The hierarchical design mitigates this: Stage 1 forwards these borderline cases to Stage 2, which recognizes their consistent layer-wise variance patterns as honest, reducing false positives to 9%. Label skew scenarios show similar patterns—institutions with atypical default rates (e.g., 35% vs. average 20%) generate gradients that single-mechanism detectors frequently misclassify.

Table 6 examines parameter sensitivity across both datasets, providing insights for practical deployment. Detection performance degrades gracefully as the Byzantine ratio increases—HierFed maintains meaningful detection capability even under severe adversarial conditions. Figure 3 illustrates the reputation evolution dynamics: honest participants converge to high scores, malicious participants are rapidly suppressed, while strategic attackers exhibit characteristic oscillatory patterns that the cross-temporal integration mechanism exploits for detection. The committee ratio  $r = 0.5$  provides

Table 5: Ablation study under MinMax attack. F1: detection F1 for identifying malicious clients; Acc: credit scoring model accuracy (%) on test set.

Config.	UCI Credit (F1 / Acc)				Xinwang Bank (F1 / Acc)			
	IID	Label	Feat.	Qty.	IID	Label	Feat.	Qty.
Full	.77/ <b>82.6</b>	.74/ <b>82.0</b>	.70/ <b>81.3</b>	.67/ <b>80.7</b>	.78/ <b>83.8</b>	.75/ <b>83.2</b>	.72/ <b>82.5</b>	.69/ <b>81.9</b>
w/o S1	.69/81.8	.58/80.5	.54/79.8	.51/79.2	.71/83.0	.60/81.7	.56/81.0	.53/80.4
w/o S2	.58/80.5	.52/79.8	.49/79.1	.46/78.6	.60/81.7	.54/81.0	.51/80.3	.48/79.8
w/o S3	.74/82.3	.70/81.7	.68/81.0	.65/80.5	.76/83.5	.72/82.9	.70/82.2	.67/81.6
None	.00/78.2	.00/77.5	.00/76.8	.00/76.3	.00/79.4	.00/78.7	.00/78.0	.00/77.5

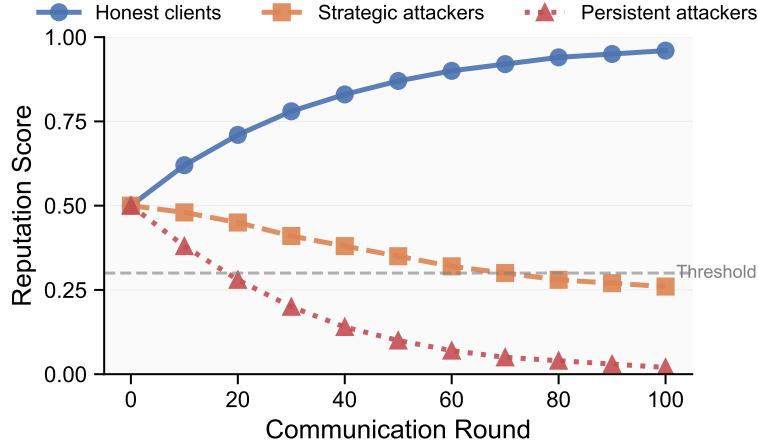


Figure 3: Reputation evolution over 100 federated rounds for three participant categories: honest clients converge to high scores ( $>0.9$ ), persistent attackers are rapidly suppressed ( $<0.2$ ), and strategic attackers exhibit oscillatory patterns exploited by cross-temporal detection.

an optimal balance between coverage and selectivity; the MAD coefficient  $k = 2.0$  offers a balanced precision-recall trade-off appropriate for security-critical financial applications. These findings confirm that HierFed achieves robust performance across diverse institutional configurations without requiring extensive hyperparameter tuning.

#### 4.4. Business Metrics and Scalability

Beyond detection accuracy, we evaluate the framework’s impact on credit decision quality using business-relevant metrics. Table 7 presents cost-sensitive evaluation assuming asymmetric misclassification costs (false negatives cost  $5\times$  more than false positives).

Table 6: Parameter sensitivity analysis on both datasets (label skew, MinMax).

Parameter	Value	UCI Credit			Xinwang Bank		
		P	R	F1	P	R	F1
Byzantine Ratio	20%	.70	.95	.81	.72	.96	.83
	30% (default)	.62	.90	.74	.64	.91	.75
	40%	.55	.82	.66	.57	.84	.68
	50%	.48	.70	.57	.50	.72	.59
Committee $r$	0.3 / 0.5 / 0.7	.69 / .73 / .74			.71 / .75 / .76		
MAD $k$	1.5 / 2.0 / 2.5	.67 / .73 / .74			.69 / .75 / .76		

Table 7: Decision quality metrics under MinMax attack (label skew). Cost ratio 5:1 (FN:FP).

Method	UCI Credit				Xinwang Bank			
	Loss	Appr.	Bad	KS	Loss	Appr.	Bad	KS
No Defense	1.42	78.2%	28.5%	0.38	1.38	79.5%	24.2%	0.41
Median	1.18	77.5%	24.1%	0.42	1.15	78.8%	20.8%	0.45
FLTrust	1.08	77.8%	22.3%	0.45	1.06	79.1%	19.5%	0.48
<b>HierFed</b>	<b>1.02</b>	78.1%	<b>21.2%</b>	<b>0.47</b>	<b>1.01</b>	79.3%	<b>18.6%</b>	<b>0.50</b>

Table 8: Scalability analysis: per-round overhead and detection performance.

$N$ (clients)	Time (seconds)			Detection		
	S1	S2	S3	P	R	F1
10	0.12	2.45	0.01	.62	.90	.74
20	0.48	2.52	0.02	.60	.88	.72
30	1.08	2.61	0.03	.58	.86	.70
50	3.02	2.78	0.05	.55	.83	.67

HierFed reduces expected loss compared to both undefended systems and FLTrust, while maintaining competitive approval rates. The improved KS statistic indicates enhanced discrimination between creditworthy and high-risk applicants.

To evaluate deployment feasibility for larger federations, we conduct scalability experiments varying the number of participating institutions from  $N = 10$  to  $N = 50$ . Table 8 reports the per-round computational overhead and detection performance.

Stage 1’s  $O(N^2d)$  similarity computation dominates overhead growth. For larger federations ( $N > 50$ ), we recommend distributed similarity compu-

Table 9: Reputation mechanism sensitivity (UCI Credit, label skew, MinMax).

$\alpha$	$\beta$	$\tau_\rho$	Exclusion (rounds)	Oscillation (rounds)	Recovery (rounds)
0.05	0.1	0.3	28	42	18
0.1	0.2	0.3	12	25	8
0.15	0.3	0.3	7	18	5
0.1	0.2	0.2	8	22	6
0.1	0.2	0.4	16	28	11

Table 10: Example audit log entry structure for round  $t = 50$  (illustrative).

Field	Client 1	Client 2	Client 3	...	Client 10
Committee member	✓	✓	×	...	✓
Stage 1 status	Trusted	Trusted	Uncertain	...	Suspicious
Stage 2 score	—	—	0.42	...	—
Final decision	Trusted	Trusted	Trusted	...	Excluded
Reputation $\rho^{(t)}$	1.82	1.75	1.21	...	0.18
Cumulative contrib.	47/50	46/50	38/50	...	12/50

tation or locality-sensitive hashing approximations. Detection performance degrades gracefully as federation size increases (F1: 0.74→0.67).

We further examine the sensitivity of the reputation mechanism to its key parameters:  $\alpha$  (reward rate),  $\beta$  (penalty rate), and exclusion threshold  $\tau_\rho$ . Table 9 reports the empirical results.

The results confirm that the reputation mechanism behaves as theoretically expected: higher  $\beta/\alpha$  ratios accelerate attacker exclusion. The default configuration ( $\alpha = 0.1$ ,  $\beta = 0.2$ ,  $\tau_\rho = 0.3$ ) provides a balanced trade-off between responsiveness and false positive tolerance.

**Governance outputs: Audit logs and credit footprints.** A distinguishing feature of HierFed is its generation of structured governance outputs beyond detection decisions. Table 10 illustrates the audit log structure generated per round, documenting committee composition, per-client anomaly scores, detection decisions, and reputation updates. These logs support three deployment requirements: (1) post-hoc investigation, wherein auditors can trace which clients contributed during affected rounds and examine their historical patterns when model performance degrades; (2) partner evaluation, wherein reputation trajectories provide quantitative evidence for admission or exclusion decisions in ongoing collaborations; and (3) incentive allocation, wherein contribution quality metrics enable fair reward distribution proportional to verified honest participation.

Figure 3 visualizes reputation trajectories over 100 rounds, demonstrating how honest clients converge to high scores while attackers are progressively excluded—this visualization itself constitutes a governance artifact for partner assessment.

Detection boundaries merit explicit discussion. IPM attacks represent fundamental detection limits when attacks become statistically indistinguishable from honest gradients. HierFed serves as the primary detection layer; for residual risk, we recommend combining detection with robust aggregators (e.g., geometric median), server-side validation, and human-in-the-loop escalation for high-stakes decisions.

## 5. Conclusion

This paper proposes HierFed, a hierarchical detection framework for Byzantine-resilient federated credit scoring that integrates graph-based committee election, contrastive autoencoder detection, and cross-temporal reputation tracking. Experiments on both public and commercial bank datasets demonstrate that HierFed achieves broader detection coverage than single-mechanism defenses while maintaining competitive model accuracy, with each stage targeting distinct attack categories that escape detection by other mechanisms.

Several limitations warrant acknowledgment: HierFed assumes a trusted central server, targets horizontal federated learning exclusively, and exhibits quadratic computational overhead in Stage 1 that constrains scalability beyond hundreds of participants. These boundaries suggest that HierFed is best suited for cross-institutional financial collaborations with moderate participant counts and trusted aggregation infrastructure.

Future research directions include extending HierFed to vertical federated learning settings, integrating differential privacy mechanisms for enhanced privacy guarantees, and investigating adaptive attacks that dynamically evolve evasion strategies over time.

## References

- L. C. Thomas, Credit scoring and its applications, SIAM Monographs on Mathematical Modeling and Computation (2017).
- S. Lessmann, B. Baesens, H.-V. Seow, L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European Journal of Operational Research 247 (2015) 124–136.

- B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.
- Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology 10 (2019) 1–19.
- P. Blanchard, E. M. El Mhamdi, R. Guerraoui, J. Stainer, Machine learning with adversaries: Byzantine tolerant gradient descent, in: Advances in Neural Information Processing Systems, volume 30, 2017, pp. 119–129.
- L. Lamport, R. Shostak, M. Pease, The byzantine generals problem, ACM Transactions on Programming Languages and Systems 4 (1982) 382–401.
- D. Yin, Y. Chen, R. Kannan, P. Bartlett, Byzantine-robust distributed learning: Towards optimal statistical rates, in: International Conference on Machine Learning, PMLR, 2018, pp. 5650–5659.
- E. M. El Mhamdi, R. Guerraoui, S. Rouault, The hidden vulnerability of distributed learning in byzantium, in: International Conference on Machine Learning, PMLR, 2018, pp. 3521–3530.
- X. Cao, M. Fang, J. Liu, N. Z. Gong, Fltrust: Byzantine-robust federated learning via trust bootstrapping, in: Network and Distributed Systems Security Symposium, 2021, pp. 1–18.
- G. Baruch, M. Baruch, Y. Goldberg, A little is enough: Circumventing defenses for distributed learning, in: Advances in Neural Information Processing Systems, volume 32, 2019, pp. 1871–1881.
- V. Shejwalkar, A. Houmansadr, Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning, in: Network and Distributed Systems Security Symposium, 2021, pp. 1–18.
- Z. Allen-Zhu, F. Ebrahimianghazani, J. Li, D. Alistarh, Byzantine-resilient non-convex stochastic gradient descent, in: International Conference on Learning Representations, 2021, pp. 1–25.

- M. Fang, X. Cao, J. Jia, N. Gong, Local model poisoning attacks to byzantine-robust federated learning, in: USENIX Security Symposium, 2020, pp. 1605–1622.
- C. Xie, O. Koyejo, I. Gupta, Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation, in: Uncertainty in Artificial Intelligence, PMLR, 2019, pp. 261–270.
- V. Tolpegin, S. Truex, M. E. Gursoy, L. Liu, Data poisoning attacks against federated learning systems, in: European Symposium on Research in Computer Security, Springer, 2020, pp. 480–501.
- E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 2938–2948.
- K. Pillutla, S. M. Kakade, Z. Harchaoui, Robust aggregation for federated learning, arXiv preprint arXiv:1912.13445 (2019).
- C. Xie, S. Koyejo, I. Gupta, Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance, in: International Conference on Machine Learning, PMLR, 2019, pp. 6893–6901.
- C. Fung, C. J. Yoon, I. Beschastnikh, Mitigating sybils in federated learning poisoning, arXiv preprint arXiv:1808.04866 (2020).
- T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Processing Magazine 37 (2020) 50–60.
- T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
- Z. Wang, J. Xiao, L. Wang, J. Yao, A novel federated learning approach with knowledge transfer for credit scoring, Decision Support Systems 177 (2024) 114084.
- F. Zheng, E. Erihe, K. Li, J. Tian, X. Xiang, A vertical federated learning method for interpretable scorecard and its application in credit scoring, arXiv preprint arXiv:2009.06218 (2020).

- J. Kang, Z. Xiong, D. Niyato, S. Xie, J. Zhang, Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory, *IEEE Internet of Things Journal* 6 (2019) 10700–10714.
- T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, *Proceedings of Machine Learning and Systems* 2 (2020) 429–450.
- K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, H. V. Poor, Federated learning with differential privacy: Algorithms and performance analysis, in: *IEEE Transactions on Information Forensics and Security*, volume 15, 2020, pp. 3454–3469.
- K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, Practical secure aggregation for privacy-preserving machine learning, in: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.