Towards Collaborative Fairness in Federated Learning Under Imbalanced Covariate Shift

Tianrun Yu
The Pennsylvania State University
University Park, PA, USA
tvy5242@psu.edu

Jiaqi Wang
The Pennsylvania State University
University Park, PA, USA
jqwang@psu.edu

Haoyu Wang
State University of New York at Albany
Latham, NY, USA
hwang28@albany.edu

Mingquan Lin
University of Minnesota Twin Cities
Minneapolis, MN, USA
lin01231@umn.edu

Han Liu
Dalian University of Technology
Dalian, Liaoning, China
liu.han.dut@gmail.com

Nelson S. Yee
The Pennsylvania State University
Hershey, PA, USA
nyee@pennstatehealth.psu.edu

Fenglong Ma*
The Pennsylvania State University
University Park, PA, USA
fenglong@psu.edu

ABSTRACT

Collaborative fairness is a crucial challenge in federated learning. However, existing approaches often overlook a practical yet complex form of heterogeneity: *imbalanced covariate shift*. We provide a theoretical analysis of this setting, which motivates the design of FedAKD (Federated Asynchronous Knowledge Distillation)—a simple yet effective approach that balances accurate prediction with collaborative fairness. FedAKD consists of client and server updates. In the client update, we introduce a novel asynchronous knowledge distillation strategy based on our preliminary analysis, which reveals that while correctly predicted samples exhibit similar feature distributions across clients, incorrectly predicted samples show significant variability. This suggests that imbalanced covariate shift primarily arises from misclassified samples. Leveraging this insight, our approach first applies traditional knowledge distillation to update client models while keeping the global model fixed. Next, we select correctly predicted high-confidence samples and update the global model using these samples while keeping client models fixed. The server update simply aggregates all client models. We further provide a theoretical proof of FedAKD's convergence. Experimental results on public datasets (FashionMNIST and CIFAR10) and a real-world Electronic Health Records (EHR) dataset demonstrate that FedAKD significantly improves collaborative fairness, enhances predictive accuracy, and fosters client participation even under highly heterogeneous data distributions. [1]

*Corresponding author.

[1]Source code is available at https://github.com/Tianrun-Yu/FedAKD.

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737161

CCS CONCEPTS

• **Information systems** → **Information systems applications**.

KEYWORDS

federated learning, collaborative fairness, covariate shift, knowledge distillation, imbalanced data

ACM Reference Format:
Tianrun Yu, Jiaqi Wang, Haoyu Wang, Mingquan Lin, Han Liu, Nelson S. Yee, and Fenglong Ma. 2025. Towards Collaborative Fairness in Federated Learning Under Imbalanced Covariate Shift. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3711896.3737161

1 INTRODUCTION

Federated Learning (FL) has emerged as a promising distributed paradigm that enables multiple participants (or clients) to collaboratively train a global model without sharing their raw local data [6, 9, 15, 24, 31]. However, disparities in data quality, quantity, and distribution among clients make uniform treatment in global model aggregation unfair, particularly for those with higher-quality or larger datasets. To address this, **collaborative fairness** (CF) has been introduced to ensure that each client's final reward or benefit is commensurate with its contribution to the global model [13]. In other words, clients with a greater impact on model performance should receive proportionally higher gains.

Several fairness-aware FL approaches, such as CGSV [23], CFFL [13], FedAve [19], and FedSAC [20], have been developed to assign different contribution-based rewards (i.e., weights) during model aggregation. While these methods help mitigate fairness disparities, they still face the following challenges:

• **Unrealistic assumptions about data heterogeneity.** Existing CF approaches primarily assume that data heterogeneity arises from imbalanced data sizes [13, 23], imbalanced class distributions [13, 23], or both [1, 28]. However, real-world datasets exhibit greater complexity. Beyond these imbalances, client data

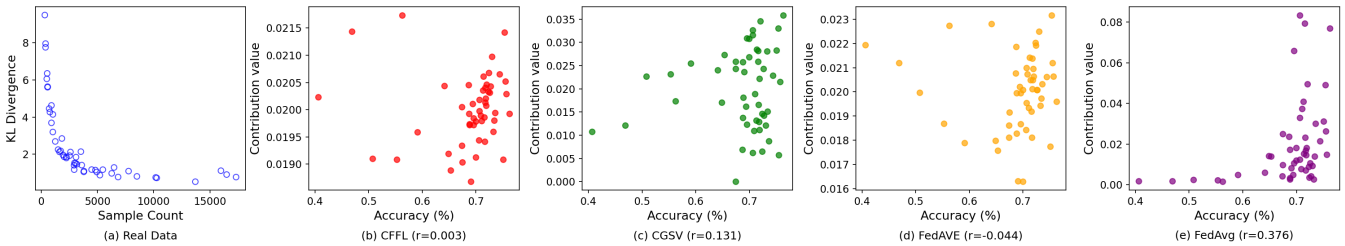


Figure 1: (a) KL divergence vs. sample size for each client's local data, revealing both data imbalance and feature covariate shifts. (b)-(e) compare different fairness methods' contribution definitions to each client's standalone training accuracy. Their poor correlation highlights the limitations of explicit contribution metrics under extra covariate shifts.

often differ significantly in feature distributions, leading to the *covariate shift* problem [1, 2, 5, 9]. Figure 1(a) presents a preliminary analysis on a real-world electronic health record (EHR) dataset for pancreatic cancer prediction[2]. The x-axis represents client dataset sizes, while the y-axis shows the Kullback–Leibler (KL) divergence between each client's fitted latent feature distribution and that of the entire dataset. Each circle represents a client, i.e., a U.S. state in the EHR dataset. The preliminary data analysis reveals that not only do clients have varying dataset sizes, but their latent feature distributions also significantly diverge from the global reference distribution. Thus, a more realistic FL heterogeneity setting should account for **imbalanced covariate shift** rather than just data quantity or class imbalance.

• **Weak correlation between client accuracy and assigned contributions.** Beyond unrealistic data distribution assumptions, existing approaches to collaborative fairness typically follow a two-step pipeline: *(1) explicitly defining a contribution metric and (2) allocating rewards based on this metric*. For example, CGSV [23] estimates client contributions via gradient similarity, CFFL [13] and FedAVE [19] rely on performance improvements measured by a global validation set, and FedSAC [20] bases contributions on standalone local training results. However, in real-world datasets characterized by imbalanced covariate shift, these approaches fail to establish a strong correlation between client accuracy and assigned contributions, contradicting their underlying design assumptions.

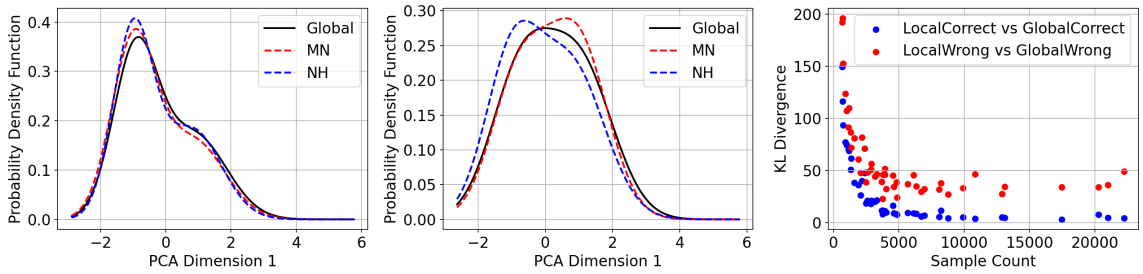
Figures 1(b)–(e) illustrate the relationship between client standalone accuracy (x-axis) on the testing set and the learned contribution value (y-axis) either on the training set or validation set under different methods – CFFL, CGSV, FedAVE, and the baseline FedAvg [15] – using the real-world EHR dataset, which is the same as we analyzed in Figure 1(a). Each dot represents an individual client, and in the case of FedAvg, the contribution score is simply the proportion of data owned by the client. The results show that existing CF-based methods exhibit significantly lower Pearson correlation scores compared to the simple FedAvg baseline. These findings highlight the limitations of existing approaches in handling realistic imbalanced covariate shift settings.

Theoretical analysis on imbalanced covariate shift. This paper aims to develop a novel model to ensure collaborative fairness in federated learning under the realistic yet challenging imbalanced covariate shift setting. To achieve this, in Section 2, we first provide a theoretical analysis in Theorem 2.1 demonstrating that imbalanced covariate shift – quantified as the KL divergence between each client's empirical data distribution and the ideal global distribution – is primarily influenced by the perturbation δ. Specifically, if the underlying data distributions of individual clients and the entire dataset follow a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, where μ is the mean vector, we show that δ and the covariance Σ play key roles in determining the extent of distributional divergence in Theorem 2.2. The two theorems motivate us to mitigate the imbalanced covariate shift to achieve collaborative fairness by correctly qualifying δ and Σ. However, directly calculating these values is infeasible, as the true data distributions are inherently *unknown*.

Motivations of model design. To address this challenge, we conducted a preliminary analysis on the results of FedAvg applied to the entire EHR dataset. For each client k, we categorized the correctly and incorrectly predicted samples as $\mathcal{I}_k$ and $\mathcal{D}_k - \mathcal{I}_k$, respectively, and aggregated these categories across all clients to form the global sets $\{\mathcal{I}_k\}_{k=1}^K$ and $\{\mathcal{D}_k - \mathcal{I}_k\}_{k=1}^K$, where $\mathcal{D}_k$ denotes the client dataset, and K is the number of clients. Next, we applied principal component analysis (PCA) to project each sample's latent representation (i.e., the encoder output from each client model) onto a 1-D space and used kernel density estimation (KDE) to estimate the probability density function (PDF) for each set. Figures 2(a) and (b) show a comparison of the global density function with the density functions of two clients (*Minnesota* and *New Hampshire*). The x-axis represents the projected PCA 1-D values, and the y-axis represents the estimated PDF values. Similar to Figure 1(a), we also analyze the distribution differences between local and global data in terms of correct and incorrect classifications, as illustrated in Figure 2(c). These results indicate that *the primary source of imbalanced covariate shift lies in the "incorrect" samples, as clients generally show agreement on the distribution of "correct" predictions.*

Our approach. Building on our theoretical and empirical analysis, we propose FedAKD, a novel framework designed to address the imbalanced covariate shift challenge while ensuring collaborative fairness. FedAKD leverages a new **fed**erated **a**synchronous **k**nowledge **d**istillation approach, comprising client updates and server updates in each communication round t. Specifically, the *client update* includes three key steps: (1) **Global → Local Distillation:** Using traditional knowledge distillation [4], we employ the global model $\mathbf{w}_g^t$ as a teacher to guide the client model $\mathbf{w}_k^t$ in learning from its full training set $\mathcal{D}_k$. (2) **High-confidence Sample Selection:** Inspired by our observations in Figure 2, correctly

[2]Details on the EHR dataset and the preliminary analysis are provided in Sections 5.1 and 2, respectively.



(a) Correct Data Subset (b) Wrong Data Subset (c) KL divergence for "right" vs.\ "wrong" samples

Figure 2: (a) Distribution of locally versus globally correct samples. (b) Distribution of locally versus globally incorrect samples. (c) KL divergence for "right" vs. "wrong" samples. We observe that the feature distributions of correctly classified samples closely resemble the global distribution, whereas those of misclassified samples deviate significantly. This suggests that imbalanced covariate shift primarily arises from incorrectly classified samples.

predicted samples positively contribute to the global model update. Thus, we first identify the correctly classified samples from the updated local model, denoted as $\mathcal{I}_k^t$. (3) **Local → Global Distillation**: The global model $\mathbf{w}_g^t$ is then refined by distilling "high-confidence" client knowledge from $\mathbf{w}_k^t$ using each client's correctly predicted set $\mathcal{I}_k^t$. This design helps mitigate distortions caused by misclassified data under covariate shift conditions. The updated global model for each client (i.e., $\mathbf{w}_{g,k}^t$) is uploaded to the server, where it is aggregated following FedAvg [15] in the *server update* step. These two updates iterate until FedAKD converges. The theoretical convergence of FedAKD is established in Theorem 3.2.

Through this two-stage asynchronous distillation process, FedAKD effectively encourages fair collaboration even when clients have highly divergent feature distributions. High-quality participants benefit by sharing more correct samples, whereas lower-quality participants gain from the improved global knowledge, thus collectively promoting *collaborative fairness* without imposing rigid or impractical contribution measurements.

Contributions. The main contributions of this work include:

- **A new heterogeneity setting.** We introduce collaborative fairness under the imbalanced covariate shift, a practical challenge in real-world medical datasets where both sample-size imbalance and feature-distribution mismatch significantly hinder the effectiveness of existing collaborative fairness metrics.
- **A simple yet effective solution.** We propose FedAKD, a novel asynchronous distillation framework that first distills knowledge from correctly predicted local samples to improve the global model quality, followed by an inverse distillation step to enhance client learning across the full dataset.
- **Theoretical guarantees.** We provide a rigorous theoretical analysis of imbalanced covariate shift, expanding the KL divergence parametrically to model real-world heterogeneity. Additionally, we prove the convergence of FedAKD under broad heterogeneity conditions, ensuring both theoretical soundness and improved collaborative fairness.
- **Promising results.** We conduct extensive experiments on three datasets, evaluating FedAKD against ten baselines across four heterogeneity settings using three evaluation metrics. The results demonstrate that FedAKD effectively addresses imbalanced covariate shift and outperforms all baselines across diverse heterogeneity scenarios.

2 IMBALANCED COVARIATE SHIFT ANALYSIS

Imbalanced covariate shift presents a significant and open challenge in federated learning, requiring a deeper mathematical understanding to effectively mitigate disparities in collaborative fairness. To formalize this, we assume the existence of a global data feature distribution p_ω. Each client's data distribution is modeled as a small parametric perturbation, denoted as $p_{\omega+\delta}$, where $\delta \in \mathbb{R}^N$ represents the perturbation vector and N is the model size. When a client draws an i.i.d. sample set of size A from $p_{\omega+\delta}$, it results in the empirical distribution $\widehat{p}_{\omega+\delta}$. Consequently, the imbalanced covariate shift can be quantitatively assessed by measuring the KL divergence between the empirical distribution $\widehat{p}_{\omega+\delta}$ and the original global distribution p_ω.

Theorem 2.1 (Ideal Imbalance Covariate Shift Quantification). *Let $\{p_\theta\}_{\theta\in\Theta}$ be a smooth parametric family of probability distributions, and let $\omega \in \Theta$ denote a baseline parameter. Suppose that a perturbed distribution $p_{\omega+\delta}$ is defined by a small perturbation $\delta \in \mathbb{R}^N$. Given that $p_{\omega+\delta}$ is estimated via an empirical distribution $\widehat{p}_{\omega+\delta}$ using A i.i.d. samples and R (free) parameters, we have the following approximation under standard regularity conditions and a large-sample limit:*

$$D_{\mathrm{KL}}\Big(\widehat{p}_{\omega+\delta} \,\Big\|\, p_\omega\Big) \;\approx\; \frac{1}{2}\delta^\top I(\omega)\,\delta + \frac{1}{2}\delta^\top(\nabla_\omega I(\omega)\,\delta)\delta \;+\; \frac{R}{2A},$$

where $I(\omega)$ is the Fisher information matrix at ω, and $\nabla_\omega I(\omega)$ denotes the gradients with respect to ω.

We assume that the probabilistic distribution follows an M-dimensional Gaussian distribution. We then extend Theorem 2.1 as follows:

Theorem 2.2 (Imbalance Covariate Shift Quantification Under Gaussian Distribution). *Let $p_{\mu,\Sigma}(x) = \mathcal{N}(x;\mu,\Sigma)$ be a M-dimensional Gaussian distribution, where $\mu \in \mathbb{R}^M$ is the mean vector and $\Sigma \in \mathbb{R}^{M\times M}$ is a symmetric positive-definite covariance matrix. Suppose $\omega = (\mu_0, \Sigma_0)$ is a baseline parameter, and consider a small perturbation $\delta = (\delta_\mu, \delta_\Sigma), \theta' = \omega + \delta \approx (\mu_0 + \delta_\mu, \Sigma_0 + \delta_\Sigma)$. Under a large-sample limit, the Kullback-Leibler divergence between*

the empirical distribution $\widehat{p}_{\theta'}$ (fitted from A i.i.d. samples drawn from $p_{\theta'}$) and the baseline model p_ω can be approximated by:

$$D_{\mathrm{KL}}(\widehat{p}_{\theta'} \| p_\omega) \approx \frac{1}{4}\|\Sigma^{-1}\delta_\Sigma\Sigma^{-1}\|_F^2 - \frac{1}{2}\mathrm{trace}\Big((\delta_\Sigma\Sigma^{-1})^3\Big) + \frac{1}{2}(\delta_\mu)^\top(\Sigma^{-1}(I-\delta_\Sigma\Sigma^{-1}))\delta_\mu + \frac{M(M+3)}{4A}. \tag{1}$$

Together, these two theorems provide a unified mathematical framework to model the combined effects of *imbalanced sample sizes* and *covariate shift*. This framework provides a principled approach to analyzing how local client distributions diverge from the global baseline, offering deeper insights into collaborative fairness in federated learning. Detailed proofs are provided in **Appendix A**. We also provide a theoretical approximation validation experiment in **Appendix B** to validate the correctness of our theorems.

3 THE PROPOSED FEDAKD

While our theoretical analysis in Section 2 provides valuable insights into imbalanced covariate shift, it cannot be directly applied to model design, as the global distribution remains ***unknown*** in federated learning. However, these theorems reveal that the imbalanced covariate shift arises due to small perturbations in client distributions. Our preliminary analysis (Figure 2 in Section 1) further suggests that these perturbations predominantly stem from incorrectly classified samples. This observation motivates us to develop an effective collaborative fairness approach named FedAKD that mitigates the imbalanced covariate shift by addressing the impact of misclassified samples via a simple asynchronous knowledge distribution strategy. The algorithm is shown in Algorithm 1.

Similar to existing collaborative fairness approaches in federated learning, FedAKD comprises both client and server updates. However, unlike prior methods that require carefully designing reward weights for each client [13, 19, 20, 23], FedAKD simplifies aggregation by directly following the standard FedAvg [15] in the server update. The novelty of FedAKD lies in the client update, where we introduce a new **asynchronous knowledge distillation** strategy, inspired by our preliminary analysis. Specifically, the client update consists of three key steps: (1) global → local distillation, (2) high-confidence sample selection, and (3) local → global distillation. Next, we provide the details of these three steps.

Algorithm 1 FedAKD

Require: K clients; local datasets $\{\mathcal{D}_k\}$; total rounds T; learning rate η; distillation coefficients α and β.

Initialization:
  Generate an initial model $\mathbf{w}^0$ (e.g., randomly);
  Client side (for each k): set each local model $\mathbf{w}_k^0 = \mathbf{w}^0$;
  Server side: set the global model $\mathbf{w}_g^1 = \mathbf{w}^0$ and distribute $\mathbf{w}_g^1$ to clients;
for $t = 1, \cdots, T$ **do**
  // **Client Update**
  for $k = 1, 2, \cdots, K$ **do**
    // Step 1: Global → Local Distillation
    $\mathbf{w}_{g,k}^t \leftarrow \mathbf{w}_g^t$;
    // Loss computation by fixing $\mathbf{w}_{g,k}^t$ using $\mathcal{D}_k$
    $\overrightarrow{\mathcal{L}_t} = \mathrm{CE}(\mathbf{w}_k^{t-1}; \mathcal{D}_k) + \alpha\mathrm{KD}(\mathbf{w}_k^{t-1}; \mathbf{w}_{g,k}^t, \mathcal{D}_k)$;
    //update model parameters
    $\mathbf{w}_k^t \leftarrow \mathbf{w}_k^{t-1} - \eta\nabla\overrightarrow{\mathcal{L}_t}$;
    // Step 2: High-confidence Sample Selection
    $\mathcal{I}_k^t = \Big\{(\mathbf{x}, y) \in \mathcal{D}_k \,\Big|\, \mathrm{Pred}(\mathbf{w}_k^t, \mathbf{x}) = y\Big\}$;
    // Step 3: Local → Global Distillation
    // Loss computation by fixing $\mathbf{w}_k^t$ using $\mathcal{I}_k^t$
    $\overleftarrow{\mathcal{L}_t} = \mathrm{CE}(\mathbf{w}_{g,k}^t; \mathcal{I}_k^t) + \beta\mathrm{KD}(\mathbf{w}_{g,k}^t; \mathbf{w}_k^t, \mathcal{I}_k^t)$;
    // Update model parameters
    $\mathbf{w}_{g,k}^{t+1} \leftarrow \mathbf{w}_{g,k}^t - \eta\nabla\overleftarrow{\mathcal{L}_t}$;
    Upload $\mathbf{w}_{g,k}^{t+1}$ to the server;
  end for
  // **Server Update**
  $\mathbf{w}_g^{t+1} = \frac{1}{\sum_{k=1}^K|\mathcal{D}_k|}\sum_{k=1}^K\Big(|\mathcal{D}_k|\,\mathbf{w}_{g,k}^{t+1}\Big)$;
  Distribute $\mathbf{w}_g^{t+1}$ to each client;
end for
Output: The global model $\mathbf{w}_g^T$ and local models $\{\mathbf{w}_k^T\}$.

3.1 Step 1: Global → Local Distillation

The global model $\mathbf{w}_g^t$ contains aggregated knowledge, but forcing all clients to adopt $\mathbf{w}_{g,k}^t = \mathbf{w}_g^t$ directly may harm local performance due to the imbalanced covariate shift. To avoid this issue, we propose global → local distillation, enabling each client to *selectively* adopt global insights while retaining local specialization by optimizing the following loss:

$$\overrightarrow{\mathcal{L}_t} = \mathrm{CE}(\mathbf{w}_k^{t-1}; \mathcal{D}_k) + \alpha\,\mathrm{KD}(\mathbf{w}_k^{t-1}; \mathbf{w}_{g,k}^t, \mathcal{D}_k), \tag{2}$$

where CE denotes the cross-entropy loss, KD is the knowledge distillation loss, and α is the hyperparameter. The gradient update yields:

$$\mathbf{w}_k^t = \mathbf{w}_k^{t-1} - \eta\nabla\overrightarrow{\mathcal{L}_t}, \tag{3}$$

where η is the learning rate. In this step, we fix the global model parameters $\mathbf{w}_{g,k}^t$ and only update the client model parameters $\mathbf{w}_k^{t-1}$ using the full training set $\mathcal{D}_k$. This procedure allows each client to merge the updated global knowledge with its local parameters, safeguarding performance for distribution-mismatched (yet high-quality) clients. Consequently, no client is penalized for joining the federation, reinforcing the incentives for collaborative fairness under the imbalanced covariate shift.

3.2 Step 2: High-confidence Sample Selection

Our preliminary analysis (Figures 2 in Section 1) suggests that imbalanced covariate shift primarily arises from misclassified samples on each client, whereas correctly classified samples positively contribute to global model learning. To address this, we select high-confidence samples (i.e., correctly classified samples) to update the global model $\mathbf{w}_{g,k}^t$, which is denoted as:

$$\mathcal{I}_k^t = \Big\{(\mathbf{x}, y) \in \mathcal{D}_k \,\Big|\, \mathrm{Pred}(\mathbf{w}_k^t, \mathbf{x}) = y\Big\}. \tag{4}$$

3.3 Step 3: Local → Global Distillation

Unlike existing bidirectional knowledge distillation [8, 17] that updates two models simultaneously using the same dataset, we

propose an asynchronous knowledge distillation approach for this step. Additionally, we leverage only the selected high-confidence samples $\mathcal{I}_k^t$, enabling the local model $\mathbf{w}_k^t$ to guide the learning of the global model $\mathbf{w}_{g,k}^t$ by optimizing the following loss:

$$\overleftarrow{\mathcal{L}_t} = \mathrm{CE}(\mathbf{w}_{g,k}^t; \mathcal{I}_k^t) + \beta\,\mathrm{KD}(\mathbf{w}_{g,k}^t; \mathbf{w}_k^t, \mathcal{I}_k^t), \tag{5}$$

where β is the hyperparameter. The gradient update yields:

$$\mathbf{w}_{g,k}^{t+1} = \mathbf{w}_{g,k}^t - \eta\,\nabla\overleftarrow{\mathcal{L}_t}, \tag{6}$$

The proposed asynchronous knowledge distillation offers three key benefits: (1) *Robustness to Noisy Updates.* It protects the global model from noisy or erroneous updates by discarding locally misclassified samples. (2) *Fair Collaboration.* It promotes fairness by allowing high-quality clients—those that classify more samples correctly—to have a stronger influence without explicitly revealing their accuracy or contributions. (3) *Adaptability to Imbalanced Covariate Shift.* It ensures that even if a client's data distribution differs significantly from the global average, it can still contribute reliable knowledge.

3.4 Convergence Analysis

3.4.1 Notions and Assumptions. In this section, we consider a binary classification problem following [16] with input space $X \in \mathbb{R}^d$ and label space $Y = \{0, 1\}$. We employ a linear classification setting: for each local sample $\mathbf{x} \in X$, the logits are $z = \mathbf{x}^\top \mathbf{w}$, and the predicted probability is $\hat{y}(\mathbf{x}) = \sigma(z) = \frac{1}{1+e^{-z}}$. In our codistillation setup, the distillation temperature is set to $\tau = 1$, keeping the standard sigmoid form. We denote the cross-entropy loss by $\mathcal{L}(\mathbf{w}; \mathcal{D}) = \mathrm{CE}(\mathbf{w}; \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}_i \in \mathcal{D}}\Big[-y_i \log\big(\hat{y}(\mathbf{x}_i)\big) - \big(1 - y_i\big)\log\big(1 - \hat{y}(\mathbf{x}_i)\big)\Big]$. Here, $\mathbf{w}$ is the model parameter vector, and $\mathcal{D}$ is the training dataset consisting of samples $\mathbf{x}_i$ with labels y_i. We further introduce the KD loss, denoted by $\mathrm{KD}(\mathbf{w}, \mathbf{w}_0; \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}_i \in \mathcal{D}}\Big[-\hat{y}_0(\mathbf{x}_i)\,\log\big(\hat{y}(\mathbf{x}_i)\big) - \big(1-\hat{y}_0(\mathbf{x}_i)\big)\,\log\big(1-\hat{y}(\mathbf{x}_i)\big)\Big]$. This can also be written in expectation form as $\mathbb{E}_{\mathbf{x}_i \in \mathcal{D}}\Big[-\hat{y}_0(\mathbf{x}_i)\,\log\big(\hat{y}(\mathbf{x}_i)\big) - \big(1 - \hat{y}_0(\mathbf{x}_i)\big)\,\log\big(1 - \hat{y}(\mathbf{x}_i)\big)\Big]$. Here, $\hat{y}_0(\mathbf{x}_i)$ is the (fixed) teacher model's output, $\hat{y}_0(\mathbf{x}_i) = \sigma\big(\mathbf{x}_i^\top \mathbf{w}_0\big)$, and $\mathbf{w}_0$ denotes the teacher's parameter vector. This KD objective is equivalent (up to a constant) to minimizing the KL divergence from the teacher distribution [10, 16].

We define the notion of a γ-inexact solution following [10, 16], which quantifies the improvement made by local updates:

Definition 3.1 (γ_1-inexact solution [10, 16]). For a function $\overrightarrow{\mathcal{L}}(\mathbf{w}, \mathbf{w}_0; \mathcal{D}) = \mathcal{L}(\mathbf{w}; \mathcal{D}) + \alpha\mathrm{KD}(\mathbf{w}, \mathbf{w}_0; \mathcal{D})$, and let $\gamma_1 \in [0, 1]$. Suppose $\mathbf{w}_0$ is an initial point for $\min_{\mathbf{w}} \overrightarrow{\mathcal{L}}(\mathbf{w}; \mathbf{w}_0; \mathcal{D})$. We say $\mathbf{w}^*$ is a γ_1-inexact solution if $\big\|\nabla\overrightarrow{\mathcal{L}}(\mathbf{w}^*, \mathbf{w}_0; \mathcal{D})\big\| \le \gamma_1\big\|\nabla\overrightarrow{\mathcal{L}}(\mathbf{w}_0, \mathbf{w}_0; \mathcal{D})\big\|$.

A smaller γ_1 indicates a greater reduction in the gradient norm relative to the initial point, implying more significant local improvement. Conversely, a larger γ_1 indicates a less complete local optimization. Similarly, for the function $\overleftarrow{\mathcal{L}}$ with coefficient β. We define $\overleftarrow{\mathcal{L}}(\mathbf{w}, \mathbf{w}_0; \mathcal{D}) = \mathcal{L}(\mathbf{w}; \mathcal{D}) + \beta\,\mathrm{KD}(\mathbf{w}, \mathbf{w}_0; \mathcal{D})$, and let $\gamma_2 \in [0, 1]$. we say $\mathbf{w}^*$ is a γ_2-inexact solution with $\big\|\nabla\overleftarrow{\mathcal{L}}(\mathbf{w}^*, \mathbf{w}_0; \mathcal{D})\big\| \le \gamma_2\big\|\nabla\overleftarrow{\mathcal{L}}(\mathbf{w}_0, \mathbf{w}_0; \mathcal{D})\big\|$.

Assumption 1 (L-smoothness [10, 11]). *There exists $L > 0$ such that for all $\mathbf{w}, \mathbf{w}'$, $\|\nabla\mathcal{L}(\mathbf{w}; \mathcal{D}) - \nabla\mathcal{L}(\mathbf{w}'; \mathcal{D})\| \le L\|\mathbf{w} - \mathbf{w}'\|$.*

Assumption 2 (μ-strong convexity [10, 11]). *There exists $\mu > 0$ such that for all $\mathbf{w}, \mathbf{w}'$, $\mathcal{L}(\mathbf{w}; \mathcal{D}) \ge \mathcal{L}(\mathbf{w}'; \mathcal{D}) + \nabla\mathcal{L}(\mathbf{w}'; \mathcal{D})^\top(\mathbf{w} - \mathbf{w}') + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}'\|^2$, $\overrightarrow{\mathcal{L}}(\mathbf{w}, \mathbf{w}_0; \mathcal{D}) \ge \overrightarrow{\mathcal{L}}(\mathbf{w}', \mathbf{w}_0; \mathcal{D}) + \nabla\overrightarrow{\mathcal{L}}(\mathbf{w}', \mathbf{w}_0; \mathcal{D})^\top(\mathbf{w} - \mathbf{w}') + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}'\|^2$, and $\overleftarrow{\mathcal{L}}(\mathbf{w}, \mathbf{w}_0; \mathcal{D}) \ge \overleftarrow{\mathcal{L}}(\mathbf{w}', \mathbf{w}_0; \mathcal{D}) + \nabla\overleftarrow{\mathcal{L}}(\mathbf{w}', \mathbf{w}_0; \mathcal{D})^\top(\mathbf{w} - \mathbf{w}') + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}'\|^2$.*

Assumption 3 (Bounded Gradient Dissimilarity [10, 16]). *Let $\mathcal{D}_g := \bigcup_k \mathcal{D}_k$ be the global dataset consisting of all local datasets $\mathcal{D}_k$. For some $\epsilon > 0$, define $\mathcal{S}_c^\epsilon = \{\mathbf{w} \mid \|\nabla\mathcal{L}(\mathbf{w}; \mathcal{D}_g)\|^2 > \epsilon\}$. There exists B_ϵ such that for all $\mathbf{w} \in \mathcal{S}_c^\epsilon$, $B(\mathbf{w}) = \sqrt{\frac{\mathbb{E}_k\left[\|\nabla\mathcal{L}(\mathbf{w}; \mathcal{D}_k)\|^2\right]}{\|\nabla\mathcal{L}(\mathbf{w}; \mathcal{D}_g)\|^2}} \le B_\epsilon$.*

Here, $B(\mathbf{w})$ measures data heterogeneity across devices. If data are IID and $n_k \to \infty$, then $B(\mathbf{w}) \to 1$. Generally, $B_\epsilon \ge 1$, and larger values capture more dissimilar data distributions.

Assumption 4 (Bounded Gradient Dissimilarity on Subset). *Let $\mathcal{I} \subseteq \mathcal{D}$ be a subset. There exists $\theta \ge 0$ such that for all $\mathbf{w}$, $\big\|\nabla\mathcal{L}(\mathbf{w}; \mathcal{I}) - \nabla\mathcal{L}(\mathbf{w}; \mathcal{D})\big\| \le \theta\big\|\nabla\mathcal{L}(\mathbf{w}; \mathcal{D})\big\|$.*

This condition ensures that the gradient on a chosen subset does not deviate excessively from the gradient on the entire local dataset, thus quantifying the heterogeneity between these two distributions.

3.4.2 Main Results.

Theorem 3.2 (FedAKD Convergence). *Under Assumptions 1–4, assume that $\mathbf{w}^t$ is not a stationary solution and the loss function $\mathcal{L}$ is B-dissimilar, i.e., $B(\mathbf{w}^t) \le B$. If α, β and $\gamma := \max\{\gamma_1, \gamma_2\}$ are chosen such that $r = (\frac{4}{\beta||\Omega_2||} + \frac{4}{\alpha||\Omega_1||})B - \frac{LB^2}{2}((L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu} + (1+\theta) + \frac{(1+\gamma)\beta||\Omega_2||}{4\mu})^2 - (\frac{4}{\beta||\Omega_2||} + \frac{4}{\alpha||\Omega_1||})(r_1 + r_2)B > 0$, where $r_1 = ((L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu} + (1+\theta) + \frac{(1+\gamma)\beta||\Omega_2||}{4\mu})(1+\theta)L + \theta + (L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu})$, $r_2 = \frac{L(1+\gamma)}{\mu} + \gamma$, $\Omega_1 = \mathbb{E}_k[\mathbb{E}_{\mathbf{x}_{k,i} \in \mathcal{D}_k}[\mathbf{x}_{k,i}\,\mathbf{x}_{k,i}]]$ and $\Omega_2 = \mathbb{E}_k[\mathbb{E}_{\mathbf{x}_{k,i} \in \mathcal{I}_k^t}[\mathbf{x}_{k,i}\,\mathbf{x}_{k,i}]]$, then FedAKD satisfies*

$$\mathcal{L}(\mathbf{w}_g^{t+1}; \mathcal{D}_g) - \mathcal{L}(\mathbf{w}^*; \mathcal{D}_g) \le (1 - 2\mu r)[\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_g) - \mathcal{L}(\mathbf{w}^*; \mathcal{D}_g)].$$

Theorem 3.2 shows that the global model in FedAKD converges effectively. The detailed proof is provided in Appendix C.

4 SIMULATION EXPERIMENTS

4.1 Non-IID Settings

Since most real-world federated applications involve non-IID data distributions, particularly in the imbalanced covariate shift setting, we consider the following three non-IID client partitions in our simulation experiments:

- **Imbalanced Dataset Sizes (POW) [13, 23]:** Each client's dataset size $|\mathcal{D}_k|$ follows a power-law distribution, leading to significant disparities in data quantity across clients. However, the feature distributions remain similar across clients.
- **Balanced Covariate Shift (BCS):** Clients exhibit covariate shifts in their data feature distributions while maintaining a similar number of samples.

Table 1: Performance evaluation of simulation experiments on two datasets (mean ± std across three runs).

(a) Max Client Accuracy (%)

Method	FashionMNIST					CIFAR10				
	POW	BCS	ICS(2.0)	ICS(5.0)	ICS(10.0)	POW	BCS	ICS(2.0)	ICS(5.0)	ICS(10.0)
Standalone	88.09±0.06	96.06±0.16	96.93±0.16	96.80±0.05	96.94±0.02	54.86±1.93	54.53±0.66	58.82±1.19	59.95±1.29	56.20±1.71
FedAvg	94.47±0.46	97.50±0.14	94.57±0.32	96.23±0.16	94.32±0.13	67.08±0.17	65.80±0.65	66.08±0.15	64.68±0.60	66.67±0.95
CFFL	94.66±0.42	99.61±0.16	97.41±2.44	98.18±1.59	98.63±1.19	67.76±1.87	70.71±3.34	71.37±0.07	64.34±4.98	69.27±2.03
CGSV	96.33±0.31	98.64±0.11	98.29±0.52	98.51±0.25	97.97±0.09	76.16±2.13	76.31±0.26	76.04±2.28	72.29±2.05	78.40±1.37
FedAVE	92.42±0.93	98.72±0.75	96.44±0.76	95.46±1.83	96.78±0.14	62.39±1.56	58.27±0.09	60.53±3.10	60.79±0.86	60.96±1.26
FedSAC	96.42±0.56	96.73±0.45	95.33±0.64	97.31±1.34	96.34±0.33	65.27±0.22	63.04±2.57	65.42±0.06	64.13±0.84	65.59±0.08
pFedCK	94.15±0.18	97.52±0.37	97.11±0.15	98.62±0.54	98.25±0.87	80.13±0.87	79.54±0.31	80.88±1.24	79.99±1.46	80.71±0.98
FedDC	91.71±0.08	96.22±0.40	94.13±0.42	94.45±0.39	94.37±0.13	67.10±1.06	64.67±0.25	66.11±0.64	65.63±1.34	65.81±0.65
FedAS	88.20±0.38	98.89±0.21	96.08±0.23	96.76±0.20	96.37±0.33	75.56±1.23	73.13±0.66	74.59±0.79	74.25±1.27	75.58±1.61
FedMPR	94.80±0.23	97.72±0.21	95.20±0.67	95.93±0.33	95.88±0.43	69.06±0.36	66.40±0.57	66.03±0.54	67.84±1.43	69.03±0.80
FedAKD	**97.45**±0.09	**99.67**±0.14	**99.56**±0.04	**99.66**±0.04	**99.56**±0.11	**83.80**±1.12	**81.27**±0.19	**81.93**±1.03	**82.73**±0.38	**81.30**±1.43

(b) Average Client Accuracy (%)

Method	FashionMNIST					CIFAR10				
	POW	BCS	ICS(2.0)	ICS(5.0)	ICS(10.0)	POW	BCS	ICS(2.0)	ICS(5.0)	ICS(10.0)
Standalone	86.26±0.06	92.04±0.03	89.15±0.19	89.50±0.05	90.01±0.13	45.44±0.13	50.03±0.45	47.25±0.38	47.20±0.96	46.33±0.75
FedAvg	91.16±0.24	92.98±0.06	91.60±0.13	91.98±0.08	91.28±0.03	63.95±1.00	62.37±0.66	63.10±0.16	62.03±0.55	62.14±0.69
CFFL	87.79±0.91	93.92±0.04	89.43±0.56	88.39±0.83	87.36±0.36	53.91±3.31	62.88±4.27	46.14±2.92	48.27±3.29	54.18±2.71
CGSV	91.23±0.15	92.21±0.35	90.52±1.35	89.32±0.36	91.14±0.15	66.68±1.50	67.00±2.93	64.32±2.32	63.65±2.11	65.56±1.35
FedAVE	89.44±0.45	92.36±0.19	89.83±0.47	88.99±0.64	89.66±0.04	55.85±1.44	52.19±1.19	55.32±2.09	51.85±2.41	53.42±1.84
FedSAC	89.54±0.63	88.92±0.83	89.75±0.92	88.65±0.54	89.76±0.63	54.63±0.32	60.53±0.81	60.52±0.29	57.52±0.72	60.84±0.64
pFedCK	90.83±0.18	93.63±0.28	91.82±0.28	91.04±0.77	91.98±0.25	67.51±0.52	64.18±0.84	65.19±0.27	65.65±0.71	65.88±0.39
FedDC	87.83±0.46	91.08±1.46	90.06±0.56	90.20±0.24	90.00±0.42	55.34±0.65	53.13±1.31	52.98±0.05	52.30±1.73	55.83±0.73
FedAS	84.36±0.22	91.47±0.23	88.27±0.40	87.84±0.27	88.22±0.24	64.64±0.75	64.68±0.30	63.75±0.60	64.39±0.32	63.81±0.54
FedMPR	91.13±0.04	92.90±0.07	91.74±0.05	91.73±0.13	91.71±0.19	64.95±0.56	62.79±0.48	63.77±0.06	63.20±0.58	64.67±0.07
FedAKD	**93.50**±0.19	**95.68**±0.15	**92.62**±0.20	**92.47**±0.14	**92.92**±0.12	**70.96**±0.58	**67.59**±0.49	**68.96**±0.31	**68.50**±0.42	**68.55**±1.35

(c) Collaborative Fairness (CF) Coefficient

Method	FashionMNIST					CIFAR10				
	POW	BCS	ICS(2.0)	ICS(5.0)	ICS(10.0)	POW	BCS	ICS(2.0)	ICS(5.0)	ICS(10.0)
FedAvg	45.45±3.60	62.43±1.11	71.76±2.14	87.60±4.47	74.43±2.37	3.62±25.88	26.00±9.58	67.60±6.97	33.97±27.58	74.79±3.52
CFFL	33.85±15.44	79.80±10.74	60.80±20.34	74.22±6.33	67.01±8.69	45.90±31.33	16.94±32.09	55.03±15.49	13.17±7.51	26.82±11.03
CGSV	38.11±14.34	69.03±6.63	52.58±7.38	57.32±11.69	68.46±16.67	36.25±26.04	45.87±19.76	18.50±21.40	-4.15±18.53	5.54±22.85
FedAVE	21.43±23.31	65.10±1.47	78.15±8.32	80.02±7.00	65.79±4.98	16.60±18.27	49.08±15.03	17.07±22.03	24.56±34.43	25.51±27.50
FedSAC	28.10±7.93	79.54±5.13	48.78±14.97	83.30±5.13	71.26±4.76	62.04±12.03	43.61±10.38	56.89±7.30	23.90±12.70	20.71±19.95
pFedCK	23.15±5.26	41.09±13.56	34.15±9.36	36.65±8.61	15.27±15.31	51.14±9.15	70.78±11.15	24.64±21.54	8.76±19.93	36.83±10.66
FedDC	-40.85±18.38	-11.12±22.49	2.39±15.37	18.37±9.41	-15.32±38.39	-4.22±32.08	-9.54±31.25	-21.57±5.97	-9.38±27.68	-0.52±42.33
FedAS	45.09±2.61	75.81±4.81	27.74±4.12	79.61±3.11	70.23±3.12	61.34±6.55	74.28±9.43	70.01±6.07	60.54±4.05	76.54±4.10
FedMPR	31.33±14.55	83.20±2.59	70.43±2.21	76.66±5.21	61.59±3.03	50.93±14.09	40.63±4.25	29.78±18.15	44.96±6.39	53.06±9.69
FedAKD	**70.61**±5.82	**86.88**±0.81	**78.25**±2.02	**89.51**±2.40	**79.72**±1.47	**88.02**±2.09	**82.15**±2.23	**81.25**±1.55	**82.53**±2.42	**84.17**±2.77

- **Imbalanced Covariate Shift (ICS):** ICS combines the characteristics of POW and BCS, where client dataset sizes follow a power-law distribution (POW), and different institutions experience significant variations in feature distributions.

Additionally, we evaluate the proposed FedAKD framework under two ideal yet commonly used **label shift** settings: *Imbalanced Class Distributions (CLA)* [13, 23] and *Imbalanced Sizes & Class Distributions (DIR)* [1, 28]. The experimental setting and results for these two non-IID partitions are presented in **Appendix D**.

4.2 Federated Data Simulation

In the simulation experiments, we use two image classification datasets: **Fashion MNIST** [22] and **CIFAR10** [7]. **Fashion MNIST** contains 70,000 grayscale images (28×28) evenly split into 10 classes (e.g., T-shirt/top, trousers). **CIFAR10** consists of 60,000 color images (32×32) across 10 classes (e.g., airplane, bird). We partition the datasets into training, validation, and testing in a ratio of 7:1:2. In addition, we set the number of clients as $K = 10$. To simulate the **POW** partition, we follow a power law with an exponent of 1 to divide the global data into 10 clients. For the k-th client, its data size is $|\mathcal{D}_k| = \frac{1}{kZ}|\mathcal{D}_g|$, where $Z = \sum_{k=1}^{10} \frac{1}{k}$.

Simulating the **covariate shift** setting is nontrivial, and as far as we know, no existing work provides an automated way to generate such federated datasets. To fill this research gap, we design a novel algorithm based on Theorem 2.2 for covariate shift data generation, as shown in **Appendix E** Algorithm 2. To stimulate the **BCS** partition, we set $C = 5$ and add a perturbation δ (satisfying $\delta^\top \Sigma^{-1} \delta = 5$) to the global mean. In such a way, each client receives an equal number of samples (i.e., balanced), thus focusing on the covariate shift while keeping dataset sizes uniform. To stimulate the **ICS** partition, we run Algorithm 2 again to produce three variants with $C = 2$, $C = 5$, and $C = 10$, representing increasing levels of covariate shift. Meanwhile, the total number of samples is partitioned among the 10 clients according to the power law distribution with the exponent as 1, thus coupling imbalanced dataset sizes with covariate shifts.

4.3 Baselines

We compare our method against two categories of baselines: *Collaborative Fairness* algorithms designed for non-IID data, and *Covariate Shift* algorithms focusing on feature-level discrepancies. Specifically, we include **CGSV** [23], **CFFL** [13], **FedAVE** [19], and **FedSAC** [20], which explicitly address fairness by measuring client contributions or customizing reward allocations. Meanwhile, **FedAS** [25], **FedDC** [1], **pFedCK** [30], and **FedMPR** [2] focus on mitigating feature-level drift across clients. We also include two traditional baselines: **Standalone** (each client trains independently without aggregation) and the classic **FedAvg** [15] for federated averaging. The details of the baselines can be found in **Appendix F**.

4.4 Implementation

We implement a network consisting of two convolutional layers (each followed by batch normalization and ReLU activation), interleaved with max-pooling, and ending with a fully connected output layer for the simulation evaluation. We implement all baselines and our model in PyTorch and train them on an NVIDIA RTX A6000 GPU. All details of parameter setting can be found in **Appendix G**.

4.5 Evaluation Metrics

Due to the imbalanced covariate shift setting, each client data has a unique distribution. Thus, we conduct local evaluations and then report the average values of all the clients for **three runs**. Let $\mathbf{Acc}_p$ denote the prediction accuracy of clients after federation. Following [20], we use three metrics:

- *Average Client Accuracy*, i.e., $\frac{\sum_k \mathbf{Acc}_p[k]}{K}$;
- *Maximum Client Accuracy*, i.e, $\max(\mathbf{Acc}_p)$;
- *Collaborative Fairness (CF) Coefficient*, which reflects how uniformly performance is distributed across clients. We use the CF defined in [13], i.e, $CF = 100 \times \rho(\mathbf{Acc}_s, \mathbf{Acc}_p) \in [-100, 100]$, where $\rho(\cdot, \cdot)$ is Pearson's correlation coefficient, and $\mathbf{Acc}_s$ represents the standalone accuracy of clients.

The greater these three metric values, the better the performance.

4.6 Results of Simulation Experiments

Table 1 presents the results on FashionMNIST and CIFAR10 under three types of non-IID partitions. As shown in Table 1(a), our method consistently achieves superior or highly competitive *Max Client Accuracy* across all partitions for both datasets. Table 1(b) further highlights our approach's advantage in *Avg Client Accuracy*, particularly on the more challenging CIFAR10 dataset. The most critical metric, *Collaborative Fairness* (CF), is reported in Table 1(c), where FedAKD demonstrates significantly higher fairness under varying degrees of non-IID settings. These results collectively validate the effectiveness of FedAKD in improving both accuracy and fairness.

Table 2: Experimental results on the EHR dataset.

Method	CF	Max Acc	Avg. Acc
Standalone	–	76.27±0.13	70.41±0.04
FedAvg	-12.63±5.22	74.75±1.24	70.13±0.21
CFFL	52.63±12.43	73.26±2.49	69.57±0.92
CGSV	46.39±4.24	72.55±1.71	69.31±0.42
FedAVE	38.24±17.24	75.66±1.59	69.24±0.81
FedSAC	70.54±1.24	74.21±0.32	68.33± 0.56
pFedCK	35.28±13.74	76.87±0.18	70.01±0.10
FedDC	17.10±3.21	75.08±0.32	69.61±0.29
FedAS	8.50±3.28	74.89 ±0.29	69.05±0.25
FedMPR	-5.59±2.31	76.39±0.25	70.10±0.13
FedAKD	**78.42**±1.09	**78.98**±1.01	**71.23**±0.27

Table 3: Performance of all methods when constrained to the same wall-clock time as FedAKD (40 rounds).

Method	Round	CF	Max Acc	Avg. Acc
FedAvg	80	-12.63±5.22	74.75±1.24	70.13±0.21
CFFL	43	48.63±11.43	75.25±1.42	67.31±0.95
CGSV	61	33.11±6.31	73.62±0.55	68.84±0.84
FedAVE	39	58.62±9.35	70.35±1.24	66.52±0.99
FedSAC	65	60.22±3.21	74.02±0.44	68.89±0.21
pFedCK	58	33.51±8.51	75.01±0.43	69.11±0.30
FedDC	51	30.10±5.32	73.54±1.35	68.53±0.88
FedAS	59	15.89±4.09	72.88±1.09	68.66±0.11
FedMPR	47	8.22±4.24	74.39±0.21	69.28±0.26
FedAKD	40	**71.89**±1.82	**77.10**±0.58	**71.02**±0.53

Table 4: Ablation study on three variants of FedAKD versus the full method. CF denotes a fairness metric among clients (higher is better), whereas Max Acc and Avg. Acc refer to the maximum and average accuracies (in %), respectively, across global rounds.

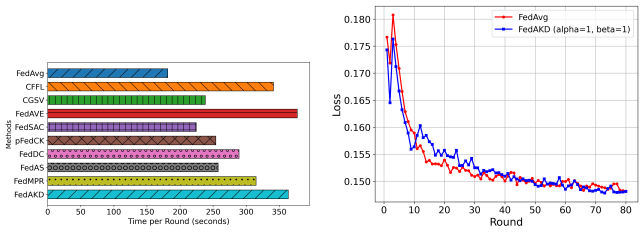
Method	CF	Max Acc	Avg. Acc
FedAKD (All-Data)	66.02±2.93	75.20±0.82	70.66±0.42
FedAKD (Single-Dist)	58.49±4.32	73.88±1.61	64.02±2.82
FedAKD (Correct-Agg.)	77.31±1.98	78.22±0.91	71.17±0.66
FedAKD (Full)	**78.42**±1.09	**78.98**±1.01	**71.23**±0.27

5 REAL-WORLD EXPERIMENTS

5.1 Experimental Settings

The EHR Dataset is extracted from the TriNetX database[3], which contains patients' claims data from all 50 states in the USA. This dataset is curated for the early prediction of pancreatic cancer and includes 259,480 *de-identified* patient records (161,345 negative vs. 98,135 positive). It consists of both static features (*sex*, *zip code*) and time-series events (*medications*, *lab tests*, *vital signs*, etc.). Further details on the EHR dataset and experimental settings are provided in **Appendix H**. In this experiment, we adopt a two-layer bidirectional GRU with attention mechanisms [3, 14, 18, 21, 26, 27, 29] to predict whether a patient eventually develops pancreatic cancer. We use the same baselines and evaluation metrics as the simulation experiments.

[3] https://trinetx.com/



(a) Average per-round runtime. **(b) Training loss on global dataset $\mathcal{D}_g$.**

Figure 3: Efficiency comparison of FedAKD and baselines.

5.2 Performance Analysis

Table 2 presents the results on the real-world EHR dataset, which naturally follows an **ICS** (Imbalanced Covariate Shift) pattern due to varying sample sizes and feature distributions across different medical institutions. As shown in the table, FedAKD significantly outperforms the baselines, particularly in the CF metric, demonstrating a more equitable distribution of benefits among clients. While FedMPR achieve competitive Max Acc, their fairness metrics remain comparatively limited.

5.3 Ablation Studies

To investigate how each component of the proposed FedAKD (Algorithm 1) contributes to its final performance, we conduct three ablation experiments. In each variant, we selectively remove or modify part of the procedure to gauge its impact: **(1) Local → Global Distillation Using All Data $\mathcal{D}_k$.** The biggest key finding of this work is to use the correctly predicted subset $\mathcal{I}_k^t$ to guide the update of $\mathbf{w}_{g,k}^t$. This baseline uses *all* local samples $\mathcal{D}_k$ for local → global distillation to test whether restricting to accurately labeled data is necessary for effective knowledge distillation. **(2) Only Local → Global Distillation.** After receiving the aggregated global model, each client simply overwrites its local model with $\mathbf{w}_g^t$, instead of performing an additional global → local distillation. This variant helps isolate the effect of double-direction distillation. **(3) Correct Sample Count for Aggregation.** In the standard procedure, the server weights each client's update by the total local dataset size. In this ablation, we replace that term with the number of *correctly predicted* local samples ($|\mathcal{I}_k^t|$), thus investigating whether "correct sample counts" lead to better aggregation fairness or performance.

The results of these ablation studies are summarized in Table 4. Our findings indicate that the proposed asynchronous knowledge distillation strategy has the most significant impact on both performance and fairness. Additionally, conducting knowledge distillation on the entire dataset negatively affects model performance. However, using the size of either the full dataset or only correctly classified data does not lead to notable performance differences. These results validate the rationale behind our model design.

5.4 Computational Cost Analysis

In our model design, the client training contains three steps, while baselines also use other techniques to calculate rewards. To validate the efficiency of our model, we show the average per-round running times (in seconds) as depicted in Figure 3a. Although FedAKD shows a slightly higher time consumption per round, primarily due to asynchronous knowledge distillation, its computational overhead is still comparable to other baselines (e.g., **CFFL** and **FedAVE**, which involve additional validation or sparsification steps).To further evaluate compute efficiency, we fixed **FedAKD** to 40 rounds, recorded its total wall-clock time, and then allowed every baseline to train for the *same* duration. Table 3 shows that—even under this strict budget—FedAKD achieves the largest fairness improvement and the highest predictive performance.

5.5 Convergence Analysis

We next analyze the convergence behavior of these methods by evaluating the global parameter $\mathbf{w}_g^t$ on the global EHR dataset, i.e., computing $\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_g)$ at each round. Figure 3b shows two representative convergence curves, comparing **FedAvg** and **FedAKD**. We observe that FedAKD converges slightly more slowly in the early stages; however, it remains stable and ultimately achieves a low global loss. This result empirically verifies Theorem 3.2, demonstrating that the global model in FedAKD converges effectively to a desirable minimum on $\mathcal{D}_g$.

6 RELATED WORK

This work mainly focuses on **collaborative fairness (CF)** in federated learning, which regards the global model as the core reward and seeks to ensure that the final performance of each client reflects its actual contribution. For instance, CGSV [23] computes a cosine gradient Shapley value to measure how closely each client's local gradient aligns with the global gradient, and allocates model updates based on this similarity. CFFL [13] relies on a *public validation set* to evaluate each client's data diversity and local-model performance, then allocates rewards accordingly. FedAVE [19] computes reputation by examining each client's local model performance and data distribution, offering better adaptability to various distributional scenarios. FedSAC [20] avoids the need for a global validation set by distributing varying submodels to high-contribution clients. However, it depends on *pre-known standalone training results*, which may be unrealistic in real-world deployments. Moreover, submodel-based pruning alone cannot fully address feature-distribution mismatch, as high-quality but distribution-mismatched data may still be underrepresented.

7 CONCLUSION

This paper investigates a practical yet challenging form of heterogeneity that impacts collaborative fairness: *imbalanced covariate shift*. To address this issue, we propose a novel approach, FedAKD (Federated Asynchronous Knowledge Distillation), which mitigates the effects of imbalanced covariate shift by excluding incorrectly predicted samples from the global model update—an insight derived from our preliminary findings. Experimental results on three datasets compared against ten baselines demonstrate the effectiveness and fairness of FedAKD across various heterogeneity settings in federated learning.

REFERENCES

[1] Liang Gao, Hongchao Fu, Lili Li, Yanyan Chen, Min Xu, and Cheng-Zhong Xu. 2022. FedDC: Federated Learning with Non-iid Data via Local Drift Decoupling and Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10112–10121.

[2] Ozgu Goksu and Nicolas Pugeault. 2024. Robust Federated Learning in the Face of Covariate Shift: A Magnitude Pruning with Hybrid Regularization Framework for Enhanced Model Aggregation. *arXiv preprint arXiv:2412.15010* (2024). https://arxiv.org/abs/2412.15010

[3] Wei Guo, Wei Ge, Longbo Cui, Hua Li, and Li Kong. 2019. An interpretable disease onset predictive model using crossover attention mechanism from electronic health records. *IEEE Access* 7 (2019), 134236–134244.

[4] Geoffrey Hinton. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[5] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sanjiv Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic Controlled Averaging for Federated Learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR, 5132–5143.

[6] Jakub Konecný. 2016. Federated Learning: Strategies for Improving Communication Efficiency. *arXiv preprint arXiv:1610.05492* (2016). arXiv:1610.05492 [cs.LG]

[7] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning Multiple Layers of Features from Tiny Images*. Technical Report. University of Toronto. Technical Report.

[8] Wonbin Kweon, SeongKu Kang, and Hwanjo Yu. 2021. Bidirectional distillation for top-K recommender system. In *Proceedings of the Web Conference 2021*. 3861–3871.

[9] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine* 37, 3 (2020), 50–60.

[10] Tian Li, Anit K. Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, Vol. 2. 429–450.

[11] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the Convergence of FedAvg on Non-IID Data. arXiv:1907.02189 [cs.LG] arXiv preprint arXiv:1907.02189.

[12] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, Piotr Dollár, and Serge Belongie. 2017. Focal Loss for Dense Object Detection. *arXiv preprint arXiv:1708.02002* (2017).

[13] Lingjuan Lyu, Xinyang Xu, Qiang Wang, Han Yu, et al. 2020. Collaborative Fairness in Federated Learning. In *Federated Learning: Privacy and Incentive*. 189–204.

[14] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1903–1911.

[15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (Proceedings of Machine Learning Research)*. PMLR, 1273–1282.

[16] Xuanming Ni, Xinyuan Shen, and Huimin Zhao. 2022. Federated optimization via knowledge codistillation. *Expert Systems with Applications* 191 (2022), 116310. https://doi.org/10.1016/j.eswa.2021.116310

[17] Ertong Shang, Hui Liu, Zhuo Yang, Junzhao Du, and Yiming Ge. 2023. FedBiKD: Federated Bidirectional Knowledge Distillation for Distracted Driving Detection. *IEEE Internet of Things Journal* (2023).

[18] Qingxiong Tan, Min Ye, Bin Yang, S. Liu, A. J. Ma, T. C. F. Yip, Y. Zhao, S. C. Hui, T. M. F. Chan, F. K. Chan, J. J. Y. Sung, E. C. Cheung, and P. Yuen. 2020. Data-GRU: Dual-Attention Time-Aware Gated Recurrent Unit for Irregular Multivariate Time Series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 930–937.

[19] Zihui Wang, Zhe Peng, Xinyu Fan, Zheng Wang, Siyang Wu, Rui Yu, ..., and Chunyan Wang. 2024. FedAVE: Adaptive data value evaluation framework for collaborative fairness in federated learning. *Neurocomputing* 574 (2024), 127227.

[20] Zihui Wang, Zheng Wang, Lingjuan Lyu, Zhigang Peng, Zhiquan Yang, Chuan Wen, and Xiaohui Fan. 2024. FedSAC: Dynamic Submodel Allocation for Collaborative Fairness in Federated Learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3299–3310.

[21] Sajila D. Wickramaratne and Md Shaad Mahmud. 2020. Bi-directional gated recurrent unit based ensemble model for the early detection of sepsis. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 70–73.

[22] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[23] Xinyi Xu, Lingjuan Lyu, Xiaofeng Ma, Chunyan Miao, Chee Seng Foo, and Bo An Kiat Huat Low. 2021. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Advances in Neural Information Processing Systems*, Vol. 34. 16104–16117.

[24] Gang Yan, Haiyan Wang, Xue Yuan, and Jia Li. 2023. Criticalfl: A critical learning periods augmented client selection framework for efficient federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, 2898–2907.

[25] Xiyuan Yang, Wenke Huang, and Mang Ye. 2024. FedAS: Bridging Inconsistency in Personalized Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 11986–11995. https://doi.org/10.1109/CVPR52733.2024.01139

[26] Yang Yang, Xiangwei Zheng, and Cun Ji. 2019. Disease prediction model based on bilstm and attention mechanism. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 1141–1148.

[27] Xiangyang Ye, Q. T. Zeng, Julio C. Facelli, Diana I. Brixner, Mike Conway, and Bradley E. Bray. 2020. Predicting optimal hypertension treatment pathways using recurrent neural networks. *International Journal of Medical Informatics* 139 (2020), 104122.

[28] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Natesh Hoang, and Yasaman Khazaeni. 2019. Bayesian Nonparametric Federated Learning of Neural Networks. In *International Conference on Machine Learning (Proceedings of Machine Learning Research)*. PMLR, 7252–7261.

[29] Jinghe Zhang, Kamran Kowsari, James H Harrison, Jason M Lobo, and Laura E Barnes. 2018. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* 6 (2018), 65333–65346.

[30] Jianfei Zhang and Yongqiang Shi. 2024. A Personalized Federated Learning Method Based on Clustering and Knowledge Distillation. *Electronics* 13, 5 (2024), 857. https://doi.org/10.3390/electronics13050857

[31] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Dave Civin, and Vikas Chandra. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582* (2018).

A IMBALANCED COVARIATE SHIFT PROOF

A.1 Proof of Theorem 2.1

We first aim to expand

$$D_{\mathrm{KL}}\big(p_{\omega+\delta}\,\big\|\,p_\omega\big) \;:=\; \mathbb{E}_{X\sim p_{\omega+\delta}}\Big[\log p_{\omega+\delta}(X) \;-\; \log p_\omega(X)\Big].$$

Observe that

$$\log p_\omega(x) \;=\; \log\big(p_{(\omega+\delta)-\delta}(x)\big) \;=\; \ell\big((\omega+\delta)-\delta,\,x\big).$$

We make a second-order Taylor expansion of $\ell(\theta-\delta,x)$ around $\theta=\omega+\delta$ at $\delta=0$. Concretely, with $\ell(\theta,x)=\log p_\theta(x)$, we have

$$\begin{aligned}\ell\big((\omega+\delta)-\delta,\,x\big) \;\approx\; \ell(\omega+\delta,x) \;-\; \delta^\top\,\nabla_\theta\ell(\theta,x)\big|_{\theta=\omega+\delta}\\ +\;\frac{1}{2}\,\delta^\top\,\nabla^2_\theta\ell(\theta,x)\big|_{\theta=\omega+\delta}\,\delta,\end{aligned}$$

$$\begin{aligned}\log p_{\omega+\delta}(x) \;-\; \log p_\omega(x) \;\approx\; \delta^\top\,\nabla_\theta\ell(\theta,x)\big|_{\theta=\omega+\delta}\\ -\;\frac{1}{2}\,\delta^\top\,\nabla^2_\theta\ell(\theta,x)\big|_{\theta=\omega+\delta}\,\delta.\end{aligned}$$

We set

$$I(\omega+\delta) \;:=\; -\,\mathbb{E}_{p_{\omega+\delta}}\Big[\nabla^2_\theta\log p_{\omega+\delta}(X)\Big].$$

Taking expectation under $X\sim p_{\omega+\delta}$ gives

$$\begin{aligned}D_{\mathrm{KL}}\big(p_{\omega+\delta}\,\big\|\,p_\omega\big) &= \mathbb{E}_{p_{\omega+\delta}}\Big[\log p_{\omega+\delta}(X)-\log p_\omega(X)\Big] &(7)\\ &\approx \mathbb{E}_{p_{\omega+\delta}}\Big[-\tfrac{1}{2}\,\delta^\top\nabla^2_\theta\ell(\theta,X)\big|_{\theta=\omega+\delta}\,\delta\Big] &(8)\\ &= \tfrac{1}{2}\,\delta^\top I(\omega+\delta)\,\delta &(9)\\ &\approx \tfrac{1}{2}\,\delta^\top I(\omega)\,\delta+\tfrac{1}{2}\,\delta^\top\nabla_\omega I(\omega)\,\delta\delta &(10)\end{aligned}$$

We want to compute or approximate the following KL divergence:

$$D_{\mathrm{KL}}\big(\widehat{p}_{\omega+\delta}\,\big\|\,p_{\omega+\delta}\big),$$

where $\widehat{p}_{\omega+\delta}$ is the empirical distribution (or fitted model) from A i.i.d. samples $X_1,\ldots,X_A$ drawn from the "true" distribution $p_{\omega+\delta}$. To simplify notation, let us define:

$\theta^* \;:=\; \omega+\delta,\quad$ and $\quad\widehat{\theta}$ be the MLE fitted using A samples from p_{θ^*}.

Thus the distribution $\widehat{p}_{\omega+\delta}$ can be written as $p_{\widehat{\theta}}$ (i.e. parameterized by $\widehat{\theta}$). We often denote

$$\Delta \;:=\; \widehat{\theta} - \theta^* \;=\; \widehat{\theta} - (\omega + \delta).$$

Hence the divergence of interest is

$$D_{\mathrm{KL}}\big(p_{\widehat{\theta}} \,\big\|\, p_{\theta^*}\big).$$

Let $\ell(\theta, x) := \log p_\theta(x)$. Taylor expand $\log p_{\widehat{\theta}}(x)$ around θ^*.

$$\begin{aligned}\ell(\widehat{\theta}, x) \;&= \ell(\theta^*, x) \;+\; (\widehat{\theta} - \theta^*)^\top \nabla_\theta \ell(\theta^*, x) \\ &\quad +\; \tfrac{1}{2}\,(\widehat{\theta} - \theta^*)^\top \nabla_\theta^2 \ell(\widetilde{\theta}, x)\,(\widehat{\theta} - \theta^*) \\ &= \ell(\theta^*, x) \;+\; \Delta^\top \nabla_\theta \ell(\theta^*, x) \;+\; \tfrac{1}{2}\,\Delta^\top \nabla_\theta^2 \ell(\widetilde{\theta}, x)\,\Delta,\end{aligned}$$

where $\widetilde{\theta}$ is between θ^* and $\widehat{\theta}$.

$$\begin{aligned}\log p_{\widehat{\theta}}(x) \;-\; \log p_{\theta^*}(x) \;&= (\widehat{\theta} - \theta^*)^\top \nabla_\theta \ell(\theta^*, x) \\ &\quad +\; \tfrac{1}{2}\,(\widehat{\theta} - \theta^*)^\top \nabla_\theta^2 \ell(\widetilde{\theta}, x)\,(\widehat{\theta} - \theta^*).\end{aligned}$$

$$\mathbb{E}_{p_{\widehat{\theta}}}\big[\log p_{\widehat{\theta}}(X) - \log p_{\theta^*}(X)\big] = \mathbb{E}_{p_{\widehat{\theta}}}\Big[\Delta^\top \nabla_\theta \ell(\theta^*, X) + \tfrac{1}{2}\,\Delta^\top \nabla_\theta^2 \ell(\widetilde{\theta}, X)\,\Delta\Big]. \tag{11}$$

We deal with the first-order term $\Delta^\top \nabla_\theta \ell(\theta^*, X)$. Since Δ does not depend on X, we have

$$\Delta^\top \mathbb{E}_{p_{\widehat{\theta}}}\big[\nabla_\theta \ell(\theta^*, X)\big] \approx \Delta^\top \mathbb{E}_{p_{\theta^*}}\big[\nabla_\theta \ell(\theta^*, X)\big] = \Delta^\top \mathbf{0} \;=\; 0,$$

where we used $\mathbb{E}_{p_{\theta^*}}[\nabla_\theta \ell(\theta^*, X)] = 0$ and $p_{\widehat{\theta}} \approx p_{\theta^*}$ for large a. Consider the remaining piece in Eq. (11):

$$\tfrac{1}{2}\,\mathbb{E}_{p_{\widehat{\theta}}}\big[\Delta^\top \nabla_\theta^2 \ell(\widetilde{\theta}, X)\,\Delta\big].$$

When $\widehat{\theta} \approx \theta^*$, we have $\nabla_\theta^2 \ell(\widetilde{\theta}, X) \approx \nabla_\theta^2 \ell(\theta^*, X)$, and

$$\mathbb{E}_{X \sim p_{\theta^*}}\big[\nabla_\theta^2 \ell(\theta^*, X)\big] = -\,I(\theta^*).$$

Hence

$$\mathbb{E}_{p_{\widehat{\theta}}}\big[\Delta^\top \nabla_\theta^2 \ell(\widetilde{\theta}, X)\,\Delta\big] \;\approx\; -\,\Delta^\top I(\theta^*)\,\Delta.$$

Recall the asymptotic distribution:

$$\Delta \;=\; \widehat{\theta} - \theta^* \;\approx\; \frac{1}{\sqrt{A}}\,\mathcal{N}\big(0,\; I(\theta^*)^{-1}\big).$$

Thus

$$\begin{aligned}\mathbb{E}\big[\Delta^\top I(\theta^*)\,\Delta\big] &= \mathrm{trace}\Big[I(\theta^*)\,\mathbb{E}(\Delta\Delta^\top)\Big] \\ &= \mathrm{trace}\Big[I(\theta^*)\,\tfrac{1}{A}\,I(\theta^*)^{-1}\Big] \\ &= \tfrac{1}{A}\,\mathrm{trace}(I_R) \\ &= \tfrac{R}{A}.\end{aligned}$$

Therefore,

$$\tfrac{1}{2}\big[-\,\Delta^\top I(\theta^*)\,\Delta\big] \approx -\,\tfrac{1}{2}\,\frac{R}{A}.$$

So we conclude the derivation that

$$D_{\mathrm{KL}}\big(\widehat{p}_{\omega+\delta} \,\big\|\, p_{\omega+\delta}\big) \;\approx\; \frac{R}{2A} \quad \text{for large sample size } A. \tag{12}$$

Due to Eq. (12) and Eq. (10), we can get

$$D_{\mathrm{KL}}\big(\widehat{p}_{\omega+\delta} \,\big\|\, p_{\omega}\big) \;\approx\; \tfrac{1}{2}\,\delta^\top I(\omega)\,\delta + \tfrac{1}{2}\,\delta^\top \nabla_\omega I(\omega)\,\delta\delta + \frac{R}{2A} + e \tag{13}$$

$$\approx\; \tfrac{1}{2}\,\delta^\top I(\omega)\,\delta + \tfrac{1}{2}\,\delta^\top \nabla_\omega I(\omega)\,\delta\delta + \frac{R}{2A}, \tag{14}$$

where $e = \int \big[\widehat{p}_{\omega+\delta}(x) \;-\; p_{\omega+\delta}(x)\big]\,\log\!\Big(\frac{p_\omega(x)}{p_{\omega+\delta}(x)}\Big)\,\mathrm{d}x \approx 0$.

A.2 Proof of Theorem 2.2

For the family $p_{\mu,\Sigma}(x) = \mathcal{N}(x;\,\mu,\Sigma)$, the Fisher information matrix $I(\mu,\Sigma)$ can be written in a block-diagonal form:

$$I(\mu,\Sigma) = \begin{pmatrix}\Sigma^{-1} & 0 \\ 0 & \tfrac{1}{2}\,(\Sigma^{-1} \otimes \Sigma^{-1})\end{pmatrix},$$

where $\otimes$ is the Kronecker product. Thus, if we set

$$\delta \;=\; (\delta_\mu,\; \mathrm{vec}(\delta_\Sigma)),$$

$$I(\omega) \;=\; \begin{pmatrix}I_{\mu,\mu} & 0 \\ 0 & I_{\Sigma,\Sigma}\end{pmatrix} \;=\; \begin{pmatrix}\Sigma^{-1} & 0 \\ 0 & \tfrac{1}{2}\,(\Sigma^{-1} \otimes \Sigma^{-1})\end{pmatrix},$$

We decompose the following expression into three parts:

$$\underbrace{\frac{1}{2}\,\delta^\top I(\omega)\,\delta}_{\textbf{(A)}} \;+\; \underbrace{\frac{1}{2}\,\delta^\top \nabla_\omega I(\omega)\,\delta\,\delta}_{\textbf{(B)}} \;+\; \underbrace{\frac{R}{2A}}_{\textbf{(C)}}$$

For the term (A) $= \tfrac{1}{2}\,\delta^\top I(\omega)\,\delta$, since $I(\omega)$ is block-diagonal in (μ,Σ), then we have

$$\tfrac{1}{2}\,\delta^\top I(\omega)\,\delta = \tfrac{1}{2}\,(\delta_\mu)^\top \Sigma^{-1}\,\delta_\mu \;+\; \tfrac{1}{2}\,\big(\mathrm{vec}(\delta_\Sigma)\big)^\top \Big[\tfrac{1}{2}\big(\Sigma^{-1} \otimes \Sigma^{-1}\big)\Big]\mathrm{vec}(\delta_\Sigma) \tag{15}$$

$$= \tfrac{1}{2}\,(\delta_\mu)^\top \Sigma^{-1}\,\delta_\mu \;+\; \tfrac{1}{4}\,\big(\mathrm{vec}(\delta_\Sigma)\big)^\top \big(\Sigma^{-1} \otimes \Sigma^{-1}\big)\mathrm{vec}(\delta_\Sigma). \tag{16}$$

Recall that $\delta_\Sigma \in \mathbb{R}^{d\times d}$ is the matrix-form perturbation to Σ, while $\mathrm{vec}(\delta_\Sigma) = \mathrm{vec}(\delta_\Sigma) \in \mathbb{R}^{d^2}$ is its vectorization. A standard identity for the Frobenius norm is

$$\big(\mathrm{vec}(\delta_\Sigma)\big)^\top \big(\Sigma^{-1} \otimes \Sigma^{-1}\big)\mathrm{vec}(\delta_\Sigma) \;=\; \|\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\|_F^2.$$

Therefore,

$$\tfrac{1}{2}\,\delta^\top I(\omega)\,\delta = \tfrac{1}{2}\,(\delta_\mu)^\top \Sigma^{-1}\,\delta_\mu \;+\; \tfrac{1}{4}\,\|\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\|_F^2. \tag{17}$$

For the term (B) $= \tfrac{1}{2}\,\delta^\top(\nabla_\omega I(\omega)\,\delta)\delta$, because $I(\mu,\Sigma)$ is block-diagonal, its derivative with respect to (μ,Σ) also splits into two blocks. Consequently, we can separate:

$$\nabla_\omega I(\omega) \;=\; \begin{pmatrix}\nabla_\omega\big[\Sigma^{-1}\big] & 0 \\ 0 & \nabla_\omega\big[\tfrac{1}{2}(\Sigma^{-1} \otimes \Sigma^{-1})\big]\end{pmatrix}.$$

B1 denotes contribution from (μ,μ) block:

$$B_1 = \tfrac{1}{2}\,\delta_\mu^\top \Big[\nabla_\Sigma\big(\Sigma^{-1}\big)(\delta_\Sigma)\Big]\,\delta_\mu.$$

We have $I_{\mu,\mu}(\Sigma) = \Sigma^{-1}$. Its derivative with respect to Σ (a matrix perturbation δ_Σ) is

$$\nabla_\Sigma\big[\Sigma^{-1}\big](\delta_\Sigma) \;=\; -\,\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}.$$

Inserting $\delta_\mu^\top$ and δ_μ then yields

$$B_1 \;=\; -\tfrac{1}{2}\,\delta_\mu^\top\Big(\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\Big)\,\delta_\mu. \tag{18}$$

This term vanishes if $\delta_\Sigma = 0$ or if one chooses to ignore δ_μ, but is otherwise a valid second-order effect.

B2 denotes contribution from (Σ, Σ) block:

$$B_2 = \tfrac{1}{2}\,\big(\mathrm{vec}(\delta_\Sigma)\big)^\top\Big[\nabla_\Sigma I_{\Sigma,\Sigma}(\Sigma)\big(\delta_\Sigma\big)\Big]\mathrm{vec}(\delta_\Sigma) \tag{19}$$

As in the main text, we set

$$I_{\Sigma,\Sigma}(\Sigma) \;=\; \tfrac{1}{2}\,\big(\Sigma^{-1}\otimes\Sigma^{-1}\big).$$

Hence we define

$$\nabla_\Sigma I_{\Sigma,\Sigma}(\Sigma) \;=\; \tfrac{1}{2}\,\nabla_\Sigma\big(\Sigma^{-1}\otimes\Sigma^{-1}\big).$$

Suppose $\delta_\Sigma \in \mathbb{R}^{d\times d}$ is a small matrix perturbation of Σ, and let $\mathrm{vec}(\delta_\Sigma) = \mathrm{vec}(\delta_\Sigma)$ be its vectorized form. Then the second-order expansion term of interest is

$$\tfrac{1}{2}\,\delta^\top\nabla_\omega\Big[\tfrac{1}{2}(\Sigma^{-1}\otimes\Sigma^{-1})\Big]\,\delta\,\delta \;=\; \tfrac{1}{2}\,\big(\mathrm{vec}(\delta_\Sigma)\big)^\top\Big[\nabla_\Sigma I_{\Sigma,\Sigma}(\Sigma)\big(\delta_\Sigma\big)\Big]\,\mathrm{vec}(\delta_\Sigma), \tag{20}$$

where we note that δ_Σ in the original notation is replaced by δ_Σ in matrix form (or $\mathrm{vec}(\delta_\Sigma)$ in vectorized form).

By a standard matrix-calculus result,

$$\delta\big(\Sigma^{-1}\otimes\Sigma^{-1}\big) \;=\; -\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\otimes\Sigma^{-1} \;-\; \Sigma^{-1}\otimes\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1},$$

where δ_Σ appears inside each product as the perturbation in matrix form. Hence,

$$\nabla_\Sigma\big(\Sigma^{-1}\otimes\Sigma^{-1}\big)\big(\delta_\Sigma\big) \;=\; -\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\otimes\Sigma^{-1} \;-\; \Sigma^{-1}\otimes\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}.$$

Since $I_{\Sigma,\Sigma}(\Sigma)$ has the extra factor $\frac{1}{2}$, we obtain

$$\nabla_\Sigma I_{\Sigma,\Sigma}(\Sigma)\big(\delta_\Sigma\big) \;=\; \tfrac{1}{2}\Big[-\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\otimes\Sigma^{-1} \;-\; \Sigma^{-1}\otimes\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\Big]. \tag{21}$$

We take Eq. (21) into Eq. (20),

$$\begin{aligned}
B_2 =& \tfrac{1}{2}\,\big(\mathrm{vec}(\delta_\Sigma)\big)^\top\Big[\nabla_\Sigma I_{\Sigma,\Sigma}(\Sigma)\big(\delta_\Sigma\big)\Big]\mathrm{vec}(\delta_\Sigma) \\
=& \tfrac{1}{2}\,\big(\mathrm{vec}(\delta_\Sigma)\big)^\top\Big[\tfrac{1}{2}\Big(-\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\otimes\Sigma^{-1} \qquad (22)\\
&- \Sigma^{-1}\otimes\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\Big)\Big]\mathrm{vec}(\delta_\Sigma) \\
=& \tfrac{1}{4}\,\big(\mathrm{vec}(\delta_\Sigma)\big)^\top\Big[-\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\otimes\Sigma^{-1} \qquad (23)\\
&- \Sigma^{-1}\otimes\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\Big]\mathrm{vec}(\delta_\Sigma) \\
=& -\tfrac{1}{4}\big(\mathrm{vec}(\delta_\Sigma)\big)^\top\Big[\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\otimes\Sigma^{-1} \qquad (24)\\
&+ \Sigma^{-1}\otimes\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\Big]\mathrm{vec}(\delta_\Sigma). \qquad (25)\\
\overset{(a)}{=}& -\tfrac{1}{2}\,\mathrm{trace}\Big(\delta_\Sigma\,\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\Big) \qquad (26)\\
=& -\tfrac{1}{2}\,\mathrm{trace}\Big((\delta_\Sigma\,\Sigma^{-1})^3\Big), \qquad (27)
\end{aligned}$$

where (a) is due to $(\mathrm{vec}(X))^\top(Y\otimes Z)\,(\mathrm{vec}(X)) = \mathrm{trace}[ZXYX^\top]$.

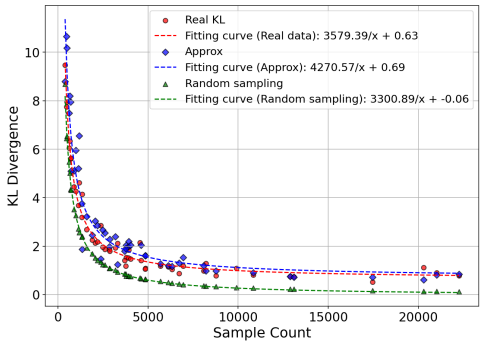


Figure 4: Comparison of KL divergences on real data (red), the Theorem 2.2 approximation (blue), and a random-subsampling baseline (green). Our approximation aligns well with the actual KL divergence, while random subsampling underestimates the real shift.

For (C), since the M-dimensional Gaussian distribution, we can obtain

$$\begin{aligned}
R &= \underbrace{M}_{\text{Mean}} + \underbrace{\frac{M(M+1)}{2}}_{\text{Covariance}} \\
&= M + \frac{M(M+1)}{2} \\
&= \frac{M^2+3M}{2}.
\end{aligned}$$

So we take this term into C,

$$C = \frac{R}{2\,A} = \frac{M(M+4)}{4A} \tag{28}$$

We take Eq. (28), Eq. (27), Eq. (18) and Eq. (17) into Eq. (15):

$$\begin{aligned}
D_{\mathrm{KL}}\big(\widehat{p}_{\omega+\delta}\,\big\|\,p_\omega\big) \;\approx\; & \tfrac{1}{2}\,(\delta_\mu)^\top\Sigma^{-1}\,\delta_\mu \;+\; \tfrac{1}{4}\,\|\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\|_F^2 + \frac{M(M+4)}{4A} \\
& - \tfrac{1}{2}\,\delta_\mu^\top\Big(\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\Big)\,\delta_\mu - \tfrac{1}{2}\,\mathrm{trace}\Big((\delta_\Sigma\,\Sigma^{-1})^3\Big) \\
=& \tfrac{1}{2}\,(\delta_\mu)^\top(\Sigma^{-1}(I-\delta_\Sigma\Sigma^{-1}))\,\delta_\mu + \tfrac{1}{4}\,\|\Sigma^{-1}\,\delta_\Sigma\,\Sigma^{-1}\|_F^2 \\
& - \tfrac{1}{2}\,\mathrm{trace}\Big((\delta_\Sigma\,\Sigma^{-1})^3\Big) + \frac{M(M+4)}{4A}
\end{aligned}$$

B THEORETICAL APPROXIMATION VALIDATION

To validate our theoretical approximations, we conduct experiments on the real EHR dataset by first training a Variational Autoencoder (VAE) to embed each client's data into an M-dimensional latent space, where $M = 100$. We then use these client representations to approximate a Gaussian distribution for each client. Similarly, we aggregate all client representations to fit a global Gaussian distribution. The results are shown in Figure 4. In Figure 4, "*Real KL*" means the empirical KL divergence between each client's approximated distribution and the global fitted distribution. "*Approx*" refers to our theoretical approximation, where small Gaussian perturbations are added to the mean vector and covariance matrix of each client's approximated distribution, allowing the KL divergence to

be computed directly via Eq. (1). "*Random sampling*" serves as a control experiment, where each client's sample size is matched by randomly subsampling from the global distribution (i.e., without introducing any intentional distributional shift). The results indicate that our theoretical approximation (blue curve) closely follows the empirical KL values (red curve) across clients. Additionally, the random sampling baseline (green curve) remains significantly lower, confirming that the real dataset exhibits substantial covariate shift effects beyond simple sample-size imbalance. These empirical findings validate the correctness of our theorems and provide a principled basis for constructing simulation datasets to evaluate model performance under the new setting.

C FEDAKD CONVERGENCE PROOF

Global → Local KD. We compute the gradient of the distillation loss where the global model serves as the teacher and the local model serves as the student. The parameters are given by:

$$\mathrm{KD} \leftarrow \mathrm{KD}\big(\mathbf{w},\, \mathbf{w}_g^{t+1};\, \mathcal{D}_k\big).$$

$$\begin{aligned}
\nabla \mathrm{KD} =& \nabla \mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}\left[-\sigma(\mathbf{x}_{k,i}^\top \mathbf{w}_g^{t+1})\log(\sigma(\mathbf{x}_{k,i}^\top \mathbf{w})) \right. \\
& \left. -(1-\sigma(\mathbf{x}_i^\top \mathbf{w}_g^{t+1}))\log(1-\sigma(\mathbf{x}_i^\top \mathbf{w}))\right] && (29)\\
\overset{(a)}{=}& \mathbb{E}_{x_{k,i}\in\mathcal{D}_k}\left[\big(\sigma\big(\mathbf{x}_{k,i}^\top \mathbf{w}\big)-\sigma\big(\mathbf{x}_{k,i}^\top \mathbf{w}_g^{t+1}\big)\big)\mathbf{x}_{k,i}\right] && (30)\\
\overset{(b)}{=}& \mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}\left[\sigma(\xi_{k,i})\left(1-\sigma(\xi_{k,i})\right)\left(\mathbf{x}_{k,i}^\top\left(\mathbf{w}-\mathbf{w}_g^{t+1}\right)\right)\mathbf{x}_{k,i}\right] && (31)\\
=& \mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}\left[\sigma(\xi_{k,i})\left(1-\sigma(\xi_{k,i})\right)\mathbf{x}_{k,i}\mathbf{x}_{k,i}^\top\right]\left(\mathbf{w}-\mathbf{w}_g^{t+1}\right) && (32)
\end{aligned}$$

where

- (a): This step employs the standard derivative of the binary cross-entropy term $-p\,\log(q) - (1-p)\,\log(1-q)$. Differentiation with respect to q, and then accounting for the chain rule via the feature vector, yields $(q-p)\,\mathbf{x}$.
- (b): This step applies the Mean Value Theorem (MVT) to the sigmoid difference: $\sigma\big(\mathbf{x}_{k,i}^\top \mathbf{w}\big) - \sigma\big(\mathbf{x}_{k,i}^\top \mathbf{w}_g^{t+1}\big) = \sigma'(\xi_1)\left[\mathbf{x}_{k,i}^\top\big(\mathbf{w}-\mathbf{w}_g^{t+1}\big)\right]$, for some ξ_1 between $\mathbf{x}_{k,i}^\top \mathbf{w}$ and $\mathbf{x}_{k,i}^\top \mathbf{w}_g^{t+1}$. Since $\sigma'(z) = \sigma(z)\,[1-\sigma(z)]$, we use $\sigma(\xi_1)\,[1-\sigma(\xi_1)]$ as a factor.

Local → Global KD. Then we compute the distillation loss definition where the local model serves as the teacher and the global model serves as the student. The parameters are given by:

$$\mathrm{KD} \leftarrow \mathrm{KD}\big(\mathbf{w},\, \mathbf{w}_k^{t};\, \mathcal{I}_k^{t}\big).$$

Based on Eq. (32), we can similarly derive the following equation:

$$\nabla \mathrm{KD} = \mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}\left[\sigma(\zeta_{k,i})\left(1-\sigma(\zeta_{k,i})\right)\mathbf{x}_{k,i}\mathbf{x}_{k,i}^\top\right]\left(\mathbf{w}-\mathbf{w}_k^{t}\right)$$

We define two errors:

$$\begin{aligned}
e_k^t =& \nabla\overrightarrow{\mathcal{L}}(\mathbf{w}_k^t, \mathbf{w}_g^t;\, \mathcal{D}_k) && (33)\\
=& \nabla\mathcal{L}(\mathbf{w}_k^t;\mathcal{D}_k) + \alpha\mathcal{L}_{KD}\big(\mathbf{w}_k^t,\, \mathbf{w}_g^t;\, \mathcal{D}_k\big) && (34)\\
=& \nabla\mathcal{L}(\mathbf{w}_k^t;\mathcal{D}_k) + \alpha\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}\left[\sigma(\xi_{k,i})\left(1-\sigma(\xi_{k,i})\right)\right. && (35)\\
& \left.\cdot\,\mathbf{x}_{k,i}\mathbf{x}_{k,i}^\top\right]\left(\mathbf{w}_k^t-\mathbf{w}_g^t\right),
\end{aligned}$$

and

$$\begin{aligned}
e_{g,k}^{t+1} =& \nabla\overleftarrow{\mathcal{L}}(\mathbf{w}_{g,k}^{t+1}, \mathbf{w}_k^t;\, \mathcal{I}_k^t) && (36)\\
=& \nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1};\mathcal{I}_k^t) + \beta\mathcal{L}_{KD}\big(\mathbf{w}_{g,k}^{t+1},\, \mathbf{w}_k^t;\, \mathcal{I}_k^t\big) && (37)\\
=& \nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1};\mathcal{I}_k^t) + \beta\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}\left[\sigma(\zeta_{k,i})\left(1-\sigma(\zeta_{k,i})\right)\right. && (38)\\
& \left.\cdot\,\mathbf{x}_{k,i}\mathbf{x}_{k,i}^\top\right]\left(\mathbf{w}_{g,k}^{t+1}-\mathbf{w}_k^t\right).
\end{aligned}$$

From these Eq. (38) and Eq. (35), we isolate expressions of $\mathbf{w}_k^t - \mathbf{w}_g^t$ and $\mathbf{w}_{g,k}^{t+1} - \mathbf{w}_k^t$:

$$\begin{aligned}
\mathbf{w}_k^t - \mathbf{w}_g^t =& \frac{1}{\alpha\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}\left[\sigma(\xi_{k,i})\left(1-\sigma(\xi_{k,i})\right)\right]}\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}\left[\mathbf{x}_{k,i}\mathbf{x}_{k,i}^\top\right]^{-1} && (39)\\
& \times\left(e_k^t - \nabla\mathcal{L}(\mathbf{w}_k^t;\mathcal{D}_k)\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{w}_{g,k}^{t+1} - \mathbf{w}_k^t =& \frac{1}{\beta\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}\left[\sigma(\zeta_{k,i})\left(1-\sigma(\zeta_{k,i})\right)\right]}\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}\left[\mathbf{x}_{k,i}\mathbf{x}_{k,i}^\top\right]^{-1} && (40)\\
& \times\left(e_{g,k}^{t+1} - \nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1};\mathcal{I}_k^t)\right).
\end{aligned}$$

Using the inexactness measure, we set $\gamma := \max\{\gamma_1, \gamma_2\}$. Then we can bound Eqs. (39) and (35) into the following form:

$$\begin{aligned}
||e_k^t|| =& ||\nabla\overrightarrow{\mathcal{L}}(\mathbf{w}_k^t, \mathbf{w}_g^t;\, \mathcal{D}_k)|| && (41)\\
\overset{(a)}{\leq}& \gamma_1||\nabla\overrightarrow{\mathcal{L}}(\mathbf{w}_g^t, \mathbf{w}_g^t;\, \mathcal{D}_k)|| && (42)\\
=& \gamma_1||\nabla\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_k)|| && (43)\\
\leq& \gamma||\nabla\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_k)||, && (44)
\end{aligned}$$

where (a) is based on **Definition 3.1**.
Next, we consider$||\mathbf{w}_k^t - \mathbf{w}_g^t||$, we bound Eq. (39) in another form:

$$\hat{\mathbf{w}}_k^t \;=\; \arg\min_{\mathbf{w}}\; \overrightarrow{\mathcal{L}}\big(\mathbf{w},\, \mathbf{w}_g^t;\mathcal{D}_k\big). \tag{45}$$

$$\begin{aligned}
||\mathbf{w}_k^t - \mathbf{w}_g^t|| \leq& ||\hat{\mathbf{w}}_k^t - \mathbf{w}_g^t|| + ||\hat{\mathbf{w}}_k^t - \mathbf{w}_k^t|| && (46)\\
\overset{(b)}{\leq}& \frac{1}{\mu}(||\nabla\overrightarrow{\mathcal{L}}(\mathbf{w}_g^t, \mathbf{w}_g^t;\, \mathcal{D}_k)|| + ||\nabla\overrightarrow{\mathcal{L}}(\mathbf{w}_k^t, \mathbf{w}_g^t;\, \mathcal{D}_k)||) && (47)\\
\leq& \frac{1+\gamma}{\mu}||\nabla\overrightarrow{\mathcal{L}}(\mathbf{w}_g^t, \mathbf{w}_g^t;\, \mathcal{D}_k)|| && (48)\\
=& \frac{1+\gamma}{\mu}||\nabla\mathcal{L}(\mathbf{w}_g^t;\, \mathcal{D}_k)||, && (49)
\end{aligned}$$

where (b) is based on **Assumption 2**.
Thus, we can bound Eq. (38) into the following form:

$$||e_{g,k}^{t+1}|| = ||\nabla\overleftarrow{\mathcal{L}}(\mathbf{w}_{g,k}^{t+1}, \mathbf{w}_k^t;\ \mathcal{I}_k^t)|| \tag{50}$$

$$\overset{(c)}{\leq} \gamma_2||\nabla\overleftarrow{\mathcal{L}}(\mathbf{w}_k^t, \mathbf{w}_k^t;\ \mathcal{I}_k^t)|| \tag{51}$$

$$= \gamma_2||\nabla\mathcal{L}(\mathbf{w}_k^t; I_k^t)|| \tag{52}$$

$$\leq \gamma||\nabla\mathcal{L}(\mathbf{w}_k^t; I_k^t)|| \tag{53}$$

$$\leq \gamma(||\nabla\mathcal{L}(\mathbf{w}_k^t; D_k)|| + ||\nabla\mathcal{L}(\mathbf{w}_k^t; D_k) - \nabla\mathcal{L}(\mathbf{w}_k^t; I_k^t)||) \tag{54}$$

$$\overset{(d)}{\leq} \gamma(1+\theta)||\nabla\mathcal{L}(\mathbf{w}_k^t; D_k)|| \tag{55}$$

$$\leq \gamma(1+\theta)(||\nabla\mathcal{L}(\mathbf{w}_g^t; D_k)|| + ||\nabla\mathcal{L}(\mathbf{w}_g^t; D_k) \tag{56}$$

$$- \nabla\mathcal{L}(\mathbf{w}_k^t; D_k)||) \tag{57}$$

$$\overset{(e)}{\leq} \gamma(1+\theta)(||\nabla\mathcal{L}(\mathbf{w}_g^t; D_k)|| + L||\mathbf{w}_g^t - \mathbf{w}_k^t||) \tag{58}$$

$$\overset{(f)}{\leq} (L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu}||\nabla\mathcal{L}(\mathbf{w}_g^t; D_k)||, \tag{59}$$

where (c) is based on **Definition 3.1**, (d) is based on **Assumption 4**, (e) is based on **Assumption 1**, and (f) is based on Eq. (49).

We denote $\Lambda_1 = \mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}[\mathbf{x}_{k,i}\,\mathbf{x}_{k,i}]$, $c_1 = \mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}\big[\sigma(\xi_{k,i})\,(1-\sigma(\xi_{k,i}))\big]$, $\Lambda_2 = \mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}[\mathbf{x}_{k,i}\,\mathbf{x}_{k,i}]$, $c_2 = \mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}\big[\sigma(\zeta_{k,i})\,(1-\sigma(\zeta_{k,i}))\big]$, $\mathbb{E}_k[\Lambda_1] = \Omega_1$, $\mathbb{E}_k[\Lambda_2] = \Omega_2$, $\mathbb{E}_k[c_1] = d_1$, $\mathbb{E}_k[c_2] = d_2$,

$$\mathbb{E}_k[\mathbf{w}_k^t] - \mathbb{E}_k[\mathbf{w}_g^t] = \frac{\mathbb{E}_k\left[\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}[\mathbf{x}_{k,i}\mathbf{x}_{k,i}^\top]^{-1}\right]}{\alpha\mathbb{E}_k\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}[\sigma(\xi_{k,i})\,(1-\sigma(\xi_{k,i}))]} \times \mathbb{E}_k(e_k^t - \mathbb{E}_k\nabla\mathcal{L}(\mathbf{w}_k^t; \mathcal{D}_k)) \tag{60}$$

$$= \frac{\Omega_1^{-1}}{\alpha d_1} \times \mathbb{E}_k(e_k^t - \mathbb{E}_k\nabla\mathcal{L}(\mathbf{w}_k^t; \mathcal{D}_k)), \tag{61}$$

and

$$\mathbb{E}_k[\mathbf{w}_{g,k}^{t+1}] - \mathbb{E}_k[\mathbf{w}_k^t] = \frac{\mathbb{E}_k\left[\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}[\mathbf{x}_{k,i}\mathbf{x}_{k,i}^\top]^{-1}\right]}{\beta\mathbb{E}_k\left[\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}[\sigma(\zeta_{k,i})\,(1-\sigma(\zeta_{k,i}))]\right]} \times (\mathbb{E}_k[e_{g,k}^{t+1}] - \mathbb{E}_k[\nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{I}_k^t)]) \tag{62}$$

$$= \frac{\Omega_2^{-1}}{\beta d_2} \times (\mathbb{E}_k[e_{g,k}^{t+1}] - \mathbb{E}_k[\nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{I}_k^t)]). \tag{63}$$

According to Eq. (61) and Eq. (63), we have

$$\mathbf{w}_g^{t+1} - \mathbf{w}_g^t = \mathbb{E}_k[\mathbf{w}_{g,k}^{t+1}] - \mathbb{E}_k[\mathbf{w}_k^t] + \mathbb{E}_k[\mathbf{w}_k^t] - \mathbb{E}_k[\mathbf{w}_g^t] \tag{64}$$

$$= \frac{\Omega_2^{-1}}{\beta d_2} \times (\mathbb{E}_k[e_{g,k}^{t+1}] - \mathbb{E}_k[\nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{I}_k^t)]) \tag{65}$$

$$+ \frac{\Omega_1^{-1}}{\alpha d_1} \times (\mathbb{E}_k[e_k^t] - \mathbb{E}_k[\nabla\mathcal{L}(\mathbf{w}_k^t; \mathcal{D}_k)]) \tag{66}$$

$$= -\frac{\Omega_2^{-1}}{\beta d_2} \times (\mathbb{E}_k[\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k) + M]) \tag{67}$$

$$- \frac{\Omega_1^{-1}}{\alpha d_1} \times (\mathbb{E}_k[\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k) + N]), \tag{68}$$

where we set

$$M = \nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{I}_k^t) - \nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k) - e_{g,k}^{t+1} \tag{69}$$

$$N = \nabla\mathcal{L}(\mathbf{w}_k^t; \mathcal{D}_k) - \nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k) - e_k^t \tag{70}$$

Firstly, we bound the following term:

$$||\nabla\overleftarrow{\mathcal{L}}(\mathbf{w}_g^t, \mathbf{w}_k^t;\ \mathcal{I}_k^t)|| = ||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{I}_k^t) + \beta c_2 \cdot \Lambda_2\left(\mathbf{w}_g^t - \mathbf{w}_k^t\right)|| \tag{71}$$

$$\overset{(a)}{\leq} ||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{I}_k^t)|| + \frac{\beta||\Lambda_2||}{4}||\mathbf{w}_g^t - \mathbf{w}_k^t|| \tag{72}$$

$$\overset{(b)}{\leq} (1+\theta)||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)|| \tag{73}$$

$$+ \frac{(1+\gamma)\beta||\Lambda_2||}{4\mu}||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)|| \tag{74}$$

$$\leq ((1+\theta) + \frac{(1+\gamma)\beta||\Lambda_2||}{4\mu})||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)||, \tag{75}$$

where (a) is due to $0 < c_2 \leq \frac{1}{4}$, and (b) is due to **Assumption 4** and Eq. (49).

We consider $||\mathbf{w}_{g,k}^{t+1} - \mathbf{w}_g^t||$:

$$\hat{\mathbf{w}}_{g,k}^{t+1} = \arg\min_{\mathbf{w}} \overleftarrow{\mathcal{L}}(\mathbf{w},\ \mathbf{w}_k^t; \mathcal{I}_k^t). \tag{76}$$

$$||\mathbf{w}_{g,k}^{t+1} - \mathbf{w}_g^t|| \leq ||\hat{\mathbf{w}}_{g,k}^{t+1} - \mathbf{w}_g^t|| + ||\hat{\mathbf{w}}_{g,k}^{t+1} - \mathbf{w}_{g,k}^{t+1}|| \tag{77}$$

$$\overset{(a)}{\leq} \frac{1}{\mu}(||\nabla\overleftarrow{\mathcal{L}}(\mathbf{w}_g^t, \mathbf{w}_k^t;\ \mathcal{I}_k^t)|| + ||\nabla\overleftarrow{\mathcal{L}}(\mathbf{w}_{g,k}^{t+1}, \mathbf{w}_k^t;\ \mathcal{I}_k^t)||) \tag{78}$$

$$= \frac{1}{\mu}(||e_{g,k}^{t+1}|| + ||\nabla\overleftarrow{\mathcal{L}}(\mathbf{w}_g^t, \mathbf{w}_k^t;\ \mathcal{I}_k^t)||) \tag{79}$$

$$\overset{(b)}{\leq} ((L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu} + (1+\theta) + \frac{(1+\gamma)\beta||\Omega_2||}{4\mu}) \tag{80}$$

$$\cdot\ ||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)||, \tag{81}$$

where (a) is due to **Assumption 2**, and (b) is based on Eq. (38) and Eq. (74).

We can bound $||M||$ and $||N||$ in the following form:

$$||M|| \leq ||\nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{D}_k) - \nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)|| \tag{82}$$

$$+ ||\nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{I}_k^t) - \nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{D}_k)|| + ||e_{g,k}^{t+1}|| \tag{83}$$

$$\leq ||\nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{D}_k) - \nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)|| \tag{84}$$

$$+ \theta||\nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{D}_k)|| + ||e_{g,k}^{t+1}|| \tag{85}$$

$$\leq (1+\theta)||\nabla\mathcal{L}(\mathbf{w}_{g,k}^{t+1}; \mathcal{D}_k) - \nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)|| \tag{86}$$

$$+ \theta||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)|| + ||e_{g,k}^{t+1}|| \tag{87}$$

$$\leq (1+\theta)L||\mathbf{w}_{g,k}^{t+1} - \mathbf{w}_g^t|| + \theta||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)|| + ||e_{g,k}^{t+1}|| \tag{88}$$

$$\overset{(a)}{\leq} r_1'||\nabla\mathcal{L}(\mathbf{w}_g^t; \mathcal{D}_k)|| \tag{89}$$

where (a) is based on Eq. (79) and Eq. (38), and $r_1' = ((L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu} + (1+\theta) + \frac{(1+\gamma)\beta||\Lambda_2||}{4\mu})(1+\theta)L + \theta + (L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu})$ and $r_1 = ((L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu} + (1+\theta) + \frac{(1+\gamma)\beta||\Omega_2||}{4\mu})(1+\theta)L +$

$\theta + (L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu}$).

$$||N|| \leq ||\nabla\mathcal{L}(\mathbf{w}_k^t;\mathcal{D}_k) - \nabla\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_k)|| + ||e_k^t|| \tag{89}$$

$$\leq L||\mathbf{w}_k^t - \mathbf{w}_g^t|| + ||e_k^t|| \overset{(a)}{\leq} r_2||\nabla\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_k)|| \tag{90}$$

where (a) is based on Eq. (39) and Eq. (35), and $r_2 = \frac{L(1+\gamma)}{\mu} + \gamma$.

We set $< \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g), \mathbf{w}_g^{t+1} - \mathbf{w}_g^t >= U$, and according to Eq. (87), Eq. (90) and **Assumption 3**, we have

$$U = (-\frac{\Omega_2^{-1}}{\beta d_2} - \frac{\Omega_1^{-1}}{\alpha d_1}) < \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g), \mathbb{E}_k[\nabla\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_k)] > \tag{91}$$

$$+ (-\frac{\Omega_2^{-1}}{\beta d_2} - \frac{\Omega_1^{-1}}{\alpha d_1}) < \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g), \mathbb{E}_k[M] + \mathbb{E}_k[N] > \tag{92}$$

$$\leq ||\frac{\Omega_2^{-1}}{\beta d_2} + \frac{\Omega_1^{-1}}{\alpha d_1}||(-B||\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g)||^2 \tag{93}$$

$$+ ||\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g)||(\mathbb{E}_k[||M||] + \mathbb{E}_k[||N||])) \tag{94}$$

$$\leq (||\frac{\Omega_2^{-1}}{\beta d_2}|| + ||\frac{\Omega_1^{-1}}{\beta d_1}||)(r_1 + r_2 - 1)B \cdot \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g) \tag{95}$$

$$\leq 4(||\frac{\Omega_2^{-1}}{\beta}|| + ||\frac{\Omega_1^{-1}}{\alpha}||)(r_1 + r_2 - 1)B \cdot \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g) \tag{96}$$

$$\leq (\frac{4}{\beta||\Omega_2||} + \frac{4}{\alpha||\Omega_1||})(r_1 + r_2 - 1)B \cdot \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g) \tag{97}$$

From Eq. (79) and **Assumption 3**, we can get

$$||\mathbf{w}_g^{t+1} - \mathbf{w}_g^t|| = \mathbb{E}_k[||\mathbf{w}_{g,k}^{t+1} - \mathbf{w}_g^t||] \tag{98}$$

$$\leq B((L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu} \tag{99}$$

$$+ (1+\theta) + \frac{(1+\gamma)\beta||\Omega_2||}{4\mu}) \cdot ||\nabla\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g)||$$

According to Eqs. (99), (97), **Assumption 2**, and **Assumption 1**, we can get

$$\mathcal{L}(\mathbf{w}_g^{t+1};\mathcal{D}_g) \leq \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g) + < \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g), \mathbf{w}_g^{t+1} - \mathbf{w}_g^t > \tag{100}$$

$$+ \frac{L}{2}||\mathbf{w}_g^{t+1} - \mathbf{w}_g^t||^2 \tag{101}$$

$$\leq \mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g) - r||\nabla\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g)||^2 \tag{102}$$

$$\mathcal{L}(\mathbf{w}_g^{t+1};\mathcal{D}_g) - \mathcal{L}(\mathbf{w}^*;\mathcal{D}_g) \leq (1-2\mu r)[\mathcal{L}(\mathbf{w}_g^t;\mathcal{D}_g) - \mathcal{L}(\mathbf{w}^*;\mathcal{D}_g)].$$

where we set $r = (\frac{4}{\beta||\Omega_2||} + \frac{4}{\alpha||\Omega_1||})B - \frac{LB^2}{2}((L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu} + (1+\theta) + \frac{(1+\gamma)\beta||\Omega_2||}{4\mu})^2 - (\frac{4}{\beta||\Omega_2||} + \frac{4}{\alpha||\Omega_1||})(r_1 + r_2)B$, $r_1 = ((L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu} + (1+\theta) + \frac{(1+\gamma)\beta||\Omega_2||}{4\mu})(1+\theta)L + \theta + (L(1+\gamma)+\mu)\frac{\gamma(1+\theta)}{\mu})$, $r_2 = \frac{L(1+\gamma)}{\mu} + \gamma$, $\Omega_1 = \mathbb{E}_k[\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{D}_k}[\mathbf{x}_{k,i}\,\mathbf{x}_{k,i}]]$ and $\Omega_2 = \mathbb{E}_k[\mathbb{E}_{\mathbf{x}_{k,i}\in\mathcal{I}_k^t}[\mathbf{x}_{k,i}\,\mathbf{x}_{k,i}]]$.

D TRADITIONAL NON-IID SETTING EXPERIMENTS

D.1 Non-IID Settings

To ensure fair comparisons and consistent experimental setups with existing collaborative fairness approaches, we adopt the same environment and configurations as FedAVE [19] and FedSAC [20]. We adopt the following traditional non-IID setting. **CLA** (Imbalanced Class Distributions) used in [13, 23] and **DIR** (Imbalanced Sizes + Class Distributions) used in [1, 28] are commonly used non-IID settings in federated learning. Under these scenarios, the label types across clients become severely imbalanced, resulting in skewed partitions that pose significant challenges for model training and fairness. **POW** (Imbalanced Dataset Sizes) was also used in this experiment.

D.2 Federated Data Simulation

In the simulation experiments, we use two image classification datasets: **Fashion MNIST** [22] and **CIFAR10** [7]. **Fashion MNIST** contains 70,000 grayscale images (28×28) evenly split into 10 classes (e.g., T-shirt/top, trousers). **CIFAR10** consists of 60,000 color images (32×32) across 10 classes (e.g., airplane, bird). We partition the datasets into training, validation, and testing in a ratio of 7:1:2. In addition, we set the number of clients as $K = 10$. To simulate the **POW** partition, we follow a power law with an exponent of 1 to divide the global data into 10 clients. For the k-th client, its data size is $|\mathcal{D}_k| = \frac{1}{kZ}|\mathcal{D}_g|$, where $Z = \sum_{k=1}^{10}\frac{1}{k}$. For the **CLA** partition, we fix the local data size and assign labels in an imbalanced manner: the first client receives data from 1 class, the second client from 2 classes, the third client from 3 classes, and so on until the tenth client, which receives data from 10 classes. For the **DIR** partition, we construct three different splits using a Dirichlet distribution with concentration parameters $\alpha = 1$, $\alpha = 2$, and $\alpha = 3$. Concretely, suppose there are C classes in the global dataset $\mathcal{D}_g$. For each class c, let $|\mathcal{D}_c|$ denote the total number of samples in class c, and sample a probability vector $\mathbf{p}_c = (p_{c,1}, p_{c,2}, \ldots, p_{c,K}) \sim \text{Dirichlet}(\alpha, \alpha, \ldots, \alpha)$, where $K = 10$ is the number of clients. We then allocate $\lfloor p_{c,k}\,|\mathcal{D}_c| \rfloor$ samples of class c to client k. Hence, the local dataset for client k is $\mathcal{D}_k = \bigcup_{c=1}^{C}\mathcal{D}_{c,k}$, where $|\mathcal{D}_{c,k}| = \left\lfloor p_{c,k} \cdot |\mathcal{D}_c| \right\rfloor$.

D.3 Baselines

We compare our method against two categories of baselines: *Collaborative Fairness* algorithms designed for non-IID data, and two standard references without fairness considerations. Specifically, we include **CGSV** [23], **CFFL** [13], **FedAVE** [19], and **FedSAC** [20], which explicitly address fairness by measuring client contributions or customizing reward allocations. Additionally, we compare with **Standalone** (each client trains independently without aggregation) and the classic **FedAvg** [15] for federated averaging. Unlike personalized federated learning approaches that tailor models to each client's local distribution, our setting uses a global test set and aims to evaluate overall model performance under traditional non-IID data partitions. Hence, we do not include personalized FL baselines here, since they focus on fitting each client's individual feature space. Instead, we follow the conventional setup in collaborative fairness (CF) research and compare only with CF-oriented baselines under these non-IID settings.

D.4 Implementation

Following the literature [19, 20], we adopt a 2-layer Convolutional Neural Network (CNN) for the *Fashion MNIST* dataset, with mini-batch size $B = 32$ and learning rate $\text{lr} = 0.15$. Then, for *CIFAR10*, we

employ a 3-layer CNN [19, 20] with mini-batch size $B = 128$ and learning rate lr = 0.015. We implement all baselines and our model in PyTorch and train them on an NVIDIA RTX A6000 GPU.

D.5 Evaluation Metrics

Since CLA and DIR settings only allow data with partial labels on each client, they cannot use the local evaluation as we did in the main experiments for the imbalanced covariate shift setting. Thus, following existing work [19, 20], we conduct global evaluations, i.e., all clients are evaluated on a single global test set. We still report the average values of *Maximum Client Accuracy* and *Collaborative Fairness (CF) Coefficient* for **three runs**.

D.6 Results of Experiments

In the traditional federated learning setting (using a single global test set under non-IID data partitions), Table 5 compares both *maximum accuracy* and *collaborative fairness* of our method against various baselines. We observe that our approach consistently outperforms the baselines in terms of both metrics. Specifically, for the POW partition on FashionMNIST, our method achieves **87.93** ± 0.25% in max accuracy, surpassing FedSAC (87.83 ± 0.07%) and other baselines. A similar advantage is seen on CIFAR10, where FedAKD outperforms FedSAC and FedAVE under the same partition. For the DIR partition, our method FedAKD attains higher maximum accuracy on both FashionMNIST and CIFAR10 with different Dirichlet parameters ($\alpha = 1, 2, 3$). When $\alpha = 1.0$, we consistently see around +0.2% to +1.2% improvement over the best baseline in maximum accuracy. For the collaborative fairness metric, our approach yields higher fairness scores, e.g., 98.93 ± 0.14% (POW/FashionMNIST) compared with 96.53 ± 1.20% by FedSAC. This indicates that the performance gap among clients is smaller under our method, demonstrating better collaborative fairness.

E COVARIATE SHIFT DATA GENERATION

In this section, we detail how to construct a covariate-shifted dataset from a baseline distribution with Algorithm 2. We assume the underlying global dataset $\mathcal{D}_g$ approximately follows a single Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, from which we draw subsets for different clients. For each client, we shift the *features* by a fixed Mahalanobis distance.

As discussed in Theorem 2.2, under a large-sample limit, the Kullback-Leibler (KL) divergence can be used to measure how much the client-specific distribution $p_{\theta'}$ deviates from the baseline model p_ω. By fixing $\|\delta_k\|^2_{\Sigma^{-1}} = \delta_k^\top \Sigma^{-1} \delta_k = C$ for each client, we ensure a controlled covariate shift in the mean. The covariance Σ remains the same, and the degree of shift is comparable across all clients. This construction thus provides a systematic way to generate imbalanced covariate shifts, enabling direct measurement and comparison of fairness or performance in federated learning experiments.

Algorithm 2 Covariate Shift Data Generation

Require: Global labeled dataset $\mathcal{D}_g = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_g|}$; constant $C > 0$; number of clients K; desired sample counts $\{n_k\}_{k=1}^K$ (such that $\sum_{k=1}^K n_k = \lfloor |\mathcal{D}_g|/2 \rfloor$); feature dimension d.

Ensure: Covariate-shifted client datasets $\{\mathcal{D}_k\}_{k=1}^K$, where each $\mathcal{D}_k$ retains the original labels but has shifted features (via importance sampling with respect to a Gaussian distribution with mean μ_k and covariance Σ).

Vectorize Features: Convert each feature x_i in $\mathcal{D}_g$ into a vector in $\mathbb{R}^d$. Let $\mathbf{X}$ be the resulting $|\mathcal{D}_g| \times d$ matrix.

Estimate Baseline Gaussian:

$$\mu = \frac{1}{|\mathcal{D}_g|} \sum_{i=1}^{|\mathcal{D}_g|} x_i, \quad \Sigma = \mathrm{Cov}(\mathbf{X}).$$

Half-sampling setup:

Let $M = \lfloor |\mathcal{D}_g|/2 \rfloor$. We only use half of the global dataset for experiments. Hence, set $\sum_{k=1}^K n_k = M$. Adjust $\{n_k\}$ if needed.

for $k = 1, \ldots, K$ **do**

Compute mean shift δ_k:

Sample a vector $\delta_k \in \mathbb{R}^d$ such that $\delta_k^\top \Sigma^{-1} \delta_k = C$.

$\mu_k \leftarrow \mu + \delta_k$ ▷ Shifted mean for client k.

Construct Shifted Dataset for client k:

(1) Compute importance weights

For each (x_i, y_i) in $\mathcal{D}_g$, compute:

$$w_i = \exp\Big(-\tfrac{1}{2}(x_i - \mu_k)^\top \Sigma^{-1}(x_i - \mu_k)\Big).$$

Then normalize:

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^{|\mathcal{D}_g|} w_j}.$$

(2) Sample from $\mathcal{D}_g$ by $\tilde{w}_i$

Sample n_k pairs (x_i, y_i) from $\mathcal{D}_g$ *with replacement* according to probabilities $\tilde{w}_i$.

Store these sampled pairs in $\mathcal{D}_k$.

end for

return $\{\mathcal{D}_k\}_{k=1}^K$.

F BASELINES

In this appendix, we present the details of the baseline methods used in our experiments. We classify them into two main categories and also include two additional traditional baselines: **Standalone** training and the classic **FedAvg**.

Standalone and FedAvg. The first traditional baseline is **Standalone**, in which each client trains independently without any model aggregation. This approach ignores the potential benefits of federated collaboration, providing a lower-bound performance reference. Next, we employ **FedAvg** [15], the standard federated averaging method. FedAvg updates the global model by performing weighted averaging of the locally trained models from all clients, thereby enabling knowledge sharing while minimizing data exchange.

Collaborative Fairness Baselines. We next consider a set of baselines specifically aimed at improving collaborative fairness. **CFFL** [13] focuses on distributing rewards proportionally by measuring each client's contribution differences. In practice, it evaluates client updates on a held-out validation set (or an equivalent performance

Table 5: Comparison of fairness and accuracy for traditional non-IID setting experiments.

(a) Maximum Client Accuracy (%)

Method	FashionMNIST					CIFAR10				
	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
Standalone	84.10±0.22	79.00±0.87	81.92±0.19	83.22±0.36	84.64±0.29	59.06±0.23	32.61±0.22	44.19±0.84	45.59±0.29	47.40±0.89
FedAvg	87.57±0.23	81.04±1.22	85.68±0.77	86.16±1.14	86.80±0.30	59.87±0.95	50.84±0.14	49.82±1.63	49.00±0.76	48.16±0.97
CFFL	83.78±0.56	85.88±0.82	86.90±0.19	87.09±0.08	87.44±0.27	58.41±1.15	42.89±1.22	48.28±1.72	42.15±0.93	44.21±1.27
CGSV	82.68±0.97	82.72±0.99	81.39±1.24	84.92±0.83	86.76±1.46	53.24±0.86	43.98±1.81	47.92±1.33	48.92±1.24	46.21±0.41
FedAVE	85.99±0.82	80.26±0.64	85.29±0.73	85.60±0.44	85.27±0.22	59.11±0.54	45.16±0.91	50.11±1.93	51.24±1.14	42.14±0.33
FedSAC	87.83±0.07	85.28±0.66	87.33±0.49	87.15±0.45	87.81±0.14	58.24±0.11	47.92±0.65	52.14±0.13	50.31±1.33	49.92±0.14
FedAKD	**87.93**±0.25	**86.03**±0.48	**87.57**±0.13	**87.27**±0.38	**87.96**±0.11	**60.03**±0.38	**50.62**±0.72	**53.32**±0.24	**53.12**±0.77	**50.09**±0.15

(b) Collaborative Fairness (CF) Coefficient

Method	FashionMNIST					CIFAR10				
	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)	POW	CLA	DIR(1.0)	DIR(2.0)	DIR(3.0)
FedAvg	16.78±21.51	83.23±6.53	20.18±16.52	23.24±27.33	29.00±31.87	18.84±6.83	86.83±4.82	32.77±6.43	27.92±8.52	50.73±10.39
CFFL	89.43±1.52	91.85±2.39	88.09±3.43	87.29±4.21	76.23±10.32	83.52±4.28	70.62±3.62	66.28±4.15	70.51±8.31	71.35±2.51
CGSV	90.87±0.98	86.63±2.89	92.43±0.64	91.29±1.93	87.53±2.34	90.15±3.37	94.26±0.82	89.27±2.05	91.74±0.45	93.63±1.37
FedAVE	94.21±1.41	89.42±0.44	93.22±1.28	90.34±0.33	94.34±0.98	93.64±0.35	96.25±1.87	95.85±1.86	98.03±0.13	90.62±1.69
FedSAC	96.53±1.20	93.45±1.75	97.31±0.83	95.78±0.89	94.34±2.93	99.05±0.24	95.63±0.51	93.54±0.99	95.73±1.37	94.73±0.97
FedAKD	**98.93**±0.14	**95.13**±1.83	**99.27**±0.23	**97.87**±0.45	**98.82**±0.42	**99.78**±0.08	**97.35**±1.20	**97.05**±0.21	**98.86**±0.72	**97.76**±1.59

benchmark) before deciding how to compensate high- versus low-contribution clients. On the other hand, **CGSV** [23] elies on gradient-based importance metrics—specifically using cosine similarities of gradients—so it does not require a separate validation set. Each client's contribution is gauged by comparing its gradient direction against the aggregated global gradient. Additionally, **FedAVE** [19] incorporates explicit fairness by maintaining a global validation set at the server. It periodically tests each client's model update on this validation set to compute a "reputation" score, which then guides how the final global model is aggregated. **FedSAC** [20] similarly uses a global validation set for evaluating each client's local update. Clients receive a proportionate "reward gradient" based on their evaluated performance, ensuring that clients with higher impact on the validation set obtain a larger share of the global update.

Personalized FL (Covariate Shift) Baselines. Since our approach also falls under the umbrella of Personalized Federated Learning (FL), we include baselines originally designed to tackle feature-level (covariate) shifts, even though they do not explicitly target collaborative fairness. **FedDC** [1] tackles non-IID data by introducing a drift variable that aligns local models more closely with the global model. **FedAS** [25] alleviates intra- and inter-client inconsistencies through federated parameter alignment and client synchronization. **pFedCK** [30] clusters clients by update similarity and performs mutual knowledge distillation between interactive and personalized models to enhance robustness under data heterogeneity. Lastly, **FedMPR** [2] combines iterative magnitude pruning with regularization techniques to improve robustness under highly heterogeneous client data distributions.

By comparing these diverse baselines, we can comprehensively evaluate our proposed method from both collaborative fairness and feature-level drift perspectives.

G IMPLEMENTATION DETAILS

G.1 Common Hyperparameters of All Algorithms Used in Simulation

Number of clients (NUM_CLIENTS) is set to 10. This indicates how many total clients are simulated or trained independently in the scenario. **Global rounds (NUM_GLOBAL_ROUNDS)** is commonly set to 20. This is the total number of federated communication rounds. **Local epochs (LOCAL_EPOCHS)** is often set to 1. It denotes how many epochs each client trains on its local data per global round. **Batch size (BATCH_SIZE)** is often set to 32. **Learning rates (learning_rates)** are set to 0.001 for *FashionMNIST* and 0.005 for *CIFAR10*. These control the local step size for optimizing `SGD`.

G.2 Algorithm-Specific Hyperparameters

Table 6 details each algorithm's unique or particularly important hyperparameters, along with their default values (or ranges) and a brief explanation.

H THE EHR DATASET

We begin with a real-world healthcare dataset composed of 17 tables. To simplify preprocessing and retain the most relevant information, we keep only: *Patient Demographic Table*, *Diagnosis Table*, *Procedure Table*, *Medication Drug Table*, *Lab Result Table*, and *Vital Sign Table*.

Data Processing. We merge these tables by *patient_ID*, ensuring each patient record contains both *static features* (e.g., zipcode, sex) and a series of *events* (medical codes plus numerical attributes, and the patient's age at each event). Because multiple medical coding systems are used across the U.S. healthcare spectrum, we unify these codes into a single standardized terminology.

Next, a board-certified medical oncologist provides a set of *pancreatic cancer* diagnostic codes, which are used to extract data labels.

Table 6: Algorithm-Specific Hyperparameters (Condensed)

Algorithm	Special Hyperparameters	Value
CFFL	1) THETA_U (grad upload) 2) CLIP_NORM (clip thr.) 3) C_TH (rep. thr.) 4) ALPHA (rep. update)	(1) 0.5 (2) 5.0 (3) 0.05 (4) 1.0
CGSV	1) ALPHA_R (mov. avg) 2) BETA (sim. scaling) 3) SPARSITY 4) ALTRUISM	(1) 0.9 (2) 2.0 (3) True (4) 1.0
FedAVE	1) UPLOAD_FRAC 2) DOWNLOAD_FRAC_BASE 3) ALPHA / BETA	(1) 0.5 (2) 0.3 (3) 0.9 / 1.0
FedAvg	—	—
FedDC	1) ALPHA (penalty) 2) drift_vars	(1) 1.0 (2) init=0
FedMPR	PRUNE_PERCENT	0.1
FedProx	MU (prox. coeff.)	0.01
FedSAC	1) BETA (c_i mapping) 2) mid_round	(1) 2.0 (2) 15
SCAFFOLD	1) η_g (global LR) 2) η_l (local LR) 3) K (local steps) 4) c_global, c_local	(1) 0.005 (2) 0.1 (3) 1 (4) init=0
FedAKD	1) Distill α 2) Distill β 3) Temp T	(1) 1.0 (2) 1.0 (3) 1.0

Table 7: State Data Statistics

State	Total	Pos.	Neg.	State	Total	Pos.	Neg.
AK	558	196	362	MT	636	221	415
AL	3,410	1,292	2,118	NC	7,263	2,222	5,041
AR	2,341	842	1,499	ND	605	179	426
AZ	5,521	2,347	3,174	NE	1,429	424	1,005
CA	20,040	7,116	12,924	NH	900	344	556
CO	4,164	1,362	2,802	NJ	7,347	3,762	3,585
CT	2,595	1,180	1,415	NM	1,203	416	787
DE	805	342	463	NV	1,994	792	1,202
FL	18,898	8,158	10,740	NY	18,268	8,134	10,134
GA	7,901	2,461	5,440	OH	11,634	4,339	7,295
HI	1,060	433	627	OK	2,575	865	1,710
IA	2,951	1,093	1,858	OR	3,637	1,322	2,315
ID	1,021	361	660	PA	11,817	4,658	7,159
IL	8,932	3,497	5,435	RI	423	158	265
IN	4,388	1,523	2,865	SC	3,386	1,194	2,192
KS	1,857	524	1,333	SD	651	266	385
KY	4,110	1,396	2,714	TN	5,671	2,038	3,633
LA	3,316	1,150	2,166	TX	15,785	5,152	1,0633
MA	3,492	1,453	2,039	UT	2,117	495	1,622
MD	5,145	1,959	3,186	VA	6,057	1,924	4,133
ME	1,183	473	710	VT	445	150	295
MI	9,744	3,629	6,115	WA	6,247	2,009	4,238
MN	3,504	1,371	2,133	WI	2,893	1,281	1,612
MO	4,341	1,572	2,769	WV	1,712	617	1,095
MS	2,246	710	1,536	WY	358	113	245

In ICD-10, the codes for malignant neoplasm of the pancreas include:

- C25 (Malignant neoplasm of pancreas, general),
- C25.0 (Malignant neoplasm of head of pancreas),
- C25.1 (Malignant neoplasm of body of pancreas),
- C25.2 (Malignant neoplasm of tail of pancreas),
- C25.3 (Malignant neoplasm of pancreatic duct),
- C25.4 (Malignant neoplasm of endocrine pancreas),
- C25.7 (Malignant neoplasm of other specified parts of pancreas),
- C25.8 (Malignant neoplasm of overlapping lesions of pancreas),
- C25.9 (Malignant neoplasm of pancreas, unspecified).

In ICD-9, the corresponding codes are:

- 157 (Malignant neoplasm of pancreas, general),
- 157.0 (Malignant neoplasm of head of pancreas),
- 157.1 (Malignant neoplasm of body of pancreas),
- 157.2 (Malignant neoplasm of tail of pancreas),
- 157.3 (Malignant neoplasm of pancreatic duct),
- 157.4 (Malignant neoplasm of islets of Langerhans),
- 157.8 (Malignant neoplasm of other specified sites of pancreas),
- 157.9 (Malignant neoplasm of pancreas, unspecified).

For each patient, if any of these codes appear in the longitudinal record, we set the label $y = 1$; furthermore, we remove any events that occur at or after the time of pancreatic cancer diagnosis to prevent data leakage. If none of the pancreatic cancer codes appear for a patient, we set that patient's label to $y = 0$.

After these preprocessing steps, we obtain 265,085 de-identified patient samples spanning 50 U.S. states.[4] We randomly sample 10% (26,509) of this dataset as a *global validation set*. The remaining 90% of the data is returned to each state, where each state splits its local data into a *local training set* and a *local test set* (e.g. 7:2). The binary label y indicates whether the patient eventually develops pancreatic cancer. Table 7 summarizes, for each of the 50 states, the number of local samples, how many are used for training/testing, and the distribution of positive/negative labels. Globally, we observe 99,604 positive and 165,481 negative samples from local total data(train and test).

Model and Training Setup. To handle this longitudinal EHR data, we embed the event codes, concatenate them with numerical attributes (e.g. *age, value*), and feed the sequence into a two-layer bidirectional GRU, which captures both forward and backward dependencies. We then apply an attention mechanism over the GRU outputs, computing importance weights for each time step and aggregating them into a context vector. This vector is concatenated with the embedded static features (*sex, postal_code*), and a linear classifier predicts whether the patient will develop pancreatic cancer.

We adopt a focal loss [12] to mitigate label imbalance and train with an Adam optimizer (default PyTorch settings). The mini-batch size is set to $B = 64$, GRU hidden dimension to 256, dropout probability to 0.3, and learning rate to `lr = 0.00001`. We embed *sex* and *postal_code* into vectors of dimensions 8 and 16, respectively, zero-padding time-series inputs (with sequence lengths tracked via `pack_padded_sequence`). Specifically, the two-layer GRU produces a 2×256-dim vector per time step, mapped by the attention layer to a 256-dim energy vector and finally to scalar scores for

[4]All sensitive identifiers are removed or hashed; events after a pancreatic cancer diagnosis are excluded.

softmax weighting. Ultimately, we train and validate our model in this federated setting, treating each state as one client ($K = 50$).