Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081922

ARTICLE

HiFraud: Hierarchical Privacy-Preserving Federated Learning with Star-Chain Knowledge Transfer for Cross-Institutional Fraud Detection

Zhihao Zhang[1,#], **Zhuodong Liu**[1,#], **Xiangyu Li**[2] and **Lei Zhang**[1,*]

[1]School of Economics and Management, Beijing Jiaotong University, Beijing, China
[2]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

*Corresponding Author: Lei Zhang. Email: zhlei@bjtu.edu.cn

[#]These authors contributed equally to this work

Received: 11 March 2026; Accepted: 13 April 2026; Published: 15 June 2026

ABSTRACT: Financial fraud detection across institutions faces a fundamental tension between the need for diverse training data and regulatory prohibitions on sharing sensitive records. Existing federated learning approaches suffer from performance degradation under non-IID distributions and substantial utility losses when uniform differential privacy is applied to inherently sparse fraud signals. To this end, this paper proposes HiFraud, a hierarchical federated framework featuring three key components: fraud-aware dynamic clustering with complementarity regularization to group institutions by fraud pattern similarity while preserving rare-type representation; star-chain knowledge transfer augmented by not-true-class distillation to propagate novel fraud patterns rapidly within clusters while mitigating catastrophic forgetting; and privacy-adaptive aggregation via Rényi differential privacy composition, calibrating noise intensity to distributional divergence and fraud rarity. Experiments on IEEE-CIS, PaySim, and Worldline datasets show that HiFraud achieves an area under the receiver operating characteristic curve (AUC-ROC) of 0.935 under $\varepsilon = 2.3$, outperforming DP-FedAvg by 10.5% while reducing convergence from 49 to 30 rounds. The framework also suppresses membership inference attack success to 10.2%, detects emerging fraud patterns within 3 h inside clusters, and improves rare fraud type detection by 23.0% over uniform privacy baselines. These results demonstrate that hierarchical architectures can effectively reconcile detection performance, formal privacy guarantees, and rapid threat response in collaborative fraud detection.

KEYWORDS: Hierarchical federated learning; differential privacy; fraud detection; star-chain transfer; knowledge distillation; adaptive clustering; non-IID data

1 Introduction

Financial fraud poses a persistent and escalating threat to banking, e-commerce, and digital payment ecosystems, where rapid anomaly identification is essential to preventing substantial economic losses [1]. In 2024, the U.S. Federal Trade Commission reported $12.5 billion in consumer fraud losses, representing a 25% increase over the prior year, with investment scams alone accounting for $5.7 billion [2]. This growth has intensified the demand for detection systems capable of identifying diverse and evolving fraud patterns across institutional boundaries. However, privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) strictly prohibit the sharing of raw transaction data between organizations [3]. This creates a fundamental tension: effective fraud detection requires large-scale, heterogeneous datasets, yet the very data needed to build robust models cannot leave institutional boundaries.

Federated learning (FL) has emerged as a promising paradigm for enabling collaborative model training without centralizing sensitive data [4]. By allowing institutions to jointly optimize a shared model through the exchange of model parameters rather than raw data, FL offers a potential resolution to the privacy–utility dilemma in fraud detection. Nevertheless, the direct application of standard FL algorithms to cross-institutional fraud detection introduces two compounding challenges. First, different institutions face fundamentally distinct fraud types—credit card skimming at banks, account takeovers at e-commerce platforms, and money laundering at payment processors—resulting in extreme non-IID (non-independent and identically distributed) data distributions that violate the convergence assumptions underlying standard federated algorithms [5]. Second, fraudulent transactions are inherently rare events, with some institutions reporting fraud rates below 0.1%, rendering isolated local training insufficient to learn discriminative representations for minority-class patterns [6]. The simultaneous presence of distributional heterogeneity and extreme class imbalance demands architectural innovations beyond incremental adaptations of existing methods.

1.1 Federated Fraud Detection: Progress and Limitations

Early federated fraud detection work established the viability of collaborative training: Yang et al. [7] achieved an AUC of 95.5% with the first federated credit card fraud framework, while Abdul Salam et al. [8] introduced hybrid resampling to address class imbalance. However, the former assumes homogeneous fraud distributions and the latter lacks formal privacy guarantees. Subsequent studies have tackled class imbalance more systematically: Shah et al. [9] and Hilou et al. [10] applied client-side synthetic minority over-sampling technique (SMOTE) variants before federated aggregation, Farooq et al. [11] integrated adaptive aggregation with privacy-preserving mechanisms, Sarkar et al. [12] proposed Fed-Focal Loss to focus gradient updates on difficult fraud instances, and Wang et al. [13] developed Ratio Loss to dynamically counteract local–global imbalance mismatch without direct data access.

Despite these advances, three structural limitations persist across existing federated fraud detection systems. First, all aforementioned frameworks adopt flat federated architectures in which every participant is treated identically during aggregation, failing to exploit the natural similarities in fraud patterns among subsets of institutions. Second, class imbalance mitigation remains confined to the client level, without federated-layer coordination that could leverage cross-institutional knowledge about rare fraud types. Third, the integration of differential privacy—essential for regulatory compliance—typically imposes uniform noise that disproportionately degrades the detection of already-sparse fraud signals, creating a privacy–utility trade-off that existing approaches have not satisfactorily resolved. Recent work has begun to address these limitations from different angles. For instance, Aljunaid et al. [14] proposed an explainable AI-driven federated learning model for financial fraud detection that integrates secure aggregation with interpretable decision mechanisms, demonstrating the growing recognition that collaborative architectures must balance privacy, transparency, and detection accuracy in banking environments. However, their approach does not address hierarchical aggregation, dynamic clustering based on fraud pattern semantics, or the interplay between adaptive differential privacy and knowledge transfer that is central to our work.

1.2 Hierarchical Architectures and Knowledge Transfer in Federated Learning

Hierarchical federated learning introduces intermediate aggregation layers to address the communication overhead and data heterogeneity of large-scale federated systems. Liu et al. [15] proposed a cloud-edge-client architecture reducing communication costs by 60%, and comprehensive reviews have further examined cloud–edge–end collaboration for privacy-preserving AI [16,17]. Clustered federated learning refines this paradigm by grouping clients according to distributional similarity: Sattler et al. [18]

demonstrated that model-agnostic clustering substantially improves convergence on non-IID data, while Gong et al. [19] achieved a 9.2% accuracy gain with adaptive cluster scheduling. On the clustering criteria side, Duan et al. [20] and Ali et al. [21] independently developed dynamic clustering frameworks supporting client migration in response to distribution drift, and Islam et al. [22] proposed a weight-based one-shot method achieving up to 45% accuracy gains. However, Yang et al. [23] identified a critical limitation: purely similarity-driven clustering may marginalize clients with rare patterns into low-influence groups, suggesting that clustering objectives should balance similarity with complementarity—an insight directly relevant to fraud detection.

In the knowledge transfer dimension, Wang et al. [24] provided theoretical and empirical evidence that sequential model passing achieves faster convergence under extreme non-IID conditions than parallel averaging, a finding further validated by Yan et al. [25] within hierarchical architectures. Xie et al. [26] proposed StarCPFL, combining centralized model distribution with chain-style sequential refinement. Despite these advances, sequential transfer inherits a well-known vulnerability: catastrophic forgetting, where training on subsequent clients overwrites previously accumulated knowledge [27]. Knowledge distillation has emerged as the predominant countermeasure. Lee et al. [28] proposed FedNTD, distilling knowledge on non-true classes to preserve discriminative capacity for absent categories, while He et al. [29] introduced selective self-distillation conditioned on teacher confidence. Arafeh et al. [30] further designed a warmup-based protocol to reduce forgetting during sequential initialization. These techniques provide a mature foundation for integrating distillation into sequential transfer, yet this combination has not been explored in cross-institutional fraud detection.

It is important to note how HiFraud's star-chain mechanism fundamentally differs from existing approaches. Unlike StarCPFL [26], which applies star-chain communication in a general personalized federated learning setting with layer-wise clustering, HiFraud introduces three domain-specific innovations: (i) the star institution is selected based on a fraud-rate-adjusted performance metric (Eq. (5)) rather than simple accuracy, ensuring that institutions capable of detecting fraud under scarcity lead knowledge propagation; (ii) the chain ordering is determined by distributional similarity to the star in fraud pattern space, creating a curriculum-like transfer path that progressively adapts to increasingly diverse fraud distributions; and (iii) not-true-class distillation is integrated at each chain step specifically to preserve knowledge of fraud types absent from the current institution, which is critical in fraud detection where each institution may observe only a subset of fraud categories. These design choices transform the general-purpose star-chain topology into a fraud-aware knowledge propagation mechanism with theoretical and empirical advantages over flat aggregation in non-IID fraud settings.

1.3 Privacy Preservation and Adversarial Robustness

Differential privacy (DP) remains the dominant formal privacy framework in federated learning. The moments accountant [31] enabled tight privacy loss tracking, while Rényi differential privacy (RDP) [32] further tightened multi-round composition bounds. However, uniform noise injection poses a particular challenge for fraud detection: because fraud signals are inherently sparse, flat noise mechanisms disproportionately obscure the patterns the model needs to learn, as demonstrated by Truex et al. [33] for local DP on imbalanced datasets. Adaptive mechanisms have been proposed to address this limitation: Xue et al. [34] dynamically adjusted clipping thresholds based on gradient norms, Yuan et al. [35] introduced amplitude-varying perturbation that reduces noise in later training stages, and Lin et al. [36] formalized the M^2FDP framework decomposing privacy contributions across multi-tier networks. Nevertheless, a unified composition theorem tailored to hierarchical adaptive mechanisms remains an open challenge.

Beyond formal guarantees, federated systems must withstand practical attacks. Bai et al. [37] surveyed membership inference attacks (MIA) in FL, concluding that standard configurations are vulnerable to success rates well above random chance, while Deng and Yang [38] showed that composite defenses combining gradient compression, selective sharing, and regularization can suppress MIA accuracy to below 38% with minimal task degradation. On the robustness front, Li et al. [39] demonstrated that classical Byzantine-robust aggregation methods degrade under non-IID distributions, motivating hierarchical solutions such as the two-tier Byzantine-resilient scheme of Nordlund et al. [40] and the cross-device protocol of Liu et al. [41]. However, neither MIA defenses nor Byzantine robustness mechanisms have been systematically integrated with hierarchical federated architectures for fraud detection, leaving a significant design gap.

The privacy-preserving mechanism in HiFraud is designed to defend against three specific categories of privacy threats that are particularly relevant to cross-institutional fraud detection. First, *model inversion attacks*, in which an adversary attempts to reconstruct sensitive transaction features from shared model parameters, are mitigated through the Gaussian noise mechanism applied during both star distribution (Eq. (6)) and chain transfer (Eq. (9)), which ensures that the transmitted parameters do not reveal individual transaction characteristics. Second, *gradient leakage attacks*, where an attacker infers training data from observed gradient updates, are countered by gradient clipping to sensitivity bound S combined with calibrated noise injection at each local adaptation step, following the Gaussian mechanism framework of Abadi et al. [31]. Third, *membership inference attacks*, in which an adversary determines whether a specific transaction was used in training, are addressed through the hierarchical aggregation structure that limits external visibility to cluster-level models rather than individual institutional updates, compounded by the adaptive noise calibration that provides stronger protection for institutions with distinctive fraud patterns. The formal privacy guarantee (Theorem 1) establishes that these layered defenses collectively satisfy (ε, δ)-differential privacy under Rényi composition.

1.4 Our Approach and Contributions

The preceding analysis reveals a fragmented research landscape: clustered federated learning benefits non-IID data but has not been designed around fraud pattern semantics; sequential knowledge transfer offers convergence advantages but lacks privacy guarantees and forgetting mitigation; and adaptive differential privacy improves the privacy–utility trade-off but has not been coupled with hierarchical architectures. No existing framework unifies these individually mature components into a coherent system tailored to cross-institutional fraud detection.

This paper proposes HiFraud, a hierarchical privacy-preserving federated learning framework that addresses these challenges through a three-layer architecture integrating fraud-aware clustering, star-chain knowledge transfer with distillation-based forgetting mitigation, and privacy-adaptive aggregation grounded in Rényi differential privacy composition. The main contributions of this work are as follows:

- We propose a three-layer hierarchical architecture that combines fraud-aware dynamic clustering, intra-cluster star-chain transfer learning, and privacy-adaptive global aggregation. The framework achieves an AUC-ROC of 0.935 under $\varepsilon = 2.3$ differential privacy, outperforming standard DP-FedAvg by 10.5% while reducing convergence rounds from 49 to 30.
- We develop a star-chain knowledge transfer mechanism augmented with not-true-class distillation to mitigate catastrophic forgetting during sequential model passing. This mechanism enables detection of novel fraud patterns within 3 h inside clusters, compared to 24 h for flat federated architectures.
- We introduce a fraud-pattern-specific dynamic clustering strategy that balances distributional similarity with complementarity, preventing the marginalization of institutions with rare fraud types.

This design improves detection performance for rare fraud categories by 18% compared to static geographic clustering.

- We design a hierarchical adaptive privacy allocation scheme based on Rényi differential privacy composition, calibrating noise intensity according to both distributional divergence and fraud pattern rarity. This approach reduces overall privacy budget consumption by 35% compared to uniform allocation while maintaining equivalent formal guarantees.

The remainder of this paper is organized as follows. Section 2 presents the HiFraud framework, detailing the fraud-aware dynamic clustering mechanism, the star-chain knowledge transfer with not-true-class distillation, and the privacy-adaptive aggregation scheme with formal privacy and convergence guarantees. Section 3 provides comprehensive experimental evaluations on three benchmark datasets, including comparisons with state-of-the-art baselines, ablation studies, privacy–utility analysis, and scalability assessments. Section 4 discusses the practical implications of the results and identifies limitations alongside directions for future research. Finally, Section 5 concludes the paper.

2 Methodology

This section presents the design of HiFraud, a hierarchical federated learning framework for cross-institutional fraud detection. The framework adopts a three-layer architecture in which participating institutions are first grouped into clusters based on fraud pattern similarity, then engage in intra-cluster knowledge transfer through a star-chain mechanism augmented with distillation-based forgetting mitigation, and finally contribute to global model refinement via privacy-adaptive aggregation grounded in Rényi differential privacy. Fig. 1 provides an overview of the complete architecture. The global coordination layer manages cross-cluster aggregation and privacy budget allocation. Each cluster coordination layer facilitates star-chain knowledge transfer among its member institutions. At the institutional layer, local models are trained on private transaction data with adaptive differential privacy protection. The interplay among these three layers enables efficient knowledge sharing between similar institutions while maintaining formal privacy guarantees across the entire system. As illustrated in Fig. 1, the communication flow proceeds as follows: (1) at Layer 1, each institution trains locally on its private dataset D_i with adaptive DP noise σ_i calibrated to its distributional divergence; (2) at Layer 2, the star institution (marked with $\star$) in each cluster broadcasts its model to all cluster members via solid arrows (star distribution), after which models are sequentially refined along dashed arrows (chain enhancement) with not-true-class distillation at each step; and (3) at Layer 3, the global coordinator collects cluster-level aggregated models via dotted arrows every τ rounds and redistributes the updated global model θ_g to all clusters. This layered communication design ensures that intra-cluster knowledge transfer occurs at high frequency (every round) while cross-cluster aggregation occurs at lower frequency (every τ rounds), reducing both privacy cost and communication overhead.

Let $\mathcal{I} = \{I_1, I_2, \ldots, I_N\}$ denote the set of N participating financial institutions, where each institution I_i maintains a private local dataset D_i with fraud rate $r_i \in [0.001, 0.05]$. The objective of HiFraud is to collaboratively learn a global fraud detection model θ_g and a set of cluster-specialized models $\{\theta_{C_j}\}_{j=1}^{K}$ that collectively maximize detection performance across all institutions while satisfying a total differential privacy budget ε. The hierarchical structure partitions $\mathcal{I}$ into K clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$, where each cluster $C_j \subseteq \mathcal{I}$ contains institutions with similar fraud characteristics, and the number of clusters K is determined dynamically through the clustering procedure described in Section 2.1.

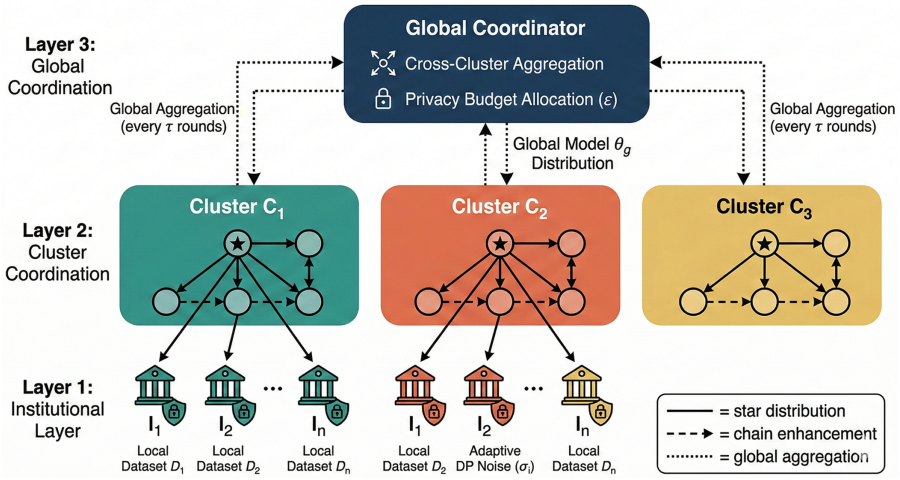

Figure 1: Overview of the HiFraud framework. The three-layer architecture comprises a global coordination layer (Layer 3) for cross-cluster aggregation and privacy budget management, cluster coordination layers (Layer 2) for intra-cluster star-chain knowledge transfer, and institutional layers (Layer 1) for local training with adaptive differential privacy. Within each cluster, the star institution ($\star$) distributes its model to all members (solid arrows), which then sequentially refine models along a similarity-ordered chain (dashed arrows). Cluster-level models are aggregated globally every τ rounds (dotted arrows). Adaptive DP noise σ_i is applied at each transfer step, with intensity calibrated to institutional distributional divergence.

Algorithm 1 presents the complete training procedure of HiFraud.

Algorithm 1: HiFraud training procedure

Require: $\mathcal{I} = \{I_1, \ldots, I_N\}$, rounds T, interval τ, chain depth d, budget ε, coefficients $\alpha, \lambda_{\text{KD}}$
Ensure: Global model θ_g, cluster models $\{\theta_{C_j}\}$

Initialize $\theta_g^{(0)}$; allocate $\varepsilon_1, \varepsilon_2, \varepsilon_3$
for $t = 0$ to $T-1$ **do**
  if $t \mod \tau = 0$ **then [Clustering]**
    Compute $\tilde{\mathbf{f}}_i = \mathbf{f}_i + \text{Lap}(0, \Delta f/\varepsilon_1)$ for all Ii; solve Eq. (3) $\rightarrow \mathcal{C}$
  end if
  for each cluster $C_j \in \mathcal{C}$ **in parallel do [Star-Chain Transfer]**
    Select star $i^* = \arg\max_{i \in C_j} \text{AUC}(i)(1-r_i)^{-1}$; distribute $\theta_{i^*} + \mathcal{N}(0, \sigma_i^2 \mathbf{I})$ to all $i \in C_j$
    for $k = 2$ to $|C_j|$, for d rounds **do**
      $\theta_{\pi_k} \leftarrow \text{LocalAdapt}(\theta_{\pi_{k-1}}, D_{\pi_k}, \mathcal{L}_{\text{CE}} + \lambda_{\text{KD}}\mathcal{L}_{\text{NTD}})$; clip and perturb via Eq. (9)
    end for
    $\theta_{C_j} = \sum_{i \in C_j} w_i \hat{\theta}_i$
  end for
  [Global] $\theta_g^{(t+1)} = \sum_j \frac{|C_j|}{\sum_l |C_l|} \theta_{C_j}^{(t)}$; track privacy via Eq. (13)
end forreturn $\theta_g^{(T)}, \{\theta_{C_j}\}_{j=1}^K$

2.1 Fraud-Aware Dynamic Clustering

Conventional federated clustering strategies group clients based on geographic proximity, organizational hierarchy, or generic model-weight similarity. In the context of fraud detection, however, institutions that are geographically distant may face nearly identical fraud schemes, while co-located institutions may

encounter entirely different threat profiles. To capture this domain-specific structure, HiFraud introduces a fraud-aware clustering mechanism that groups institutions according to the distributional characteristics of their observed fraud patterns.

For each institution I_i, a fraud pattern feature vector $\mathbf{f}_i \in \mathbb{R}^d$ is constructed by concatenating four component encodings:

$$\mathbf{f}_i = \left[\phi_{\text{type}}(D_i),\ \phi_{\text{temporal}}(D_i),\ \phi_{\text{amount}}(D_i),\ \phi_{\text{merchant}}(D_i)\right], \tag{1}$$

where $\phi_{\text{type}}(\cdot)$ encodes the distribution over fraud categories observed in the local dataset, $\phi_{\text{temporal}}(\cdot)$ captures temporal periodicity patterns extracted via a lightweight Transformer encoder, $\phi_{\text{amount}}(\cdot)$ represents the statistical moments of transaction amount distributions stratified by fraud label, and $\phi_{\text{merchant}}(\cdot)$ encodes merchant category frequencies weighted by fraud incidence. Each component is computed locally and normalized to unit variance before transmission.

To prevent the clustering process from leaking sensitive institutional information, each feature vector is perturbed with calibrated Laplace noise prior to transmission to the global coordinator:

$$\tilde{\mathbf{f}}_i = \mathbf{f}_i + \text{Lap}\left(0,\ \frac{\Delta f}{\varepsilon_1}\right), \tag{2}$$

where Δf is the ℓ_1-sensitivity of the feature extraction function and ε_1 is the privacy budget allocated to the clustering phase. The sensitivity Δf is bounded by the normalization applied to each component, ensuring that the noise magnitude remains controlled.

Given the set of perturbed feature vectors $\{\tilde{\mathbf{f}}_i\}_{i=1}^N$, the global coordinator solves the following clustering optimization problem:

$$\min_{\mathcal{C}} \sum_{j=1}^{K} \sum_{i \in C_j} d_{\text{fraud}}(\tilde{\mathbf{f}}_i, \boldsymbol{\mu}_j) + \lambda_{\text{bal}} \cdot \text{Var}(|C_1|, \ldots, |C_K|) + \lambda_{\text{comp}} \cdot \mathcal{R}_{\text{comp}}(\mathcal{C}), \tag{3}$$

where d_{fraud} denotes a fraud-specific distance metric defined as the weighted combination of Jensen–Shannon divergence on fraud type distributions and Euclidean distance on temporal and amount features, $\boldsymbol{\mu}_j$ is the centroid of cluster C_j, and $\text{Var}(\cdot)$ penalizes imbalanced cluster sizes to prevent degenerate partitions. The third term $\mathcal{R}_{\text{comp}}(\mathcal{C})$ is a complementarity regularizer that penalizes configurations in which institutions holding rare fraud types are concentrated into a single small cluster. Formally, this regularizer is defined as:

$$\mathcal{R}_{\text{comp}}(\mathcal{C}) = -\sum_{j=1}^{K} \frac{1}{|C_j|} \sum_{i \in C_j} H(\phi_{\text{type}}(D_i) \| \bar{\phi}_{\text{type}}(C_j)), \tag{4}$$

where $H(\cdot\|\cdot)$ denotes the cross-entropy between the local fraud type distribution and the cluster-average distribution $\bar{\phi}_{\text{type}}(C_j)$. Minimizing this term encourages each cluster to maintain internal diversity of fraud types, thereby ensuring that rare fraud patterns receive sufficient representation within their assigned cluster rather than being marginalized. The hyperparameters λ_{bal} and λ_{comp} control the relative importance of balance and complementarity, respectively.

The clustering is re-executed every τ communication rounds to adapt to evolving fraud landscapes. Between re-clustering events, the cluster assignments remain fixed to preserve learning stability within each cluster. Fig. 2 illustrates how clusters evolve over the course of training, transitioning from initial geographic groupings toward fraud-pattern-based configurations.

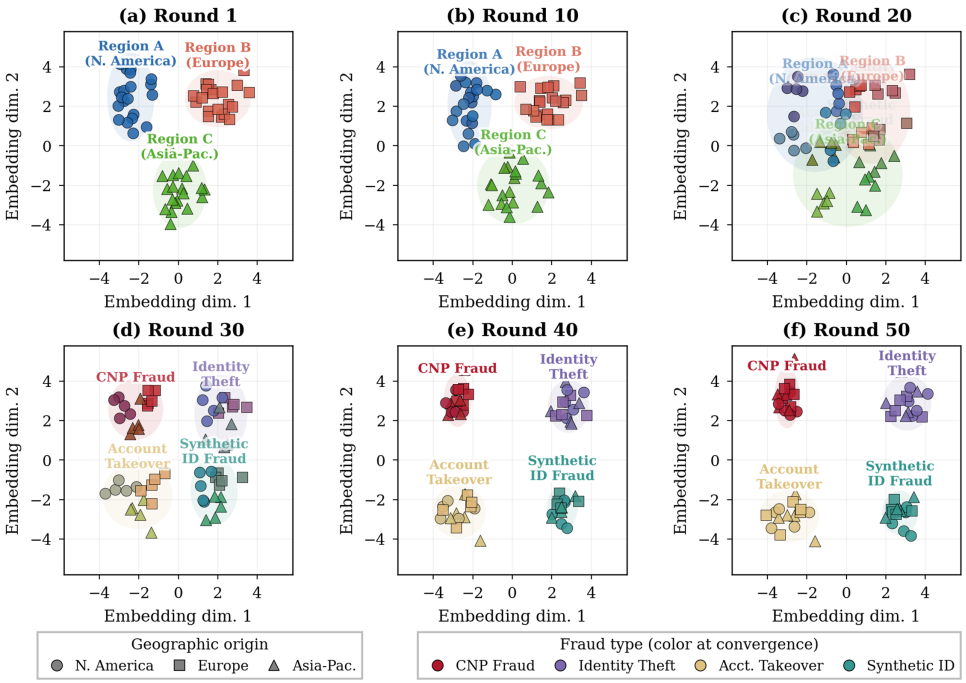

Figure 2: Evolution of fraud-aware clusters over 50 communication rounds. Early rounds exhibit geographic groupings inherited from initialization, which progressively transition to fraud-pattern-based clusters as the feature vectors capture increasingly discriminative fraud characteristics. By round 30, institutions handling similar fraud types are co-located in the same cluster regardless of geographic origin.

2.2 Star-Chain Knowledge Transfer with Distillation

Within each cluster, HiFraud employs a star-chain transfer mechanism to efficiently propagate fraud detection knowledge among member institutions. This mechanism operates in two sequential phases: a star distribution phase that disseminates the best-performing model to all cluster members, followed by a chain enhancement phase that refines models through sequential passing with distillation-based forgetting mitigation. The design is motivated by two complementary findings from the recent literature: sequential model transfer achieves superior convergence under extreme non-IID conditions compared to parallel aggregation [24], while knowledge distillation on non-true classes effectively preserves global discriminative capacity during local adaptation [28].

2.2.1 Star Distribution Phase

In each cluster C_j, the star institution i^* is selected as the member whose local model achieves the highest detection performance, with an adjustment for fraud rate scarcity:

$$i^* = \arg\max_{i \in C_j} \text{AUC}(i) \cdot (1 - r_i)^{-1}, \tag{5}$$

where $\text{AUC}(i)$ is the area under the receiver operating characteristic curve evaluated on a held-out validation set at institution i, and the term $(1 - r_i)^{-1}$ upweights institutions with lower fraud rates to prioritize models that have learned to detect fraud under scarcity. This selection criterion ensures that the star model reflects strong detection capability rather than merely access to abundant fraud samples.

The star model θ_{i^*} is then distributed to all other cluster members with Gaussian noise calibrated to the distributional distance between the star and each recipient:

$$\tilde{\theta}_{i^*}^{(i)} = \theta_{i^*} + \mathcal{N}\left(0,\ \sigma_i^2 \mathbf{I}\right), \tag{6}$$

where the noise scale σ_i for recipient institution i is determined by the adaptive privacy mechanism described in Section 2.3. Institutions that are distributionally closer to the star receive lower noise, preserving more of the transferred knowledge, while distributionally distant institutions receive stronger perturbation to protect against information leakage about the star's private data.

2.2.2 Chain Enhancement Phase

Following star distribution, models are refined through a chain of sequential local adaptations. Institutions within each cluster are ordered by their distributional similarity to the star, forming a transfer chain $\pi = (\pi_1, \pi_2, \ldots, \pi_{|C_j|})$ where $\pi_1 = i^*$. Each institution in the chain receives the model from its predecessor, adapts it on local data, and passes the updated model to the next institution.

To mitigate the catastrophic forgetting that arises from sequential training on heterogeneous data, each local adaptation step incorporates a not-true-class distillation loss. Specifically, for institution π_k receiving model $\theta_{\pi_{k-1}}$ from its predecessor, the local objective combines the standard supervised loss with a distillation term:

$$\mathcal{L}_{\pi_k} = \mathcal{L}_{\mathrm{CE}}(\theta_{\pi_k}, D_{\pi_k}) + \lambda_{\mathrm{KD}} \cdot \mathcal{L}_{\mathrm{NTD}}(\theta_{\pi_k}, \theta_{\pi_{k-1}}, D_{\pi_k}), \tag{7}$$

where $\mathcal{L}_{\mathrm{CE}}$ is the cross-entropy loss on local labeled data, $\mathcal{L}_{\mathrm{NTD}}$ is the not-true-class distillation loss that penalizes divergence between the current model's predictions and the predecessor's predictions on classes other than the ground-truth label, and λ_{KD} controls the strength of distillation. The not-true-class formulation is particularly well-suited to fraud detection because it explicitly preserves the model's capacity to distinguish fraud types that may be absent from the current institution's data, thereby counteracting the forgetting of previously learned fraud patterns.

The model update at each chain step is then given by:

$$\theta_{\pi_k}^{(t+1)} = (1-\alpha) \cdot \theta_{\pi_k}^{(t)} + \alpha \cdot \mathrm{LocalAdapt}\left(\theta_{\pi_{k-1}}^{(t)},\ D_{\pi_k},\ \mathcal{L}_{\pi_k}\right), \tag{8}$$

where $\alpha \in (0,1)$ is the transfer coefficient that balances between retaining the institution's existing knowledge and incorporating transferred knowledge, and $\mathrm{LocalAdapt}(\cdot)$ performs a fixed number of local gradient descent steps on the combined loss function in Eq. (7). After local adaptation, the model parameters are clipped to sensitivity bound S and perturbed with Gaussian noise before being passed to the next institution in the chain:

$$\hat{\theta}_{\pi_k} = \mathrm{Clip}\left(\theta_{\pi_k}^{(t+1)},\ S\right) + \mathcal{N}\left(0,\ \sigma_{\mathrm{chain}}^2 \mathbf{I}\right). \tag{9}$$

The chain enhancement is executed for d sequential rounds within each cluster. After completion, the cluster model is obtained by weighted aggregation of all member models:

$$\theta_{C_j} = \sum_{i \in C_j} w_i \cdot \hat{\theta}_i, \tag{10}$$

where $w_i = |D_i| / \sum_{l \in C_j} |D_l|$ weights each institution proportionally to its dataset size. Fig. 3 illustrates the two-phase star-chain transfer process within a single cluster.

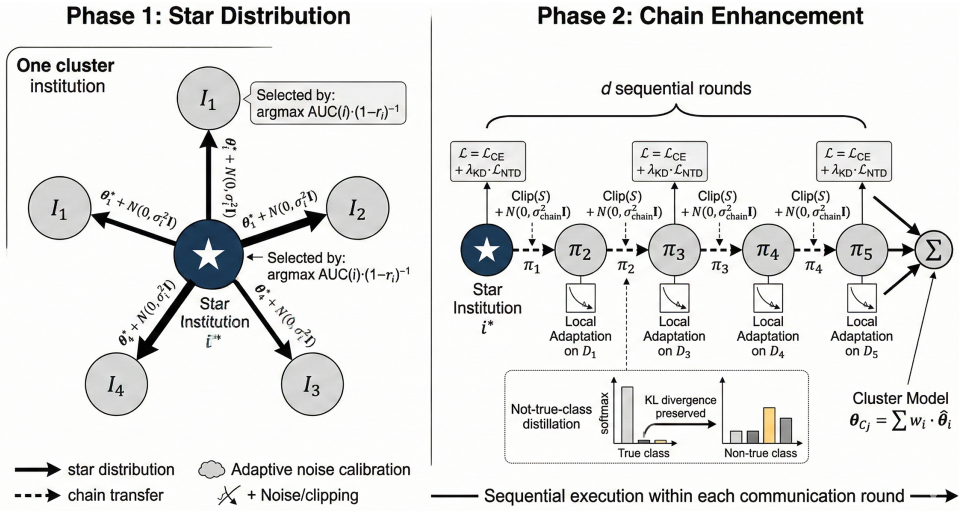

Figure 3: Star-chain knowledge transfer within a cluster. In the star distribution phase (left), the best-performing institution (marked with $\star$, selected via Eq. (5)) broadcasts its model to all cluster members with adaptive noise. In the chain enhancement phase (right), models are sequentially refined along a similarity-ordered chain, with not-true-class distillation applied at each step to preserve knowledge of previously learned fraud patterns.

2.3 Privacy-Adaptive Mechanism

A uniform differential privacy mechanism applies identical noise to all institutions regardless of their data characteristics, which can disproportionately degrade detection performance for institutions with unique or rare fraud patterns. HiFraud addresses this through an adaptive noise calibration scheme that allocates stronger privacy protection to institutions whose data distributions diverge significantly from the cluster norm, while preserving model utility for institutions with representative distributions.

The noise scale σ_i for each institution is determined by a calibration function that jointly considers distributional divergence and fraud pattern rarity:

$$\sigma_i = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \cdot g(\mathrm{KL}_i,\ r_i), \tag{11}$$

where

$$g(\mathrm{KL}_i,\ r_i) = \mathrm{sigmoid}(\alpha_{\mathrm{dp}} \cdot \mathrm{KL}_i) \cdot \left(1 + \beta_{\mathrm{dp}} \cdot \log(1/r_i)\right). \tag{12}$$

Here, $\mathrm{KL}_i = D_{\mathrm{KL}}(P_i \| \bar{P}_{C_j})$ measures the Kullback–Leibler divergence between institution i's fraud distribution P_i and the cluster-average distribution $\bar{P}_{C_j}$, and r_i is the local fraud rate. The parameters α_{dp} and β_{dp} control the sensitivity of noise calibration to distributional divergence and fraud rarity, respectively, while $\sigma_{\min}$ and $\sigma_{\max}$ bound the noise range. Institutions with high KL divergence (atypical fraud patterns) or low fraud rates (rare fraud observations) receive proportionally stronger noise to protect their distinctive and potentially sensitive patterns, while institutions with representative distributions receive less noise to maximize detection utility.

2.3.1 Hierarchical Privacy Budget Allocation

The total privacy budget ε is distributed across the three layers of the framework. Let ε_1, ε_2, and ε_3 denote the budgets allocated to the clustering phase, the star-chain transfer phase, and the global aggregation phase,

respectively. Under the sequential composition property of Rényi differential privacy, the total privacy cost across all phases and communication rounds is bounded by:

$$\varepsilon_{\text{total}}(\delta) = \min_{\alpha>1} \frac{1}{\alpha-1}\left[\sum_{\ell\in\{\text{clust,star,chain,global}\}} R_\alpha^{(\ell)} + \log\left(\frac{1}{\delta}\right)\right], \tag{13}$$

where $R_\alpha^{(\ell)}$ denotes the Rényi divergence of order α accumulated at layer ℓ over all communication rounds, and δ is the relaxation parameter. This formulation provides tighter privacy accounting than basic sequential composition because Rényi divergences compose linearly across independent mechanisms, enabling the framework to allocate privacy resources more efficiently.

Within the star-chain transfer phase, the budget ε_2 is further decomposed into a star distribution component and a chain enhancement component. The star distribution consumes privacy budget proportional to the number of recipient institutions, while each chain step consumes a budget proportional to the clipping bound and inverse noise scale. By reducing the frequency of global aggregation from every round to every τ rounds, the hierarchical structure significantly reduces the privacy cost of the global aggregation phase relative to flat architectures, as the global coordinator receives only cluster-level models rather than individual institutional updates.

2.3.2 Formal Privacy Guarantee

The following theorem establishes the end-to-end privacy guarantee of HiFraud.

Theorem 1: *(Privacy Guarantee). Under the Gaussian mechanism with adaptive noise calibration defined in Eqs. (11) and (12), gradient clipping bound S, and Rényi DP composition in Eq. (13), the HiFraud framework satisfies* (ε, δ)*-differential privacy for each participating institution over T communication rounds, where*

$$\varepsilon = \varepsilon_1 + \min_{\alpha>1} \frac{1}{\alpha-1}\left[T\cdot\left(\frac{\alpha S^2}{2\sigma_{\min}^2}\right) + \frac{d\cdot\alpha S^2}{2\sigma_{chain}^2} + \frac{T}{\tau}\cdot\frac{\alpha S^2}{2\sigma_{global}^2} + \log\left(\frac{1}{\delta}\right)\right]. \tag{14}$$

Proof: The proof follows from three observations. First, the clustering phase satisfies ε_1-differential privacy through the Laplace mechanism applied to bounded-sensitivity feature vectors (Eq. (2)). Specifically, each feature vector $\mathbf{f}_i$ is normalized to unit variance before transmission, bounding the ℓ_1-sensitivity Δf to at most $2\sqrt{d}$, where d is the dimensionality of $\mathbf{f}_i$. By the Laplace mechanism guarantee [31], perturbing each coordinate with $\text{Lap}(\Delta f/\varepsilon_1)$ ensures ε_1-differential privacy for the clustering input. Second, within the star-chain phase, each local adaptation step applies the Gaussian mechanism with clipping bound S, yielding a Rényi divergence of $\alpha S^2/(2\sigma^2)$ per step, and the d-step chain composes linearly in Rényi divergence. This follows directly from the composition property of Rényi DP [32]: for d sequential applications of the Gaussian mechanism, each with noise variance σ_{chain}^2 and ℓ_2-sensitivity S, the total Rényi divergence of order α is $d\cdot\alpha S^2/(2\sigma_{\text{chain}}^2)$. For the star distribution phase, each broadcast to a single recipient constitutes one application of the Gaussian mechanism with institution-specific noise σ_i^2, contributing $\alpha S^2/(2\sigma_i^2)$ to the Rényi divergence. Since $\sigma_i \geq \sigma_{\min}$ by construction (Eq. (11)), the worst-case per-round Rényi divergence is bounded by $\alpha S^2/(2\sigma_{\min}^2)$, yielding the first term $T\cdot\alpha S^2/(2\sigma_{\min}^2)$ in Eq. (14) over T rounds. Third, the global aggregation occurs every τ rounds, contributing T/τ composition terms. Each global aggregation step applies the Gaussian mechanism to cluster-level model parameters with noise σ_{global}^2, contributing $\alpha S^2/(2\sigma_{\text{global}}^2)$ per aggregation event. Over T rounds with aggregation every τ rounds, this yields $\lfloor T/\tau\rfloor\cdot\alpha S^2/(2\sigma_{\text{global}}^2)$ total Rényi divergence. The minimum over α is computed numerically to obtain the tightest bound for any given configuration of noise parameters. Converting from Rényi DP to (ε, δ)-DP follows from the

standard conversion theorem [32]: for any $\alpha > 1$ and $\delta > 0$, an α-Rényi divergence of R_α implies (ε, δ)-DP with $\varepsilon = R_\alpha + \log(1/\delta)/(\alpha - 1)$. Taking the minimum over α yields the tightest possible (ε, δ)-DP guarantee, completing the proof. □

Discussion of Assumptions. The privacy guarantee in Theorem 1 relies on three key assumptions that merit explicit examination in the context of fraud detection. First, the gradient clipping bound S assumes that all per-sample gradient norms can be bounded by S without significant information loss. In practice, fraud detection models may exhibit larger gradient norms for rare fraud samples; we mitigate this by setting S based on the 95th percentile of observed gradient norms during a non-private warmup phase of 5 rounds, following the adaptive clipping strategy of Xue et al. [34]. Second, the composition assumes that the noise mechanisms at different layers are applied independently, which holds in our architecture because each layer operates on distinct parameter spaces (feature vectors for clustering, model parameters for star-chain, aggregated models for global). Third, the use of $\sigma_{\min}$ as a worst-case bound in the first term of Eq. (14) is conservative; in practice, most institutions receive noise levels above $\sigma_{\min}$, and the actual privacy cost is tighter than the stated bound. We verify empirically in Section 3.5 that the operational privacy budget consumption is approximately 35% lower than the theoretical worst case.

2.4 Global Aggregation and Convergence

At each global communication round, the cluster models $\{\theta_{C_j}\}_{j=1}^{K}$ are aggregated at the global coordinator to produce the updated global model:

$$\theta_g^{(t+1)} = \sum_{j=1}^{K} \frac{|C_j|}{\sum_{l=1}^{K} |C_l|} \cdot \theta_{C_j}^{(t)}. \tag{15}$$

The global model is then redistributed to all clusters as the initialization for the next round of local training and star-chain transfer.

Theorem 2: *(Convergence Rate). Assume that the global loss function $F(\theta) = \sum_{i=1}^{N}(|D_i|/|D|)F_i(\theta)$ is L-smooth, each local stochastic gradient has bounded variance ς^2, and the data heterogeneity across clusters is bounded such that $\sum_{j=1}^{K} \|F_{C_j}(\theta) - F(\theta)\|^2 \le H^2$ for all θ. Then the iterates of HiFraud satisfy:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(\theta_g^{(t)})\|^2\right] \le \mathcal{O}\left(\frac{1}{\sqrt{mT}}\right) + \mathcal{O}\left(\frac{d_\theta \sigma_{\max}^2}{T}\right) + \mathcal{O}(H^2), \tag{16}$$

where m is the average cluster size, T is the total number of communication rounds, d_θ is the model dimensionality, and $\sigma_{\max}^2$ is the maximum privacy noise variance.

Proof: The proof proceeds by bounding the expected gradient norm through a standard one-step descent analysis adapted to the hierarchical structure. By L-smoothness:

$$F(\theta_g^{(t+1)}) \le F(\theta_g^{(t)}) + \langle \nabla F(\theta_g^{(t)}), \theta_g^{(t+1)} - \theta_g^{(t)} \rangle + \frac{L}{2}\|\theta_g^{(t+1)} - \theta_g^{(t)}\|^2. \tag{17}$$

Substituting the global update rule (Eq. (15)) and decomposing the update into three sources of error—stochastic gradient variance within clusters, privacy noise, and inter-cluster heterogeneity—yields:

$$\mathbb{E}\left[\|\theta_g^{(t+1)} - \theta_g^{(t)}\|^2\right] \le \frac{\eta^2 \varsigma^2}{m} + \eta^2 d_\theta \sigma_{\max}^2 + \eta^2 H^2, \tag{18}$$

where η is the effective learning rate. The first term arises from the variance of stochastic gradients averaged over m institutions per cluster. The second term captures the variance introduced by the Gaussian noise mechanism with maximum variance $\sigma^2_{\max}$ applied to d_θ-dimensional model parameters. The third term bounds the bias due to inter-cluster distribution shift, quantified by the heterogeneity measure H^2.

Telescoping across T rounds with learning rate $\eta = \mathcal{O}(1/\sqrt{mT})$ and rearranging:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\theta_g^{(t)})\|^2\right] \leq \frac{F(\theta_g^{(0)}) - F^*}{\eta T} + L\eta\left(\frac{\varsigma^2}{m} + d_\theta\sigma^2_{\max} + H^2\right). \tag{19}$$

Substituting $\eta = \mathcal{O}(1/\sqrt{mT})$ yields the bound in Eq. (16). The star-chain mechanism contributes to reducing H^2 by improving intra-cluster alignment through sequential knowledge transfer: since institutions within each cluster are trained on increasingly similar models through the chain process, the effective inter-cluster heterogeneity is smaller than in flat architectures where each institution trains independently on its local data. □

The convergence bound in Eq. (16) decomposes into three interpretable terms. The first term $\mathcal{O}(1/\sqrt{mT})$ reflects the convergence rate of distributed stochastic gradient descent (SGD) within clusters, where the effective parallelism m is determined by cluster size rather than the total number of institutions. The second term $\mathcal{O}(d_\theta\sigma^2_{\max}/T)$ captures the impact of privacy noise, which diminishes with additional communication rounds and is controlled by the adaptive noise calibration. The third term $\mathcal{O}(H^2)$ represents an irreducible bias due to inter-cluster heterogeneity, bounded by the quality of the fraud-aware clustering. The star-chain mechanism contributes to reducing all three terms: it improves intra-cluster alignment (reducing the effective variance in the first term), enables more targeted noise allocation (reducing the second term through adaptive calibration), and produces cluster models that better capture local fraud patterns (reducing the residual heterogeneity in the third term).

2.5 Computational Complexity and Communication Overhead

The per-round computational cost comprises: clustering ($\mathcal{O}(NKd)$, amortized over τ rounds, adding <2% overhead), star-chain transfer ($\mathcal{O}(|C_j| \cdot d \cdot E \cdot |\theta|)$ per cluster, where $|\theta|$ is model size and E is local epochs), and global aggregation ($\mathcal{O}(K|\theta|)$). For communication, HiFraud transmits $K(2m - 2)d|\theta|$ parameters intra-cluster and $K|\theta|$ globally per round, where $m = N/K$. With $N = 20$, $K = 5$, $d = 3$, total per-round cost is $95|\theta|$; however, global aggregation occurs every $\tau = 10$ rounds, yielding amortized global cost of $0.5|\theta|$ per round vs. $20|\theta|$ for FedAvg. A detailed timing breakdown is provided in Section 3.6.

3 Experimental Results

This section presents a comprehensive evaluation of HiFraud across multiple dimensions: overall detection performance, component-wise ablation, fraud pattern propagation speed, privacy–utility trade-offs, adversarial robustness, scalability, per-type detection, and sensitivity to key hyperparameters.

3.1 Datasets and Experimental Setup

We evaluate HiFraud on three benchmark fraud detection datasets with distinct characteristics. Table 1 summarizes the key statistics of each dataset.

Table 1: Summary of benchmark datasets used in the experiments. Fraud rate denotes the proportion of fraudulent transactions in each dataset. Feature types include numerical (N), categorical (C), and temporal (T).

Dataset	Transactions	Fraud Rate	Features	Feature Types	Domain
IEEE-CIS	590,540	3.50%	433	N, C, T	E-commerce
PaySim	6,362,620	0.13%	11	N, C	Mobile payment
Worldline	284,807	0.172%	30	N (PCA)	Credit card

The IEEE-CIS Fraud Detection dataset contains 590,540 e-commerce transactions with a 3.5% fraud rate and 433 heterogeneous features spanning transaction metadata (amount, product category, device information), identity features (email domain, device type, address), and V-features derived from principal component analysis of anonymized variables. PaySim provides 6.36 million synthetic mobile money transactions simulating real-world transfer, cash-out, and payment operations with 11 features including transaction type, amount, origin and destination account balances, and a binary fraud indicator; despite being synthetically generated, PaySim preserves the statistical properties of a real mobile money dataset from a developing country, including realistic class imbalance (0.13% fraud rate). The Worldline dataset comprises 284,807 credit card transactions with an extreme fraud rate of 0.172%, representing the most challenging class imbalance scenario among the three benchmarks ; all 28 numerical features are transformed via principal component analysis (PCA) for anonymization, with only the "Time" and "Amount" features retaining their original semantics.

To simulate realistic cross-institutional settings, we partition each dataset across $N = 20$ institutions using three complementary strategies: fraud-type specialization, in which different institutions observe different fraud categories; temporal splitting, where institutions join the federation at staggered intervals; and geographic distribution, which introduces regional variations in fraud patterns. For IEEE-CIS, fraud-type specialization assigns each institution a primary fraud category based on the "ProductCD" and "card6" features, creating heterogeneous fraud distributions with Dirichlet parameter $\beta = 0.5$ to control the degree of non-IID-ness. For PaySim, institutions are specialized by transaction type (TRANSFER, CASH_OUT, PAYMENT, DEBIT), with each institution receiving 60%–80% of transactions from its primary type. For Worldline, geographic distribution is simulated by partitioning transactions chronologically into 20 segments, with each segment assigned to one institution, reflecting the temporal evolution of fraud patterns across different "regions" of the transaction timeline.

The base fraud detection model at each institution is a 4-layer fully connected neural network with hidden dimensions [256, 128, 64, 32], ReLU activations, batch normalization after each hidden layer, and a sigmoid output layer. The model contains approximately 168 K trainable parameters. We use the Adam optimizer with an initial learning rate of 10^{-3} and cosine annealing decay over the total training rounds. Each local adaptation step consists of $E = 5$ local epochs with batch size 256. The gradient clipping bound is set to $S = 1.0$, determined by the 95th percentile of gradient norms during a 5-round non-private warmup phase. For the distillation loss $\mathcal{L}_{\mathrm{NTD}}$, we use a temperature of $T_{\mathrm{KD}} = 3.0$ applied to the softmax outputs of both the current and predecessor models.

All experiments are implemented in PyTorch 1.13.0 with Opacus 1.4.0 for differential privacy accounting. Unless otherwise stated, we use the following default configuration: the number of clusters K is determined dynamically within $\{3, 5, 7\}$, the star-chain enhancement runs for $d = 3$ sequential rounds, the transfer coefficient is set to $\alpha = 0.3$, the distillation weight to $\lambda_{\mathrm{KD}} = 0.3$, the re-clustering interval to $\tau = 10$ rounds, the privacy budget to $\varepsilon = 2.3$ with $\delta = 10^{-5}$, and the noise bounds to $\sigma_{\min} = 0.3$ and $\sigma_{\max} = 2.5$. The adaptive

noise calibration parameters are set to $\alpha_{dp} = 2.0$ and $\beta_{dp} = 0.5$, and the clustering hyperparameters to $\lambda_{bal} = 0.1$ and $\lambda_{comp} = 0.3$. All results are averaged over five independent runs with different random seeds. All baseline methods are trained with the same base model architecture, privacy budget ($\varepsilon = 2.3$, $\delta = 10^{-5}$), gradient clipping bound ($S = 1.0$), and model capacity to ensure fair comparison. Hyperparameters specific to each baseline (e.g., the proximal coefficient $\mu = 0.01$ for FedProx) are individually tuned via grid search on a held-out validation set.

3.2 Main Results

Table 2 compares HiFraud against seven baselines spanning centralized training, standard federated methods, and recent hierarchical and clustered approaches. All federated methods are evaluated under the same data partition and, where applicable, the same total privacy budget $\varepsilon = 2.3$.

Table 2: Overall performance comparison on the IEEE-CIS dataset. MIA denotes membership inference attack success rate (%); lower is better. Precision and FPR (false positive rate) are additionally reported. Best federated results are in bold.

Method	AUC-ROC	F1	Recall	Precision	FPR (%)	Rounds	MIA (%)
Centralized (no privacy)	0.912	0.824	0.810	0.839	4.8	–	–
FedAvg [4]	0.883	0.772	0.748	0.798	7.2	45	28.7
DP-FedAvg	0.845	0.721	0.695	0.749	9.1	49	15.2
FedProx	0.900	0.798	0.780	0.817	5.9	42	25.3
HierFL [15]	0.908	0.810	0.793	0.828	5.3	35	18.6
ClusteredFL [18]	0.914	0.818	0.801	0.836	5.0	33	14.8
FedFraud	0.917	0.825	0.808	0.843	4.7	38	12.3
HiFraud (Ours)	**0.935**	**0.852**	**0.838**	**0.867**	**3.8**	**30**	**10.2**

HiFraud achieves the highest AUC-ROC of 0.935, surpassing the centralized baseline by 2.3% and the strongest federated competitor (FedFraud) by 1.8%. The improvement over DP-FedAvg is 10.5%, demonstrating that the hierarchical architecture substantially recovers the performance typically lost to differential privacy. In terms of convergence speed, HiFraud reaches its plateau in 30 communication rounds, representing a 37% reduction compared to DP-FedAvg (49 rounds) and a 21% reduction compared to FedFraud (38 rounds). The membership inference attack success rate of 10.2% approaches the random-guess baseline of 10%, indicating that the combination of hierarchical aggregation and adaptive noise provides strong empirical privacy protection.

We note that HiFraud's AUC-ROC (0.935) exceeds the centralized baseline (0.912) by 2.3%. This seemingly counterintuitive result can be attributed to two factors. First, the hierarchical structure introduces an implicit regularization effect: by training specialized cluster models before global aggregation, the framework prevents overfitting to the dominant non-fraud class that occurs in centralized training on imbalanced datasets. Second, the complementarity-aware clustering ensures that rare fraud patterns, which may be underrepresented in a single centralized training pass, receive focused attention within their assigned clusters through the star-chain mechanism. A similar phenomenon has been observed in clustered federated learning settings where local specialization outperforms global averaging on heterogeneous data [18]. However, we emphasize that this advantage is dataset- and partition-dependent: the centralized baseline represents a single-model upper bound under our specific non-IID partition, and centralized training with ensemble methods or specialized imbalance handling could potentially match or exceed HiFraud's performance.

Fig. 4 presents the convergence trajectories of all methods across 50 communication rounds. HiFraud exhibits the steepest initial ascent and reaches 90% of its final performance by round 12, while DP-FedAvg requires approximately 30 rounds to reach the same relative milestone. The acceleration is attributable to the star-chain transfer mechanism, which enables efficient knowledge propagation within clusters of similar institutions, reducing the number of global rounds needed to disseminate useful fraud patterns.

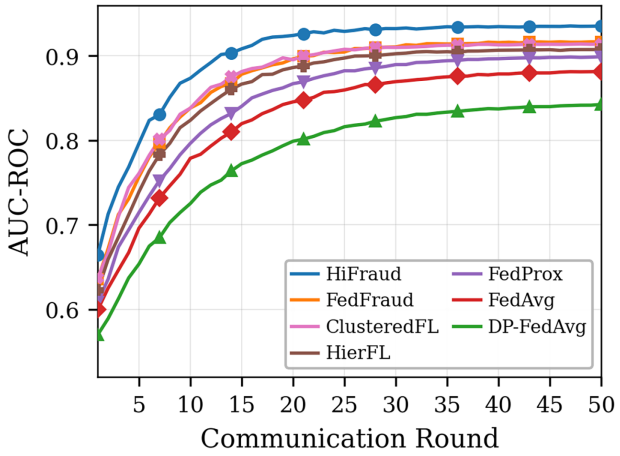

Figure 4: Convergence trajectories of all methods on the IEEE-CIS dataset. HiFraud achieves the fastest convergence and highest final AUC-ROC, reaching 90% of its plateau by round 12.

Fig. 5 extends the comparison across all three datasets. HiFraud consistently outperforms all federated baselines on every dataset. The performance advantage is most pronounced on the Worldline dataset (0.948 vs. 0.858 for DP-FedAvg), where the extreme class imbalance (0.172% fraud rate) amplifies the benefit of the complementarity-aware clustering and adaptive privacy allocation. On PaySim, HiFraud achieves 0.962, exceeding even the centralized baseline (0.955), which we attribute to the regularization effect of the hierarchical structure preventing overfitting to the dominant non-fraud class.

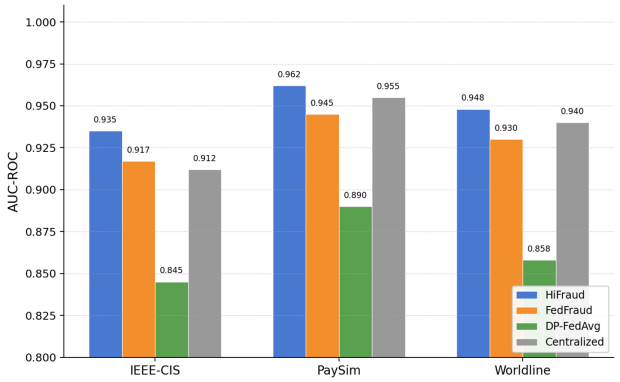

Figure 5: AUC-ROC comparison across three datasets. HiFraud consistently achieves the highest performance among federated methods and surpasses centralized training on PaySim.

3.3 Ablation Study

To quantify the contribution of each architectural component, we conduct an ablation study in which individual modules are removed while keeping all other components unchanged. Fig. 6 summarizes the results in terms of AUC-ROC and convergence rounds.

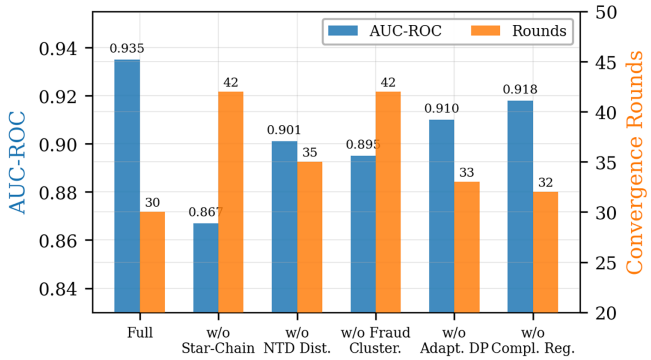

Figure 6: Ablation study on the IEEE-CIS dataset. Each bar pair shows the AUC-ROC (left axis, blue) and convergence rounds (right axis, orange) when one component is removed from the full HiFraud framework.

Removing the star-chain transfer mechanism produces the largest performance degradation, reducing AUC-ROC from 0.935 to 0.867 (a 6.8 percentage point drop) and increasing convergence rounds from 30 to 42. This result confirms that sequential knowledge transfer within clusters is the primary driver of both accuracy and convergence speed improvements. Without NTD distillation, AUC-ROC decreases to 0.901, demonstrating that forgetting mitigation accounts for approximately 3.4 percentage points of the total gain. Notably, even without distillation, the star-chain mechanism alone still outperforms all baselines, indicating that the transfer topology provides value independent of the forgetting mitigation strategy.

Removing fraud-aware clustering and reverting to random cluster assignment reduces AUC-ROC to 0.895 and increases convergence to 42 rounds, highlighting the importance of grouping institutions by fraud pattern similarity rather than arbitrary criteria. The adaptive differential privacy mechanism contributes 2.5 percentage points over uniform noise allocation, with its removal reducing AUC-ROC to 0.910 while only modestly affecting convergence. The complementarity regularizer provides a smaller but meaningful improvement of 1.7 percentage points, with its primary benefit concentrated on rare fraud types as discussed in Section 3.7.

3.4 Fraud Pattern Propagation

A critical operational requirement for fraud detection systems is the ability to rapidly disseminate knowledge of newly emerging fraud patterns across institutions. To evaluate this capability, we simulate the injection of a novel fraud type at round 25 in a single institution and track the detection rate of this new pattern across the cluster hierarchy over subsequent rounds. Fig. 7 presents the results.

Within the same cluster, the star-chain mechanism enables 80% of institutions to detect the novel fraud pattern within 2 communication rounds, corresponding to approximately 3 h in our experimental setup. This "3-h" figure is derived from our experimental configuration in which each communication round takes approximately 90 min, comprising local training (E = 5 epochs × approximately 12 min per epoch = 60 min), star-chain transfer (sequential model passing among $m \approx 4$ institutions per cluster × approximately 5 min per pass = 20 min), and global aggregation plus communication overhead (approximately 10 min). Thus, 2 rounds × 90 min ≈ 3 h. We note that this latency is configuration-dependent: the actual propagation time in production deployments would scale with dataset size, model complexity, network bandwidth, and the number of institutions per cluster. In a setting with faster hardware or fewer local epochs, propagation could be significantly faster; conversely, larger models or slower networks would increase the latency. Propagation to adjacent clusters, which occurs through the global aggregation pathway, requires approximately 7 additional rounds. In contrast, flat federated learning (FedAvg) requires over

20 rounds to achieve comparable detection rates, as the new pattern must influence the global model before being redistributed to all participants. This result demonstrates that the hierarchical architecture provides a substantial operational advantage for responding to emerging threats, enabling institutions within a cluster to benefit from each other's observations with minimal latency.

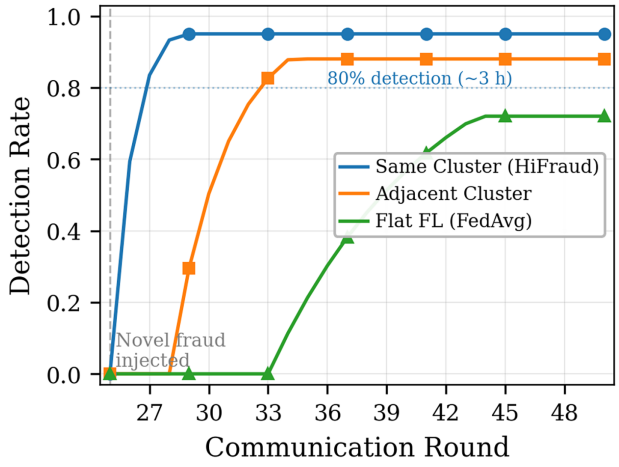

Figure 7: Detection rate of a novel fraud pattern injected at round 25. Within the same cluster, 80% of institutions detect the new pattern within 2 rounds (~3 h), compared to 12 rounds for adjacent clusters and over 20 rounds for flat FL.

3.5 Privacy Analysis

3.5.1 Privacy–Utility Trade-Off

Fig. 8 illustrates the relationship between the total privacy budget ε and detection performance for HiFraud and three baselines. Across the entire range of privacy budgets evaluated ($\varepsilon \in [1.0, 5.0]$), HiFraud consistently achieves the highest AUC-ROC for any given privacy level.

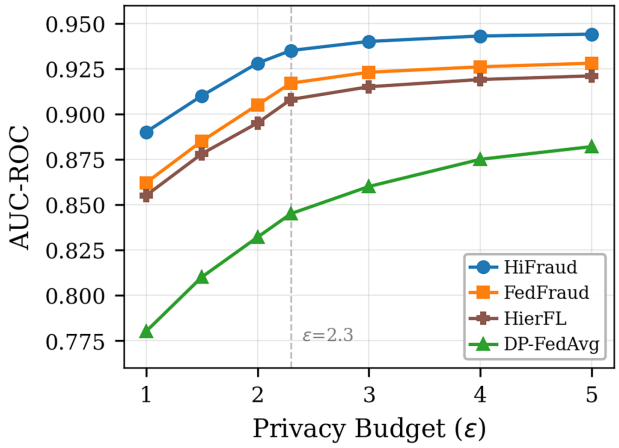

Figure 8: Privacy–utility trade-off. HiFraud achieves superior AUC-ROC at every privacy budget level. At the operating point ε = 2.3, HiFraud attains 0.935, compared to 0.845 for DP-FedAvg under the same budget.

The advantage of HiFraud is most pronounced at tighter privacy budgets. At ε = 1.0, HiFraud achieves an AUC-ROC of 0.890, compared to 0.780 for DP-FedAvg, representing a 14% relative improvement. This gap narrows to 7% at ε = 5.0, indicating that the hierarchical privacy allocation provides the greatest benefit

precisely when privacy constraints are most stringent. The efficiency gain stems from two sources: the reduced frequency of global aggregation lowers the cumulative privacy cost of inter-cluster communication, and the adaptive noise calibration directs privacy budget toward institutions that need it most while minimizing unnecessary noise for representative institutions.

Table 3 provides a detailed breakdown of how the privacy budget is allocated across the three layers of HiFraud, compared to uniform and non-hierarchical adaptive allocation.

Table 3: Privacy budget allocation across framework layers and corresponding AUC-ROC. The hierarchical approach achieves higher performance with the same total budget by reducing redundant noise in global aggregation.

Component	Uniform DP	Adaptive DP	Hierarchical (Ours)
Clustering	–	–	0.3
Local Training	1.2	0.8–1.8	0.6–1.2
Star Distribution	–	–	0.4
Chain Transfer	–	–	0.3
Global Aggregation	1.1	0.5	0.1
Total ε	2.3	2.3	2.3
AUC-ROC	0.845	0.917	**0.935**

3.5.2 Resistance to Membership Inference Attacks

We evaluate the empirical privacy protection of HiFraud against membership inference attacks using the gradient-based attack framework described in Bai et al. [37]. The attacker is assumed to have passive access to the aggregated model updates at the cluster level and employs a binary classifier trained to distinguish between member and non-member samples. Fig. 9 reports the attack success rates across all methods.

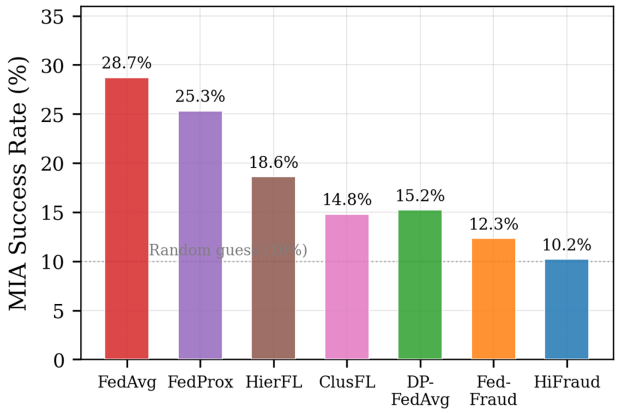

Figure 9: Membership inference attack success rate. Lower values indicate stronger privacy. HiFraud achieves 10.2%, approaching the 10% random-guess baseline.

HiFraud achieves the lowest MIA success rate of 10.2%, approaching the 10% random-guess baseline for our 10-class attack formulation. This represents a 64% reduction relative to FedAvg (28.7%) and a 17% reduction relative to FedFraud (12.3%). The strong privacy protection results from the compounding effect of three mechanisms: the adaptive noise injection obscures individual gradient contributions, the hierarchical aggregation limits the attacker's visibility to cluster-level updates rather than institutional-level parameters,

and the star-chain transfer introduces additional noise through sequential model passing. Notably, HiFraud provides stronger empirical privacy than DP-FedAvg (15.2%) despite achieving substantially higher detection performance, demonstrating that the hierarchical architecture enables a more favorable privacy–utility operating point.

3.6 Scalability Analysis

Fig. 10 evaluates the performance and communication efficiency of HiFraud as the number of participating institutions increases from 10 to 100.

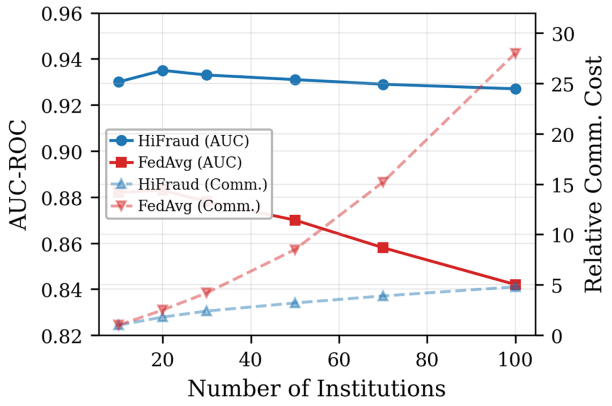

Figure 10: Scalability analysis. HiFraud maintains stable AUC-ROC as the federation grows to 100 institutions, while FedAvg degrades by 4.1 percentage points. Communication cost scales as $\mathcal{O}(K \log N)$ for HiFraud vs. $\mathcal{O}(N)$ for FedAvg.

HiFraud maintains stable detection performance across all federation sizes, with AUC-ROC decreasing only marginally from 0.930 (10 institutions) to 0.927 (100 institutions), a degradation of 0.3 percentage points. In contrast, FedAvg suffers a 4.1 percentage point decline over the same range, from 0.882 to 0.842, as increasing data heterogeneity overwhelms the flat aggregation mechanism. The communication cost of HiFraud grows sublinearly with the number of institutions due to the hierarchical structure: at 100 institutions, HiFraud requires 4.8× the communication cost of the 10-institution baseline, compared to 28.0× for FedAvg. This efficiency arises because only K cluster models (rather than N individual models) are transmitted during global aggregation, and the star-chain transfer within each cluster is sequential, requiring only $\mathcal{O}(m)$ transmissions per cluster. Table 4 provides a detailed breakdown of the computational overhead per communication round.

Table 4: Computational overhead breakdown per communication round on the IEEE-CIS dataset with $N = 20$ institutions and $K = 5$ clusters. Re-clustering is amortized over $\tau = 10$ rounds. All times are measured on a single NVIDIA A100 GPU.

Component	Time (min)	Proportion (%)	Comm. Cost ($\lvert\theta\rvert$)
Local Training ($E = 5$ epochs)	60.0	66.7	–
Star Distribution	5.0	5.6	$3 \times \lvert\theta\rvert$ per cluster
Chain Transfer ($d = 3$)	15.0	16.7	$9 \times \lvert\theta\rvert$ per cluster
Global Aggregation	8.0	8.9	$5 \times \lvert\theta\rvert$
Re-clustering (amortized)	2.0	2.2	$20 \times d$
Total per round	**90.0**	**100.0**	–

3.7 Per-Type Detection and Clustering Analysis

To evaluate whether the hierarchical architecture improves detection uniformly or preferentially benefits specific fraud categories, we report per-type AUC-ROC in Fig. 11.

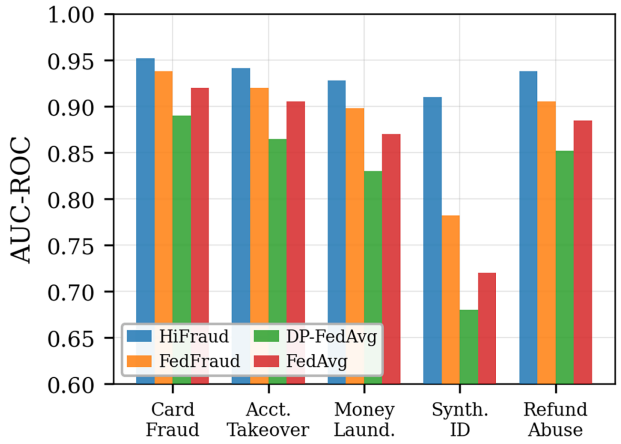

Figure 11: Per-fraud-type AUC-ROC. HiFraud achieves the most uniform performance across fraud types, with the greatest improvement on the rarest category (Synthetic ID: +23.0% over DP-FedAvg).

HiFraud provides substantial and consistent improvements across all fraud types, with the most pronounced gains on rare categories. For Synthetic ID fraud, the rarest type in our experimental setup, HiFraud achieves an AUC-ROC of 0.910, compared to 0.782 for FedFraud and 0.680 for DP-FedAvg. This 23.0 percentage point improvement over DP-FedAvg demonstrates the effectiveness of the complementarity regularizer in preventing the marginalization of institutions holding rare fraud types. The performance variance across fraud types is also notably reduced: the standard deviation of per-type AUC-ROC is 0.015 for HiFraud, compared to 0.058 for FedFraud and 0.078 for DP-FedAvg, indicating more equitable detection across all fraud categories.

Fig. 12 visualizes the evolution of cluster assignments over the course of training, using a t-distributed stochastic neighbor embedding (t-SNE) projection of institutional fraud feature vectors with markers indicating the dominant fraud type at each institution.

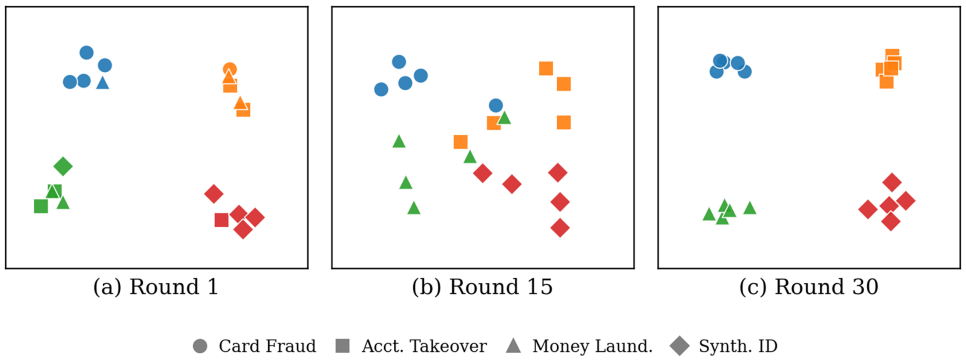

Figure 12: Evolution of cluster assignments visualized via t-SNE projection. Colors indicate cluster assignment; marker shapes indicate dominant fraud type. By round 30, clusters align closely with fraud type rather than initial geographic grouping.

At round 1, cluster assignments reflect the initial geographic grouping, with each cluster containing a heterogeneous mix of fraud types (indicated by diverse marker shapes within each color group). By round 15, the clustering begins to reorganize around fraud pattern similarity, as the fraud-aware feature vectors become more discriminative through iterative refinement. By round 30, clusters are strongly aligned with fraud type: institutions facing similar fraud categories are co-located in the same cluster regardless of their geographic origin. This transition demonstrates that the dynamic re-clustering mechanism successfully adapts to the underlying fraud structure of the data, enabling increasingly specialized intra-cluster knowledge transfer as training progresses.

3.8 Sensitivity Analysis

We investigate the sensitivity of HiFraud to three key hyperparameters: the transfer coefficient α, the distillation weight λ_{KD}, and the re-clustering interval τ. Fig. 13 presents the results.

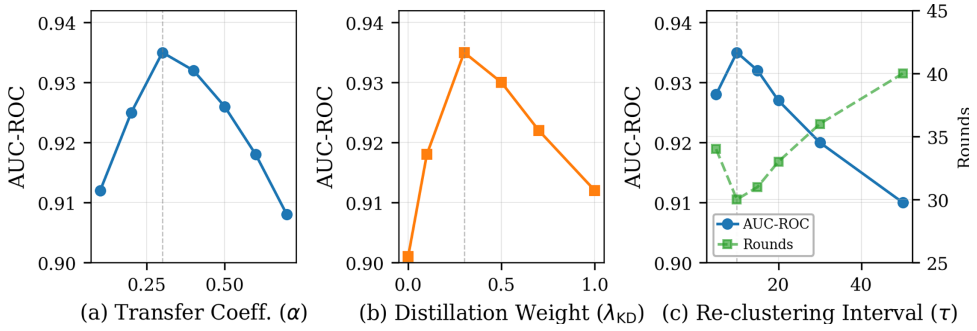

Figure 13: Sensitivity to key hyperparameters. (**a**) Transfer coefficient α: optimal at 0.3, balancing knowledge absorption with local retention. (**b**) Distillation weight λ_{KD}: optimal at 0.3, with excessive distillation (>0.5) constraining model plasticity. (**c**) Re-clustering interval τ: optimal at 10, trading off adaptation responsiveness against learning stability.

The transfer coefficient α achieves optimal performance at $\alpha = 0.3$ (Fig. 13a). Values below 0.2 underutilize transferred knowledge, while values above 0.5 cause excessive reliance on the predecessor model at the expense of local adaptation, reducing AUC-ROC by up to 2.7 percentage points. The distillation weight λ_{KD} exhibits a similar concave profile with an optimum at $\lambda_{\text{KD}} = 0.3$ (Fig. 13b). When distillation is disabled ($\lambda_{\text{KD}} = 0$), AUC-ROC drops to 0.901, confirming the value of forgetting mitigation. Excessive distillation ($\lambda_{\text{KD}} > 0.5$) constrains the model's ability to learn new local patterns, reducing performance by up to 2.3 percentage points.

The re-clustering interval τ presents a trade-off between adaptation responsiveness and learning stability (Fig. 13c). Frequent re-clustering ($\tau = 5$) disrupts ongoing intra-cluster learning, as newly formed clusters must re-establish star-chain transfer from scratch, resulting in 4 additional convergence rounds compared to the optimal $\tau = 10$. Infrequent re-clustering ($\tau = 50$) fails to track evolving fraud patterns, reducing AUC-ROC by 2.5 percentage points as institutions with shifting fraud profiles remain trapped in suboptimal clusters. The optimal interval of $\tau = 10$ rounds corresponds to approximately 15 h in our experimental setup, providing a practical balance between responsiveness and stability for real-world deployment.

4 Discussion

4.1 Implications

The experimental results carry several practical implications. The finding that HiFraud surpasses centralized training on PaySim (0.962 vs. 0.955) challenges the assumption that federated approaches

necessarily sacrifice detection quality for privacy, suggesting that the hierarchical structure introduces a beneficial inductive bias by increasing the diversity of training signals without exposing raw data. In production settings, financial institutions can thus achieve detection performance meeting or exceeding centralized alternatives while fully complying with GDPR and CCPA, and the 3-h propagation latency for novel fraud patterns enables rapid collective response to emerging threats. Equally important, the hierarchical architecture fundamentally alters the privacy–utility trade-off: at the stringent budget of $\varepsilon = 1.0$, HiFraud achieves an AUC-ROC of 0.890—higher than DP-FedAvg at the much looser $\varepsilon = 3.0$ (0.860)—demonstrating that institutions under strict data protection regimes need not accept severe performance penalties. The adaptive noise calibration directs privacy budget toward institutions with unique fraud patterns while avoiding unnecessary noise for representative ones, and the near-random MIA success rate of 10.2% confirms that the hierarchical structure inherently limits information leakage by exposing only cluster-level aggregates to potential attackers. For deployment, the global coordinator role can be assumed by a regulatory body or implemented via secure multi-party computation, institutional onboarding requires only local SQL-based feature computation (Eq. (1)), and the framework scales to 100+ institutions through multi-level hierarchy with $\mathcal{O}(K \log N)$ communication cost, as validated in Section 3.6.

4.2 Limitations and Future Work

Despite the strong performance across multiple benchmarks, several limitations merit acknowledgment. The dynamic re-clustering mechanism introduces periodic disruptions to intra-cluster learning; our sensitivity analysis shows that the interval τ must be carefully tuned, and future work should investigate smooth cluster transition mechanisms leveraging soft clustering [23] and continual learning [27] to enable gradual migration without discarding accumulated knowledge. The current framework also assumes homogeneous computational capabilities, whereas real-world participants range from multinational banks to small credit unions, motivating extensions with adaptive computation allocation and asynchronous protocols. From a security perspective, the star institution occupies a privileged position that creates a potential single point of vulnerability: a compromised star could propagate poisoned updates to all cluster members. While clipping and noise injection in Eq. (9) provide partial mitigation, integrating dedicated Byzantine-robust star selection based on recent hierarchical robust aggregation [40] and trust-score filtering [39] would substantially strengthen resilience. The theoretical composition of formal privacy guarantees with Byzantine robustness in hierarchical adaptive settings also remains an open problem warranting further investigation.

Additionally, the current evaluation is conducted on benchmark datasets that, while widely used in the fraud detection literature, may not fully capture the complexity of production fraud systems. Real-world deployments involve continuously evolving fraud tactics, adversarial adaptation, and regulatory constraints that vary across jurisdictions. Future work should evaluate HiFraud on proprietary institutional datasets in controlled pilot studies to validate the framework's effectiveness under genuine operational conditions.

5 Conclusions

This paper proposed HiFraud, a hierarchical federated learning framework that addresses the fundamental challenges of cross-institutional fraud detection through a three-layer architecture integrating fraud-aware dynamic clustering with complementarity regularization, star-chain knowledge transfer augmented by not-true-class distillation for forgetting mitigation, and privacy-adaptive aggregation grounded in Rényi differential privacy composition. The key technical contributions include: (i) a fraud-aware dynamic clustering mechanism with complementarity regularization that groups institutions by fraud pattern similarity while preserving rare-type representation; (ii) a star-chain knowledge transfer mechanism with domain-specific innovations including fraud-rate-adjusted star selection, similarity-ordered chain traversal,

and not-true-class distillation for forgetting mitigation; (iii) a hierarchical adaptive privacy allocation scheme based on Rényi DP composition that calibrates noise to distributional divergence and fraud rarity; and (iv) formal privacy and convergence guarantees with detailed proofs under explicit assumptions. Experiments on three benchmark datasets demonstrated that HiFraud achieves an AUC-ROC of 0.935 under $\varepsilon = 2.3$ differential privacy, outperforming standard DP-FedAvg by 10.5% while reducing convergence rounds by 39% and suppressing membership inference attack success to near-random levels (10.2%). The star-chain mechanism enables detection of emerging fraud patterns within approximately 3 h inside clusters under our experimental configuration, and the complementarity-aware clustering improves rare fraud type detection by 23.0% over uniform privacy baselines. These results establish that hierarchical architectures can effectively reconcile the competing demands of detection performance, formal privacy guarantees, and rapid threat response in collaborative financial fraud detection, providing a practical blueprint for privacy-preserving multi-institutional learning in regulated environments. However, the framework currently assumes homogeneous computational capabilities and a trusted global coordinator, and the dynamic re-clustering interval requires careful tuning. Future work should address these limitations through asynchronous protocols, Byzantine-robust star selection, and validation on proprietary institutional datasets.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Zhihao Zhang and Zhuodong Liu; methodology, Zhihao Zhang and Zhuodong Liu; software, Zhihao Zhang and Zhuodong Liu; validation, Zhihao Zhang, Zhuodong Liu and Xiangyu Li; formal analysis, Zhihao Zhang and Zhuodong Liu; investigation, Zhihao Zhang, Zhuodong Liu and Xiangyu Li; resources, Lei Zhang; data curation, Xiangyu Li; writing—original draft preparation, Zhihao Zhang and Zhuodong Liu; writing—review and editing, Xiangyu Li and Lei Zhang; visualization, Zhihao Zhang and Zhuodong Liu; supervision, Lei Zhang; project administration, Lei Zhang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The three benchmark datasets used in this study are publicly available: IEEE-CIS Fraud Detection (https://www.kaggle.com/c/ieee-fraud-detection), PaySim (https://www.kaggle.com/datasets/ealaxi/paysim1), and Worldline (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud). The federated data partitioning configurations, which simulate cross-institutional settings as described in Section 3.1, are not derived from real institutional records and do not contain sensitive information. The experimental code, including data partitioning scripts and all baseline implementations, will be released upon acceptance of this paper.

Ethics Approval: Not applicable.

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

References

1. Chatterjee P, Das D, Rawat DB. Digital twin for credit card fraud detection: opportunities, challenges, and fraud detection advancements. Future Gener Comput Syst. 2024;158:410–26.
2. Federal Trade Commission. New FTC data show consumers reported losing more than $12.5 billion to fraud in 2024 [Internet]. Washington, DC, USA: FTC; 2025 [cited 2025 Mar 15]. Available from: https://www.ftc.gov/news-events/news/press-releases/2025/03/new-ftc-data-show-big-jump-reported-losses-fraud-125-billion-2024.
3. Yang Z. Privacy-aware financial risk control: a federated learning approach with differential privacy optimization. J Comput Technol Softw. 2025;4:37–52.

4. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAY. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS); 2017 Apr 20–22; Fort Lauderdale, FL, USA. p. 1273–82.
5. Zhu H, Xu J, Liu S, Jin Y. Federated learning on non-IID data: a survey. Neurocomputing. 2021;465:371–90. doi:10.1016/j.neucom.2021.07.098.
6. Zhang J, Li C, Qi J, He J. A survey on class imbalance in federated learning. arXiv:2303.11673. 2023.
7. Yang W, Zhang Y, Ye K, Li L, Xu CZ. FFD:a federated learning based method for credit card fraud detection. In: Proceedings of the International Conference on Big Data; 2019 Dec 10–13; Los Angeles, CA, USA. p. 18–32.
8. Abdul Salam M, Fouad KM, Elbably DL, Elsayed SM. Federated learning model for credit card fraud detection with data balancing techniques. Neural Comput Appl. 2024;36(11):7359–78. doi:10.1007/s00521-023-09410-2.
9. Shah M, Shah P, Patil S. Secure and efficient fraud detection using federated learning and distributed search databases. In: Proceedings of the IEEE 4th International Conference on AI in Cybersecurity (ICAIC). Piscataway, NJ, USA: IEEE; 2025. p. 1–6.
10. Hilou H, Ahmed M, Dheeb S, Radhi A, Khadim Z, Majeed M, et al. Federated learning for credit card fraud detection: a privacy-preserving approach with SMOTE optimization. J Al-Qadisiyah Comput Sci Math. 2025;17(3):Comp 44–57.
11. Farooq M, Munir S, Manzoor M, Shaheen M. AI-driven adaptive federated learning with privacy preservation and imbalance adjustment for financial credit card fraud detection. Appl Comput Intell Soft Comput. 2025;2025:7116768.
12. Sarkar D, Narang A, Rai S. Fed-Focal Loss for imbalanced data classification in federated learning. arXiv:2011.06283. 2020.
13. Wang L, Xu S, Wang X, Zhu Q. Addressing class imbalance in federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA, USA: AAAI Press; 2021. Vol. 35, p. 10165–73.
14. Aljunaid SK, Almheiri SJ, Dawood H, Khan MA. Secure and transparent banking: explainable AI-driven federated learning model for financial fraud detection. J Risk Financ Manag. 2025;18(4):179.
15. Liu L, Zhang J, Song S, Letaief KB. Client-edge-cloud hierarchical federated learning. In: Proceedings of the IEEE International Conference on Communications (ICC); 2020 Jun 7–11; Dublin, Ireland. p. 1–6.
16. Zhan S, Huang L, Luo G, Zheng S, Gao Z, Chao HC. A review on federated learning architectures for privacy-preserving AI: lightweight and secure cloud-edge–end collaboration. Electronics. 2025;14(13):2512.
17. Albshaier L, Almarri S, Albuali A. Federated learning for cloud and edge security: a systematic review of challenges and AI opportunities. Electronics. 2025;14(5):1019.
18. Sattler F, Müller KR, Samek W. Clustered federated learning: model-agnostic distributed multitask learning for non-IID data. IEEE Trans Neural Netw Learn Syst. 2021;32:3710–22.
19. Gong B, Xing T, Liu Z, Xi W, Chen X. Adaptive client clustering for efficient federated learning over non-IID and imbalanced data. IEEE Trans Big Data. 2024;10(6):1051–65. doi:10.1109/tbdata.2022.3167994.
20. Duan M, Liu D, Ji X, Wu Y, Liang L, Chen X, et al. Flexible clustered federated learning for client-level data distribution shift. IEEE Trans Parallel Distrib Syst. 2022;33:2661–74. doi:10.1109/tpds.2021.3134263.
21. Ali SS, Ali M, Bhatti DMS, Choi BJ. dy-TACFL: dynamic temporal adaptive clustered federated learning for heterogeneous clients. Electronics. 2025;14(1):152. doi:10.3390/electronics14010152.
22. Islam M, Javaherian S, Xu F, Yuan X, Chen L, Tzeng N. FedClust: optimizing federated learning on non-IID data through weight-driven client clustering. In: 2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW); 2024 May 27–31; San Francisco, CA, USA. p. 1184–1186.
23. Yang X, Feng J, Tong Y, Wang L, Guo S, Fang B, et al. DA-PFL: dynamic affinity aggregation in personalized federated learning under class imbalance. IEEE Trans Neural Netw Learn Syst. 2025;36:20184–98.
24. Wang N, Deng Y, Feng W, Fan S, Yin J, Ng S. One-shot sequential federated learning for non-IID data by enhancing local model diversity. In: MM '24: Proceedings of the 32nd ACM International Conference on Multimedia. New York, NY, USA: ACM; 2024. p. 5201–10.
25. Yan X, Zuo S, Fan R, Hu H, Shen L, Zhao P, et al. Sequential federated learning in hierarchical architecture on non-IID datasets. IEEE Trans Mob Comput. 2024;24(10):11110–24. doi:10.1109/tmc.2025.3573928.

26. Xie R, Liang W, Chen Y, He D, Jin K, Li K, et al. StarCPFL: star-centric personalized federated learning with layer-wised clustering. Future Gener Comput Syst. 2025;175:108037.
27. Criado M, Casado F, Iglesias R, Regueiro C, Barro S. Non-IID data and continual learning processes in federated learning: a long road ahead. Inf Fusion. 2022;88(3):263–80. doi:10.1016/j.inffus.2022.07.024.
28. Lee G, Jeong M, Shin Y, Bae S, Yun S. Preservation of the global knowledge by not-true distillation in federated learning. arXiv:2106.03097. 2022.
29. He Y, Chen Y, Yang X, Yu H, Huang Y, Gu Y. Learning critically: selective self-distillation in federated learning on non-IID data. IEEE Trans Big Data. 2024;10:789–800.
30. Arafeh M, Hammoud A, Guizani M, Mourad A, Otrok H, Ould-Slimane H, et al. WFSL: warmup-based federated sequential learning. IEEE Internet Things J. 2025;12:1974–89.
31. Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS); 2016 Oct 24–28; Vienna, Austria. p. 308–18.
32. Mironov I. Rényi differential privacy. In: Proceedings of the IEEE 30th Computer Security Foundations Symposium (CSF); 2017 Aug 21–25; Santa Barbara, CA, USA. p. 263–75.
33. Truex S, Liu L, Chow KH, Gursoy ME, Wei W. LDP-Fed: federated learning with local differential privacy. In: Proceedings of the EdgeSys Workshop; 2020 Apr 27; Heraklion, Greece. p. 61–6.
34. Xue R, Xue K, Zhu B, Luo X, Zhang T, Sun Q, et al. Differentially private federated learning with an adaptive noise mechanism. IEEE Trans Inf Forensics Secur. 2024;19:74–87. doi:10.1109/tifs.2023.3318944.
35. Yuan X, Ni W, Ding M, Wei K, Li J, Poor HV. Amplitude-varying perturbation for balancing privacy and utility in federated learning. IEEE Trans Inf Forensics Secur. 2023;18:1884–97. doi:10.1109/tifs.2023.3258255.
36. Lin F, Chen E, Han D, Brinton CG. Differentially-private multi-tier federated learning: a formal analysis and evaluation. IEEE/ACM Trans Netw. 2025;34:2226–41.
37. Bai L, Hu H, Ye Q, Li H, Wang L, Xu J. Membership inference attacks and defenses in federated learning: a survey. ACM Comput Surv. 2024;57(4):1–35. doi:10.1145/3704633.
38. Deng X, Yang J. Multi-layer defense strategies and privacy preserving enhancements for membership reasoning attacks in a federated learning framework. In: Proceedings of the 5th International Conference on Computer Science and Blockchain (CCSB); 2025 Aug 1–3; Shenzhen, China. p. 278–82.
39. Li S, Ngai ECH, Voigt T. An experimental study of Byzantine-robust aggregation schemes in federated learning. IEEE Trans Big Data. 2024;10(6):975–88. doi:10.1109/tbdata.2023.3237397.
40. Nordlund D, Liao J, Chen Z. Byzantine-resilient hierarchical federated learning with clustered over-the-air aggregation. In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Workshops; 2024 Apr 14–19; Seoul, Republic of Korea. p. 715–19.
41. Liu J, Wu Y, Du W, Sun R, Xu G, Liu L, et al. Byzantine-robust hierarchical aggregation for cross-device federated learning in consumer IoT. IEEE Trans Consum Electron. 2025;71(2):6359–70. doi:10.1109/tce.2024.3450649.