FedCFA: Alleviating Simpson's Paradox in Model Aggregation with Counterfactual Federated Learning

Zhonghua Jiang[1]*, Jimin Xu[1]*, Shengyu Zhang[1]†, Tao Shen[1], Jiwei Li[1], Kun Kuang[1], Haibin Cai[2], Fei Wu[1]

[1]Zhejiang University
[2]East China Normal University
{jiangzhonghua, xujimin, sy_zhang, tao.shen, jiwei_li, kunkuang, wufei}@zju.edu.cn, hbcai@sei.ecnu.edu.cn

Abstract

Federated learning (FL) is a promising technology for data privacy and distributed optimization, but it suffers from data imbalance and heterogeneity among clients. Existing FL methods try to solve the problems by aligning client with server model or by correcting client model with control variables. These methods excel on IID and general Non-IID data but perform mediocrely in Simpson's Paradox scenarios. Simpson's Paradox refers to the phenomenon that the trend observed on the global dataset disappears or reverses on a subset, which may lead to the fact that global model obtained through aggregation in FL does not accurately reflect the distribution of global data. Thus, we propose FedCFA, a novel FL framework employing counterfactual learning to generate counterfactual samples by replacing local data critical factors with global average data, aligning local data distributions with the global and mitigating Simpson's Paradox effects. In addition, to improve the quality of counterfactual samples, we introduce factor decorrelation (FDC) loss to reduce the correlation among features and thus improve the independence of extracted factors. We conduct extensive experiments on six datasets and verify that our method outperforms other FL methods in terms of efficiency and global model accuracy under limited communication rounds.

Introduction

Federated learning (FL) is a technology to enable collaborative learning among multiple parties without violating data privacy (Konečnỳ et al. 2016). In FL, many clients collaborate to train a shared model under the coordination of a server while keeping data dispersed. This reduces the risk of privacy breaches generated by centralized machine learning.

FedAvg (McMahan et al. 2017), a basic FL algorithm, applies gradient descent to train models on distributed clients. In order to adapt to the special scenarios of FL, many researchers have proposed various improved algorithms (Reguieg et al. 2023; Li et al. 2021b; Rothchild et al. 2020) based on FedAvg to enhance the fairness of resource allocation (Ilhan, Su, and Liu 2023; Hao et al. 2021), communication efficiency (Liao et al. 2023; Zhao et al. 2023), privacy security (Zhang et al. 2021; Li et al. 2022) and defense attack ability (Ovi et al. 2023; Cao and Gong 2022) of FL. Among them, the imbalance and heterogeneity (Luo et al. 2023; Mendieta et al. 2022; Qu et al. 2022) of client data are one of the most prominent problems in FL.

The inherent variability in data across clients presents a formidable challenge to FL, causing locally-trained client models to overlook broader patterns evident in the global data. Such heterogeneity further results in inefficient collaborative training and increased communication rounds. To address this problem, existing FL methods mainly adopt two strategies. The first involves (Fang and Ye 2022; Seo and Elmroth 2023; Li et al. 2023a) leveraging data outside of local data to enhance local model optimization, albeit at the potential cost of compromising data privacy. The second strategy (Cheng et al. 2022; Gao et al. 2022; Li, He, and Song 2021) involves heuristically adjusting local models to

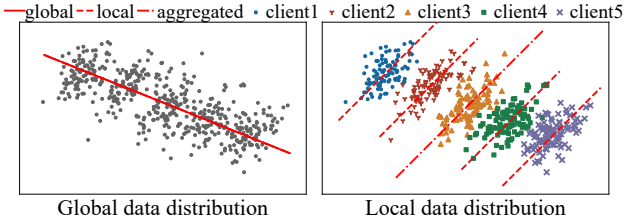

Figure 1: Simpson's Paradox.

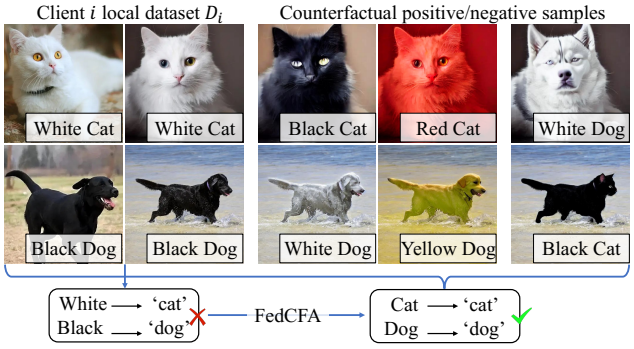

Figure 2: FedCFA can generate counterfactual samples that do not exist locally on client i, preventing the model from learning incorrect feature-label relationships.

*These authors contributed equally.
†The corresponding author.

align more closely with the global model, thereby alleviating data heterogeneity. While most existing works leverage the global model as the alignment reference, our observations, as illustrated in Figure 1, indicate that **global model may fall victim to Simpson's Paradox and potentially become untrustworthy**. That is, the data distribution captured by the aggregated model contradicts the global data distribution, rendering the aforementioned strategies ineffective.

Simpson's Paradox manifests in probability and statistics as a discrepancy where a trend in a dataset reverses or disappears upon analyzing subsets or aggregated data, as depicted in Figure 1. Consider a FL system for classifying cat and dog images, involving two clients with distinct datasets. Client i's dataset primarily includes images of white cats and black dogs, and Client j's dataset comprises images of light gray cats and brown dogs. Individually, the datasets reveal a similar trend: lighter-colored animals are categorized as 'cat', while darker-colored animals are deemed 'dog'. This leads in an aggregated global model that leans towards associating colors with classifications and assigning higher weights to color features. Nevertheless, the global data distribution introduces a number of images of cats and dogs with different colors (such as black cats and white dogs), contradicting the aggregated model. A model trained on global data can easily discover that animal colors are not related to specific classifications, thus reducing weights of color features.

In this paper, we set the goal of alleviating Simpson's Paradox problem, *i.e.*, aligning the aggregated model with global data distribution. We propose FedCFA, a novel FL framework leveraging counterfactual learning. The essence of FedCFA in confronting Simpson's Paradox problem is to identify and counterfactually manipulate critical features that dominate the deviation from global data distribution under the guidance of globally aggregated data, to make local data distribution closer to global data distribution. As shown in Figure 2, the counterfactual samples generated through the counterfactual transformation enable the local model to grasp accurate feature-label relationship and avoid local data distribution contradicting global data distribution, thereby alleviating Simpson's Paradox in model aggregation. Technically, we devise two counterfactual modules that selectively replace critical features, integrating global average data into local data, and constructs positive/negative samples for model learning. Specifically, given local data, we identify dispensable/indispensable features, performing counterfactual transformations to obtain positive/negative samples by replacing those features accordingly. Through contrastive learning on counterfactual samples which are **closer to the global data distribution**, the local model can effectively learn global data distribution. However, counterfactual transformation faces the challenge of extracting independently controllable features from data. A feature may encode multiple types of information, e.g., a pixel of an animal image may carry both color and shape information. To improve the quality of counterfactual samples, we need to ensure that the extracted features cover singular information. Therefore, we introduce factor decorrelation loss, which directly punishes the correlation coefficient between factors to achieve decoupling between feature.

We conduct in-depth experiments to verify the effectiveness of FedCFA on different datasets. To summarize, this paper makes the following key contributions:

- We point out the limitations of existing FL algorithms in dealing with Simpson's Paradox and propose a new FL framework based on counterfactual learning, namely FedCFA, for mitigating this problem.
- Based on FedCFA framework, we design and implement a baseline. This baseline can decouple the data features, obtain independent factors for counterfactual transformation, generates counterfactual samples for local model training, and avoids learning wrong data distribution. Finally, we design a factor decorrelation loss function to measure and constrain the correlation among factors, enhancing feature decoupling effectiveness.
- We conduct extensive experiments based on six datasets under Non-IID and IID data distribution settings. Compared with FedAvg and the advanced FL algorithms, the proposed FedCFA has better convergence and training efficiency, improving the accuracy of the global model.

Related Works

FL is a distributed machine learning paradigm that allows multiple clients to collaboratively train a shared model while protecting data privacy (Konečnỳ et al. 2016). FedAvg (McMahan et al. 2017) is a widely used FL algorithm that achieves model sharing by performing local gradient descent on clients and averaging parameters on server. To adapt to FL's special scenarios, many improved algorithms (Reguieg et al. 2023; Li et al. 2021b; Zhu, Ma, and Blaschko 2023; Rothchild et al. 2020; Li et al. 2021a; Chow et al. 2023; Ilhan, Su, and Liu 2023) have also been proposed based on FedAvg. Among these, training a high-quality global model amidst data heterogeneity (Li et al. 2020a; Zhu et al. 2021; Xie and Song 2023; Hamman and Dutta 2024) remains a significant FL challenge.

In order to solve the problem of data heterogeneity, some methods (Chen and Vikalo 2023; Chai, Liu, and Yang 2022) try to use global or other clients' data to aid local model training. FedBGVS (Mou et al. 2021) mitigates class bias impact through a balanced global validation set. CDFDM (Chai, Liu, and Yang 2022) designs a sharing model that dynamically allocates the amount of shared data according to user data scale, effectively alleviating the problem of data heterogeneity. However, FL methods based on shared data will inevitably leak the privacy information of data owner. Other methods (Li et al. 2020b; Karimireddy et al. 2020; Li et al. 2019) adjust local models or aggregate weights for consistency with the global model. FedProx (Li et al. 2020b) introduces the difference between the global model and the local model of the previous round as a regularization term in the objective function, so that the local update will not deviate too much from the global model. SCAFFOLD (Karimireddy et al. 2020) uses control variables (variance reduction) to correct errors in local updates caused by client drift. q-FedAvg (Li et al. 2019) considers the issue of federation fairness and reduces the variance by adjusting aggregation weight, albeit at the cost of diminished model performance

on some clients. However, when Simpson's Paradox exists in data, neither client model nor global model can accurately reflect the true distribution of global data, causing the above strategy to fail. FedMix (Yoon et al. 2021) augments local data via Mixup with global average data. It fails to disrupt spurious feature-label relationship in local data, and still cannot mitigate Simpson's Paradox. Counterfactual learning (Zhang et al. 2024; Alfeo et al. 2023; Zhang et al. 2020) partly mitigates Simpson's Paradox by generating global-distribution-aligned samples through local data intervention.

Method

Problem Statement

We consider a FL scenario with K clients, each holding a local dataset $\mathcal{D}_k$, for $k = 1, 2, ..., K$. Let $\mathcal{D}_g = \cup_{k=1}^{K} \mathcal{D}_k$ denote the global dataset, and let $|\mathcal{D}|$ denote the size of dataset $\mathcal{D}$. The primary objective in FL is to aggregate model parameters w_k from all participating clients, aiming to derive a global model w_g that fits the global data distribution. This problem is formalized as an optimization task, where $\ell(\cdot)$ denotes the loss function:

$$\begin{aligned} &\min_{w_g}\{f_g(w_g) := \frac{1}{|\mathcal{D}_g|}\sum_{(x,y)\in\mathcal{D}_g}\ell(w_g; x, y)\}, \\ &s.t.\ w_g = \sum_{k=1}^{K}\frac{|\mathcal{D}_k|}{|\mathcal{D}_g|}w_k. \end{aligned} \tag{1}$$

Overall Schema

To address the global model optimization in FL, prevalent algorithms such as FedAvg employ parameter aggregation techniques. Specifically, these algorithms train local models on individual datasets and subsequently aggregate these models' parameters via a weighted average to form a global model. This approach fundamentally assumes that optimizing the function below indirectly tackles the global model optimization issue delineated by Eq. 1, where $\mathbf{w} = (w_1, w_2, ..., w_K)$:

$$\min_{\mathbf{w}}\{F(\mathbf{w}) := \frac{1}{|\mathcal{D}_g|}\sum_{k=1}^{K}\sum_{(x,y)\in\mathcal{D}_k}\ell(w_k; x, y)\}. \tag{2}$$

The premise for the above assumption to hold is that local data distribution can approximately represent global data distribution. However, in actual FL scenarios, data heterogeneity can lead to Simpson's Paradox. The distribution of global data is different or even opposite to that of a single client. Therefore, a local model that is only optimized on local data cannot accurately reflect global data distribution. Likewise, the global model obtained by aggregating these local models cannot accurately reflect global data distribution. To address this, we use global average data to perform counterfactual transformation on local data to generate counterfactual samples, so that the local data distribution is closer to the global and used for the optimization of the local model.

The main steps of our proposed FedCFA framework are shown in Algorithm 1, where T is the total communication rounds set for FL, and w_k^t is the model parameters of client k in the t-th round. In FedCFA, we divide a network into three parts: an Encoder for extracting factors, a Decoder for decoding the input factors, and a SoftMax classifier for classifying the features decoded by the decoder. FedCFA first uses Encoder to extract factors from the image, and then replaces the original factors with factors generated by the global average data through Encoder, thereby generating counterfactual positive and negative samples. The model performs contrastive learning on counterfactual samples, so that the local model can better fit global data distribution, that is, $w_k \Rightarrow w_r$, w_r represents the right global model parameters. By narrowing the gap between w_k and w_r, it indirectly brings w_g and w_r closer, that is, $w_g \Rightarrow w_r$. In addition, we propose a factor decorrelation loss to measure the correlation between factors output by Encoder and take it as one of the optimization goals. Through such a design, we can make the factors extracted more independent, thereby improving the quality of counterfactual samples.

Algorithm 1: FedCFA.

Initialize w_g^0 for global server
for $t \leftarrow 0$ **to** $T-1$ **do**
  Server sends w_g^t and $(\bar{X}_g, \bar{Y}_g)$ to the clients
  $S_t \leftarrow K$ clients selected at random
  for $k \in S_t$ **do**
    $w_k^t \leftarrow w_g^t$
    for *batch*$(X, Y) \subseteq (X_k, Y_k)$ **do**
      Calculate $\mathcal{L}_{cls} = \ell(X, Y)$ and $\mathcal{L}_{corr}$
      Generate counterfactual samples
      Calculate $\mathcal{L}_{pos}, \mathcal{L}_{neg}$
      Calculate $\mathcal{L}_{total}$, $w_k^t \leftarrow w_k^t - \beta_t \bigtriangledown \mathcal{L}_{total}$
    end
    Compute $(\bar{X}_k, \bar{Y}_k)$
    Send $w_k^t, (\bar{X}_k, \bar{Y}_k)$ to server
  end
  Update $(\bar{X}_g, \bar{Y}_g)$, $w_g^{t+1} \leftarrow \sum_{k=1}^{K}\frac{|\mathcal{D}_k|}{|\mathcal{D}_g|}w_k^t$
end

Global Average Dataset Construction

According to Central Limit Theorem, for a random subset of size n from the original dataset, the mean $\bar{x}_i$ converges to a normal distribution as n grows large, characterized by a mean μ and variance $\frac{\sigma^2}{n}$:

$$\bar{x}_i \sim N(\mu, \frac{\sigma^2}{n}), \tag{3}$$

where μ and σ^2 are the expectation and variance of the original dataset. When n is small, $\bar{x}_1$can capture the local characteristics and changes of the dataset more finely, especially in preserving the details of the tail of the data distribution and near the outliers. On the contrary, as n increases, the stability of $\bar{x}_1$ is significantly improved, and its variance is significantly reduced, making it more robust and reliable as an estimate of the overall mean μ, and its sensitivity to outliers is greatly reduced. In addition, in distributed computing environments such as FL, in order to effectively control communication costs, choosing a larger n as the sample size is considered an optimization strategy.

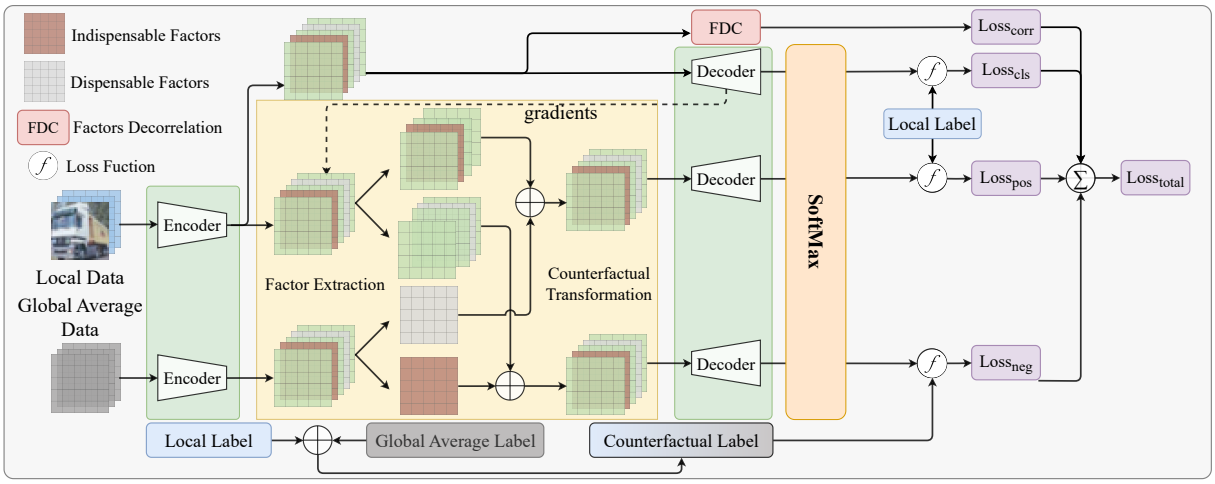

Figure 3: Local model training process in FedCFA.

Based on the above analysis, we construct a global average dataset of size B to approximate global data distribution. Specifically, each client randomly divides local data $\mathcal{D}_k$ into B subsets of $n = \lfloor |\mathcal{D}_k|/B \rfloor$ samples: $\{x_{(i-1)n+1}, x_{(i-1)n+2}, \ldots, x_{in}\}, i = 1, 2, \ldots, B$. For each subset, calculate its mean $\bar{x}_i$:

$$\bar{x}_i = \frac{1}{n} \sum_{j=(i-1)n+1}^{i \cdot n} x_j. \tag{4}$$

As a result, the client can generate a local average dataset $\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_B\}$. The server aggregates these local average data from multiple clients and uses the same method to compute a global average dataset $\bar{X}_g = \{\bar{x}_{g,1}, \bar{x}_{g,2}, \ldots, \bar{x}_{g,B}\}$ of size B, which approximates global data distribution. For label Y, we apply the identical strategy to generate global average labels. Finally, we get the complete global average dataset $\bar{\mathcal{D}}_g = \{\bar{X}_g, \bar{Y}_g\}$.

Counterfactual Transformation Modules

As shown in Figure 2, counterfactual transformation modules replace key features in local data with the corresponding features of global average data, generating counterfactual positive and negative samples, making local data distribution closer to the global, thereby alleviating Simpson's Paradox. The local model training process in FedCFA is shown in Figure 3. Similar to other algorithms, FedCFA first uses original data (X, Y) to calculate classification loss:

$$\mathcal{L}_{cls} = \ell(X, Y). \tag{5}$$

Then, FedCFA uses Encoder to extract data factors $F = Encoder(X)$. Next, we derive the derivative of each factor at the output layer of the model's Decoder to obtain the gradient value of each factor. According to the gradient value, we select the $topk$ factors with low/high gradients, select the corresponding factors from the features obtained by the global average data through the model's Encoder, and replace them to generate counterfactual positive/negative samples. We use $Mask_{pos}/Mask_{neg}$ to set selected low/high gradient factors to zero to retain the factors we need:

$$F_{pos} = Mask_{pos} * F + (1 - Mask_{pos}) * \bar{F}_g, \tag{6}$$

$$F_{neg} = Mask_{neg} * F + (1 - Mask_{neg}) * \bar{F}_g, \tag{7}$$

where $\bar{F}_g$ represents the factors obtained from global average data processed by the Encoder.

For positive samples, the labels do not change. For negative samples, a weighted average is used to generate counterfactual labels:

$$Y_{neg} = \frac{topk}{|F|} * \bar{Y}_g + (1 - \frac{topk}{|F|}) * Y, \tag{8}$$

where $\bar{Y}_g$ represents the label of the global average data.

The classification loss obtained by FedCFA using counterfactual positive and negative samples is as follows:

$$\mathcal{L}_{pos} = \ell(F_{pos}, Y), \quad \mathcal{L}_{neg} = \ell(F_{neg}, Y_{neg}). \tag{9}$$

FDC: Factor Decorrelation

The same pixel may contain multiple data features. For example, in animal images, a pixel may carry both color and appearance information. To enable Encoder to efficiently disentangle various factors, we propose a new loss function. We use Pearson correlation analysis to measure the correlation between factors and use it as a regularization term. Given a batch of data, we use F_i to represent all factors of the i-th sample. $F_{i,j}$ represents the j-th factor of the i-th sample. We regard the factors of the same index j for each sample in the same batch as a set of variables $F_{:,j}$. Finally, we use the average of the absolute values of the Pearson correlation coefficients for each pair of variables as the FDC loss:

$$\mathcal{L}_{corr} = \frac{\sum_{j=1}^{|F_0|} \sum_{j'>j}^{|F_0|} |r(F_{:,j}, F_{:,j'})|}{|F_0| * (|F_0| - 1)/2}, \tag{10}$$

$$r(F_{:,j}, F_{:,j'}) = \frac{Cov(F_{:,j}, F_{:,j'})}{\sqrt{Var(F_{:,j}) Var(F_{:,j'})}}, \tag{11}$$

where $Cov(\cdot)$ is the covariance calculation function, $Var(\cdot)$ is the Variance calculation function.

The final total loss is given:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{neg}\mathcal{L}_{neg} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{corr}\mathcal{L}_{corr}. \tag{12}$$

Proof

We define $F_k, F'_k, F_g, \bar{F}_g$ to represent the factors obtained by the Encoder for the client's local data, local data after counterfactual transformation, global data, and global average data. In this paper, we need to prove that the distribution of local data after counterfactual transformation is closer to global data distribution, that is, to prove that the distribution $P_{F'_k}$ of the local data factor F'_k is closer to the distribution P_{F_g} of the global data factor F_g.

To quantify the difference between two distributions, we use the Wasserstein distance. For any two probability distributions P and Q, the Wasserstein distance is defined as:

$$W(P,Q) = inf_{\gamma \in \Gamma(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|], \tag{13}$$

where $\Gamma(P,Q)$ is the set of all joint probability distributions with P and Q as marginal distributions.

To prove that the counterfactual transformation can reduce the distance between the local factor distribution P_{F_k} and the global factor distribution P_{F_g}, we need to verify the following inequality:

$$W(P_{F_g}, P_{F'_k}) < W(P_{F_g}, P_{F_k}). \tag{14}$$

Since we construct a global average dataset to approximate the global data distribution, we only need to prove the following relationship:

$$W(P_{\bar{F}_g}, P_{F'_k}) < W(P_{\bar{F}_g}, P_{F_k}). \tag{15}$$

The counterfactual transformation generates F'_k by replacing the key factors S_k of the local factor F_k with the corresponding factors in global average data factor $\bar{F}_g$. This means that for any $f_{k,j} \in F_k$, its transformed sample $f'_{k,j}$ has a higher similarity with $\bar{f}_{g,j}$ of the global average data factor $\bar{F}_g$ on the key feature S_k. Therefore, for any $\bar{f}_g \in \bar{F}_g$ and $f'_{k,j} \in F'_k$, we have

$$\left\|\bar{f}_g - f'_{k,j}\right\| < \left\|\bar{f}_g - f_{k,j}\right\|. \tag{16}$$

From this, we can conclude that when $\gamma \in \Gamma(P_{\bar{F}_g}, P_{F'_k})$, it is expected that $\mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]$ is smaller:

$$\begin{aligned} W_1 &= inf_{\gamma\in\Gamma(P_{\bar{F}_g}, P_{F'_k})} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|] \\ &< W_0 = inf_{\gamma\in\Gamma(P_{\bar{F}_g}, P_{F_k})} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]. \end{aligned} \tag{17}$$

In summary, $W(P_{\bar{F}_g}, P_{F'_k})$ is smaller than $W(P_{\bar{F}_g}, P_{F_k})$. This indicates that the counterfactual transformation helps to make the local data closer to the global data distribution.

Experiments

Implementation. Using FedLab (Zeng et al. 2023), we build a typical FL scenario. Unless specified otherwise, we use MLP, ResNet18 and LSTM as network model, with 60 clients, learning rate β of 0.01, one local epoch, batch size of 128, and 500 communication rounds. We conduct experiments on a NVIDIA A100 with 40GB memory.

Datasets: CIFAR10, CIFAR100 (Krizhevsky, Hinton et al. 2009), Tiny-ImageNet, FEMNIST (Caldas et al. 2018), Sent140 (Go, Bhayani, and Huang 2009), MNIST. We built a dataset with Simpson's Paradox based on MNIST.

Different Data Partition Methods. We use two different data partition methods: IID and Non-IID. IID Partition distributes samples uniformly to K clients through random sampling. We use IID_K to represent this data division. For Non-IID, we utilize Dirichlet distribution $Dir_K(\alpha)$ to simulate the imbalance of dataset. The smaller the α, the greater the data difference between clients. We try several different client numbers and data partition methods: $Dir_{60}(0.6)$, $Dir_{60}(0.2)$, $Dir_{100}(0.2)$, $Dir_{100}(0.6)$, IID_{60}, IID_{100}. We use Dirichlet distribution to adjust the frequency of different categories labels in each client to simulate label distribution $P(Y)$ heterogeneity among clients. For FEMNIST, we divide different users into different clients to simulate feature distribution $P(X)$ heterogeneity due to handwriting style variance. For binary classification text dataset Sent140, we divide it into different clients based on users and ensure consistent label distribution among clients, to simulate the heterogeneity of conditional feature distribution $P(X|Y)$.

Baseline Methods: FedAvg (McMahan et al. 2017), FedProx (Li et al. 2020b), SCAFFOLD (Karimireddy et al. 2020), FedPVR (Li et al. 2023b), q-FedAvg (Li et al. 2019) and FedMix (Yoon et al. 2021).

Main Comparison

We adopt two metrics, the global model accuracy after 500 rounds and the number of communication rounds required to achieve target accuracy, to evaluate the performance of FedCFA. Compared with six baseline methods on six datasets, FedCFA shows obvious advantages. We set the target accuracy on CIFAR10, and compare rounds needed for different FL methods. Results in Table 3 show most other methods fail to achieve FedCFA's target accuracy within 1000 rounds, illustrating superior performance and efficiency of FedCFA in FL. In addition, we analyze the effect of FedCFA on improving model accuracy under different experimental settings. The experimental results are shown in Table 1 and 2.

Different Dataset. We verify the effectiveness of FedCFA on six datasets. FEMNIST and Sent140 are relatively simple, and various algorithms can achieve high accuracy. FedCFA has only a slight advantage over other methods. On CIFAR10, CIFAR100 and Tiny-ImageNet, FedCFA shows obvious superiority. These three datasets feature more complex images and greater inter-client data distribution disparities. In this case, FedCFA can effectively mitigate data heterogeneity in FL and improve global model accuracy. For example, under the $Dir_{100}(0.6)$ partition of CIFAR100, FedCFA outperforms the top baseline by 7.75%.

Different Data Heterogeneity. In view of the heterogeneity and imbalance of client data, we model the heterogeneity of label distribution $P(Y)$, the heterogeneity of feature distribution $P(X)$, and the heterogeneity of conditional feature distribution $P(X|Y)$ among clients. We also compare the performance of each algorithm under IID partition. Through counterfactual transformation, FedCFA can narrow the gap between local and global data. Consequently, as Tables 1 and

	Method	$Dir_{60}(0.2)$	$Dir_{60}(0.6)$	IID_{60}	$Dir_{100}(0.2)$	$Dir_{100}(0.6)$	IID_{100}
CIFAR100	FedAvg	40.70±0.24	42.85±0.26	44.37±0.40	38.17±0.32	40.19±0.26	42.19±0.52
	FedProx	40.39±0.23	42.51±0.34	44.21±0.64	38.27±0.46	39.90±0.42	42.24±0.98
	SCAFFOLD	29.36±0.39	33.30±0.48	37.96±0.26	23.25±0.54	29.98±0.19	32.77±0.25
	FedPRV	38.35±1.11	42.91±0.49	45.91±0.05	30.65±0.74	36.58±0.14	39.96±0.30
	q-FedAvg	40.34±0.60	42.61±0.63	44.43±0.28	38.15±0.48	40.20±0.10	42.04±0.65
	FedMix	42.51±0.28	44.16±0.26	45.65±0.31	39.78±0.07	41.43±0.84	43.63±0.64
	FedCFA	**46.96±1.04**	**49.32±0.20**	**48.31±0.53**	**46.71±0.59**	**49.18±0.75**	**47.86±1.22**
CIFAR10	FedAvg	65.88±0.32	73.95±0.16	75.43±0.51	62.87±0.12	70.99±0.70	72.82±0.35
	FedProx	72.23±0.44	77.68±0.03	76.93±0.28	70.36±0.75	75.47±0.53	73.36±0.47
	SCAFFOLD	33.05±5.57	54.58±3.95	75.96±0.57	34.69±2.16	56.17±1.91	71.84±0.77
	FedPRV	59.42±2.26	71.52±0.21	77.42±0.04	55.43±1.74	67.11±0.76	76.16±0.37
	q-FedAvg	71.71±1.05	77.96±0.19	76.92±0.09	70.04±1.55	75.47±0.52	73.68±0.33
	FedMix	74.61±0.74	78.64±0.53	77.90±0.17	73.91±0.79	77.11±0.31	73.93±0.06
	FedCFA	**75.89±1.00**	**82.43±0.08**	**83.36±0.51**	**75.76±0.15**	**81.73±0.12**	**81.68±0.89**

Table 1: The top-1 accuracy (%) after running 500 communication rounds using different methods on CIFAR100, CIFAR10.

Method	Tiny-ImageNet		FEMNIST	Sent140
	$Dir_{60}(0.2)$	$Dir_{60}(0.6)$	Non-IID	Non-IID
FedAvg	27.39±0.13	30.90±0.29	81.31±0.94	68.10±0.48
FedProx	27.34±0.05	30.78±0.16	81.63±0.08	68.10±0.44
q-FedAvg	26.89±0.07	30.70±0.13	81.90±0.58	68.04±0.71
FedMix	28.01±0.19	32.43±0.13	82.31±0.21	67.97±0.39
FedCFA	**30.70±0.68**	**32.86±0.77**	**83.19±0.54**	**69.26±0.37**

Table 2: The top-1 accuracy (%) after running 500 communication rounds on Tiny-ImageNet, FEMNIST, Sent140.

Method	$Dir_{60}(0.2)$	IID_{60}	$Dir_{100}(0.2)$	IID_{100}
Target	75.5	78.4	73.6	75.6
FedAvg	(>,66.22)	(>,74.99)	(>,62.95)	(>,72.42)
FedProx	(>,74.03)	(>,77.16)	(>,72.37)	(>,73.46)
SCAFFOLD	(>,32.77)	(693,78.46)	(>,38.34)	(800,75.68)
q-FedAvg	(>,72.23)	(>,77.04)	(>,72.03)	(>,73.56)
FedMix	(610,75.69)	(>,77.79)	(>,72.25)	(>,74.00)
FedCFA	**(375,75.58)**	**(453,78.41)**	**(427,73.61)**	**(408,75.67)**

Table 3: Communication rounds required by each algorithm to achieve the target accuracy. (R, Acc) denotes the result, where R is the communication round and Acc is the model accuracy (%). If an algorithm cannot achieve the target accuracy within 1000 rounds, we use ">" to denotes R, and use the highest accuracy within 1000 rounds as Acc.

2 illustrate, FedCFA enhances global model accuracy under both Non-IID and IID partition.

Different Number of Clients. We consider client counts of 60 and 100. As Table 1 shows, all baseline accuracies notably drop with increasing client numbers due to intensified data distribution heterogeneity. However, FedCFA is able to maintain high model accuracy. This reveals FedCFA can effectively adapt to varying client numbers and has strong robustness and resistance to data heterogeneity.

Data with Simpson's Paradox. We built a dataset with Simpson's Paradox based on MNIST, by coloring the numbers 1 and 7 and distributing them to 5 clients according to the color depth. For each client, the color of number 1 is darker than that of number 7. For example, in client 1, number 1 is yellow and number 7 is white. This data division makes the data of each client biased, and the model may learn false color-label relationship. We pre-trained an MLP to 95% accuracy, used it as the initial weight for FL, and conducted experiments using FedCFA, FedAvg, and FedMix. The results in Figure 4 show that FedAvg and FedMix are affected by Simpson's paradox and their accuracy decreases. FedCFA eliminates false feature-label associations through counterfactual transformation, generates counterfactual samples to make local distribution close to global distribution, and improves model accuracy.

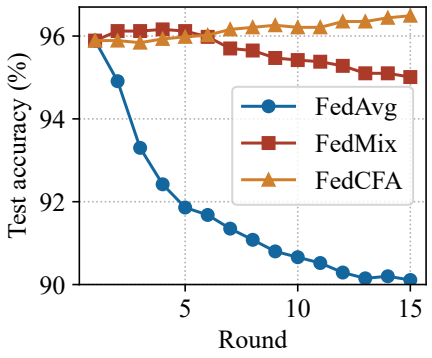

Figure 4: The top-1 accuracy (%) on MNIST data with Simpson's Paradox.

Analysis of FedCFA

This section explores the impact of hyperparameters on FedCFA by comparing FedCFA performance under different hyperparameter settings on CIFAR10.

Factor Decorrelation Module. We set λ_{corr} to 0.1 and compare FedCFA's performance with and without FDC under CIFAR10's four data partition methods. As Figure 5 illustrates, FDC regularization enhances the improvement ef-

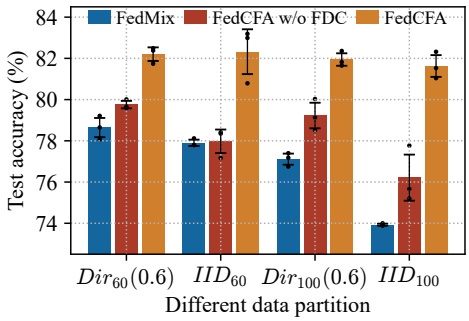

Figure 5: Error Bar Chart: At experimental settings of different random seeds and data partition.

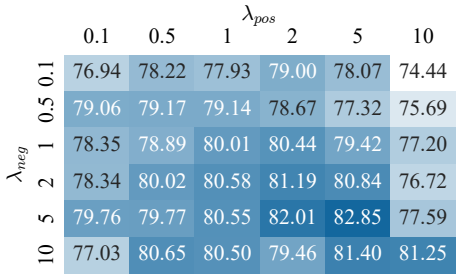

Figure 6: The top-1 accuracy (%) of FedCFA with different proportions of λ_{neg} and λ_{pos}.

fect of the counterfactual module on model accuracy.

Ablation Study of Counterfactual Modules. We conduct ablation study on two counterfactual modules under $Dir_{60}(0.6)$ setting to verify their efficacy. The two counterfactual modules are used to generate counterfactual positive and negative samples respectively. We compare four cases: FedCFA without counterfactual modules, with only positive, only negative, and with both counterfactual modules. Table 4 reveals that counterfactual module can improve FedCFA performance. However, if only one counterfactual module is used, the model cannot fully utilize the contrastive learning between positive and negative samples, and the accuracy improvement is limited. Collaborative operation of two counterfactual modules can markedly boost FedCFA's accuracy, surpassing the additive effect of individual module actions.

Proportion of Counterfactual Modules Losses. To explore the impact of the counterfactual module loss ratio on model performance, this paper conducts experiments under $Dir_{60}(0.6)$ setting. We use a set of six different coefficients $R = \{0.1, 0.5, 1, 2, 5, 10\}$, where $\lambda_{neg}, \lambda_{pos} \in R$, to obtain 36 different loss combinations. Figure 6 indicates steady model performance enhancement as λ_{neg} and λ_{pos} rise simultaneously. However, non-synchronous increases of λ_{neg} and λ_{pos} lead to unbalanced contrastive learning on counterfactual positive and negative samples, causing unstable model performance gains. Moreover, we find excessively high counterfactual module loss ratios impair model's learning on original data distributions, reducing model accuracy.

Application Position of Counterfactuals in Models. As shown in Table 5, Hook represents the layer at which the counterfactual is applied in the ResNet18 model. The smaller the Hook, the closer the counterfactual is to the input layer of the model. To explore the impact of applying counterfactuals at different layers on model performance, we conduct counterfactual experiments on different layers of ResNet18. We found that applying counterfactuals in intermediate layers achieves the best results. This is because at these levels, the model can not only extract key factors, but also avoid losing too many dispensable factors through excessive transformation of the data, making the generated counterfactual positive samples of low quality.

FedCFA	Cls.	Cls.+Pos.	Cls.+Neg.	Cls.+Pos.+Neg.
Accuracy	78.17	78.86	78.87	**80.01**

Table 4: Ablation study of counterfactual modules. Cls. denotes $\mathcal{L}_{cls}$, Pos. denotes $\mathcal{L}_{pos}$, and Neg. denotes $\mathcal{L}_{neg}$.

	$Dir_{60}(0.6)$				$Dir_{100}(0.6)$			
Hook	0	1	2	4	0	1	2	4
	77.65	77.87	**80.01**	75.31	75.89	75.96	**80.03**	73.54
Topk	8	16	24	32	8	16	24	32
	78.53	79.62	**80.01**	78.57	76.51	78.57	**80.03**	79.01

Table 5: The top-1 accuracy (%) of FedCFA under different hyperparameter settings.

Number of Factors for Counterfactual Transformation. In order to explore the effect of counterfactual module, we use different Topk values to control the degree of counterfactual transformation, that is, how many dispensable/indispensable factors are selected for counterfactual transformation. We select four Topk values in $\{8, 16, 24, 32\}$ and compare the models accuracy. As shown in Table 5, as Topk increases, model performance first increases and then decreases. When Topk is 24, the model can achieve optimal performance. This is because when topk is too small, FedCFA cannot select all dispensable/indispensable factors. When topk is too large, the selected factors are not pure enough. Both cases lead to reduce counterfactual samples quality and affect the contrastive learning effect of model.

Conclusion

This paper proposes FedCFA, a novel FL framework leveraging counterfactual learning. FedCFA effectively alleviates Simpson's Paradox impacts by integrating counterfactual samples into local model training, aligning local data distributions with the global. Additionally, FedCFA uses a factor decorrelation loss to decouple and constrain different factors in the data, ensuring the quality of counterfactual samples. We conduct extensive experiments on six datasets and demonstrate that FedCFA outperforms existing FL methods in terms of both efficiency and model accuracy. For future research, we can explore other correlation analysis techniques to improve the framework, especially methods that can capture non-linear correlation between factors to extract more independent factors.

Acknowledgments

This work was supported by the National Science and Technology Major Project (2022ZD0119100), the National Natural Science Foundation of China (No. 62402429, 62441605), the Key Research and Development Program of Zhejiang Province (No. 2024C03270). Additionally, this work was partially supported by ZJU Kunpeng&Ascend Center of Excellence, Ningbo Yongjiang Talent Introduction Programme (2023A-397-G).

References

Alfeo, A. L.; Zippo, A. G.; Catrambone, V.; Cimino, M. G.; Toschi, N.; and Valenza, G. 2023. From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. *Computer Methods and Programs in Biomedicine*, 236: 107550.

Caldas, S.; Duddu, S. M. K.; Wu, P.; Li, T.; Konečnỳ, J.; McMahan, H. B.; Smith, V.; and Talwalkar, A. 2018. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.

Cao, X.; and Gong, N. Z. 2022. Mpaf: Model poisoning attacks to federated learning based on fake clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3396–3404.

Chai, B.; Liu, K.; and Yang, R. 2022. Cross-Domain Federated Data Modeling on Non-IID Data. *Computational Intelligence and Neuroscience*, 2022.

Chen, H.; and Vikalo, H. 2023. Federated learning in non-iid settings aided by differentially private synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5026–5035.

Cheng, A.; Wang, P.; Zhang, X. S.; and Cheng, J. 2022. Differentially private federated learning with local regularization and sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10122–10131.

Chow, K.-H.; Liu, L.; Wei, W.; Ilhan, F.; and Wu, Y. 2023. STDLens: Model Hijacking-resilient Federated Learning for Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16343–16351.

Fang, X.; and Ye, M. 2022. Robust federated learning with noisy and heterogeneous clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10072–10081.

Gao, L.; Fu, H.; Li, L.; Chen, Y.; Xu, M.; and Xu, C.-Z. 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.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12): 2009.

Hamman, F.; and Dutta, S. 2024. Demystifying Local & Global Fairness Trade-offs in Federated Learning Using Partial Information Decomposition. In *The Twelfth International Conference on Learning Representations*.

Hao, W.; El-Khamy, M.; Lee, J.; Zhang, J.; Liang, K. J.; Chen, C.; and Duke, L. C. 2021. Towards fair federated learning with zero-shot data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3310–3319.

Ilhan, F.; Su, G.; and Liu, L. 2023. ScaleFL: Resource-Adaptive Federated Learning With Heterogeneous Clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24532–24541.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, 5132–5143. PMLR.

Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, A.; Sun, J.; Zeng, X.; Zhang, M.; Li, H.; and Chen, Y. 2021a. Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, 42–55.

Li, B.; Esfandiari, Y.; Schmidt, M. N.; Alstrøm, T. S.; and Stich, S. U. 2023a. Synthetic data shuffling accelerates the convergence of federated learning under data heterogeneity. *arXiv preprint arXiv:2306.13263*.

Li, B.; Schmidt, M. N.; Alstrøm, T. S.; and Stich, S. U. 2023b. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3964–3973.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10713–10722.

Li, T.; Hu, S.; Beirami, A.; and Smith, V. 2021b. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, 6357–6368. PMLR.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020a. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3): 50–60.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020b. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2: 429–450.

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2019. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*.

Li, Z.; Zhang, J.; Liu, L.; and Liu, J. 2022. Auditing privacy defenses in federated learning via generative gradient

leakage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10132–10142.

Liao, D.; Gao, X.; Zhao, Y.; and Xu, C.-Z. 2023. Adaptive Channel Sparsity for Federated Learning Under System Heterogeneity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20432–20441.

Luo, K.; Li, X.; Lan, Y.; and Gao, M. 2023. GradMA: A Gradient-Memory-based Accelerated Federated Learning with Alleviated Catastrophic Forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3708–3717.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, 1273–1282. PMLR.

Mendieta, M.; Yang, T.; Wang, P.; Lee, M.; Ding, Z.; and Chen, C. 2022. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8397–8406.

Mou, Y.; Geng, J.; Welten, S.; Rong, C.; Decker, S.; and Beyan, O. 2021. Optimized federated learning on class-biased distributed data sources. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 146–158. Springer.

Ovi, P. R.; Dey, E.; Roy, N.; and Gangopadhyay, A. 2023. Mixed Quantization Enabled Federated Learning to Tackle Gradient Inversion Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5045–5053.

Qu, L.; Zhou, Y.; Liang, P. P.; Xia, Y.; Wang, F.; Adeli, E.; Fei-Fei, L.; and Rubin, D. 2022. Rethinking architecture design for tackling data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10061–10071.

Reguieg, H.; Hanjri, M. E.; Kamili, M. E.; and Kobbane, A. 2023. A Comparative Evaluation of FedAvg and Per-FedAvg Algorithms for Dirichlet Distributed Heterogeneous Data. *arXiv preprint arXiv:2309.01275*.

Rothchild, D.; Panda, A.; Ullah, E.; Ivkin, N.; Stoica, I.; Braverman, V.; Gonzalez, J.; and Arora, R. 2020. Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning*, 8253–8265. PMLR.

Seo, E.; and Elmroth, E. 2023. MadFed: Enhancing Federated Learning with Marginal-data Model Fusion. *IEEE Access*.

Xie, Z.; and Song, S. 2023. Fedkl: Tackling data heterogeneity in federated reinforcement learning by penalizing kl divergence. *IEEE Journal on Selected Areas in Communications*, 41(4): 1227–1242.

Yoon, T.; Shin, S.; Hwang, S. J.; and Yang, E. 2021. Fedmix: Approximation of mixup under mean augmented federated learning. *arXiv preprint arXiv:2107.00233*.

Zeng, D.; Liang, S.; Hu, X.; Wang, H.; and Xu, Z. 2023. FedLab: A Flexible Federated Learning Framework. *Journal of Machine Learning Research*, 24(100): 1–7.

Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; and Gao, Y. 2021. A survey on federated learning. *Knowledge-Based Systems*, 216: 106775.

Zhang, S.; Jiang, T.; Wang, T.; Kuang, K.; Zhao, Z.; Zhu, J.; Yu, J.; Yang, H.; and Wu, F. 2020. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4373–4382.

Zhang, S.; Miao, Q.; Nie, P.; Li, M.; Chen, Z.; Feng, F.; Kuang, K.; and Wu, F. 2024. Transferring Causal Mechanism over Meta-representations for Target-Unknown Cross-domain Recommendation. *ACM Transactions on Information Systems*, 42(4): 1–27.

Zhao, J. C.; Elkordy, A. R.; Sharma, A.; Ezzeldin, Y. H.; Avestimehr, S.; and Bagchi, S. 2023. The Resource Problem of Using Linear Layer Leakage Attack in Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3974–3983.

Zhu, H.; Xu, J.; Liu, S.; and Jin, Y. 2021. Federated learning on non-IID data: A survey. *Neurocomputing*, 465: 371–390.

Zhu, J.; Ma, X.; and Blaschko, M. B. 2023. Confidence-aware personalized federated learning via variational expectation maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24542–24551.