MITIGATING DEMOGRAPHIC BIAS OF FEDERATED LEARNING MODELS VIA WORST-FAIR DOMAIN SMOOTHING

Anonymous authors
Paper under double-blind review

ABSTRACT

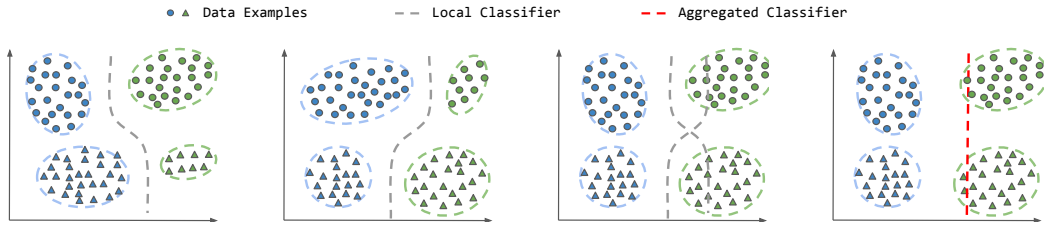
Federated learning (FL) has shown impressive performance in training modern machine learning models from distributed data sources. However, the distributed training process of FL could suffer from a non-trivial bias issue, where the trained models are affected by the imbalanced distribution of the training data on local clients, and eventually lead to a severe bias of the aggregated global model. In this paper, we propose a novel fairness-aware FL training framework **W**orst-**F**air **D**omain **S**moothing (**WFDS**) to address the bias issue of FL models from a domain-shifting perspective. Our framework consists of two concurrent components: 1) local worst-fair training, and 2) reference domain smoothing. The first module is designed to train fair local models and enforces the robustness of local fairness against domain shifts from local distribution to global distribution by performing worst-fair training. The second module simulates a reference data domain of the studied FL application for all clients, and implicitly reduces the domain discrepancy of training data among different clients. With reduced domain discrepancy, the fairness of each local model will be learned from similar training distributions despite on different clients. As such, improved global fairness can be achieved after aggregating the local models into the global model. Evaluation results on multiple real-world datasets show that WFDS can achieve significant performance gains in demographic fairness compared to state-of-the-art baselines.

1 INTRODUCTION

Federated learning (FL) has become one of the promising solutions to train modern machine learning (ML) models without directly accessing the training data on local clients (Wang et al., 2021a; McMahan et al., 2017). That is, at each communication round, each local client receives the global model from a central server, and launches the training of the model using private local data to obtain its local model. At the end of each communication round, the global model will be updated by aggregating the local models using a secure aggregation protocol. In an FL application, the local data privacy is preserved as there is no direct exchange of the data from the clients to the server or between clients McMahan et al. (2017). While FL has been successfully deployed for many privacy-sensitive applications (e.g., recidivism justice, loan approvals, and healthcare (Xu et al., 2021)), concerns about the fairness of such FL models have been raised. For instance, Larson & Kirchner (2016); Bacchini & Lorusso (2019); Buolamwini & Gebru (2018) have reported that ML algorithms deployed for commercial face recognition services or recidivism prediction systems across the US are prone to have a much higher error rate on African-Americans than Caucasians.

Recent efforts (Zhang et al., 2018; Beutel et al., 2017; Zhao et al., 2017; Kou et al., 2021; Han et al., 2021; Hashimoto et al., 2018) have been made to mitigate the bias issue of ML models, but many of them are tailored for the centralized learning setting, and only limited methods have been proposed to address the bias for FL models (Zhang et al., 2020). Moreover, as argued in (Ezzeldin et al., 2021; Cui et al., 2021), simply applying fair methods to local clients and aggregating the fair models from local clients (i.e., **local fairness**) can not ensure the fair performance of the global model (i.e., **global fairness**) and vice versa due to the disparity between the local and global data distributions. To address the bias issue of the FL models, we follow the definition of the demographic bias of an FL

model as the performance discrimination of the global model against a certain population group associated with sensitive demographic attributes (e.g., gender, race) (Ezzeldin et al., 2021)



(a) Local Classifier on Client 1 Distribution (**Fair**)

(b) Local Classifier on Client 2 Distribution (**Fair**)

(c) Local Classifier on Global Distribution (**UnFair**)

(d) Aggregated Classifier on Global Distribution (**Unfair**)

Figure 1: Domain shifts from local to global distributions and heterogeneity of local fairness.

In this work, we identify two core obstacles of addressing the bias issue in FL models: (1) the vulnerability of local fairness under the distribution shift from local to global data distributions, and (2) the discrepancy among local training data distributions, which leads to the heterogeneity of local fairness. For a better understanding of these two obstacles, an illustrative example is given in Figure1. In this example, the local fair model on client 1 (Figure 1a) or client 2 (Figure 1b) becomes unfair when directly applied to the global data distribution (Figure 1c), indicating the local fairness is not transferable due to the domain shift from the local to global data distribution. Moreover, due to the discrepancy between local data distributions on client 1 and client 2, the fairness of local models obtained by debiasing methods is conditioned on local data. As such, locally trained fair models demonstrate heterogeneous behaviors over the global data distributions (as shown in Figure 1c), which may result in an unfair global model upon aggregation (Figure 1d). Therefore, we formulate the objective of mitigating bias in FL models as a task of learning robust local fairness against domain shifts and reducing discrepancy across local data distributions. To this end, we propose a novel fair FL framework called Worst-Fair Domain Smoothing (WFDS), addressing the bias issue from domain-shifting perspective. Within WFDS, we **define each local client as an independent data domain**, where the demographic composition of the local data on this specific client is different from the demographic composition on other clients.

WFDS consists of two modules: local worst-fair training and reference domain smoothing as shown in Figure 2. The first module, local worst-fair training, optimizes a fairness-aware loss over deliberately crafted worst-fair training samples. This module is designed to train fair models on local clients and enforce the robustness of local fairness against distribution shifts from local distributions to the global distribution. To reduce the heterogeneity of local fairness and preserve the local fairness during global model aggregation, we design a second module of our framework: reference domain smoothing. This module simulates a data domain at the central server, which can be accessed by local clients as a domain reference. All local data domains will be smoothed towards the simulated reference domain. As such, domain discrepancy among local clients that leads to the heterogeneity of local fairness is effectively reduced. Consequently, the global fairness is improved when aggregating local models with reduced heterogeneity.

We summarize the main contributions of our works as follows[1]:

1. To the best of our knowledge, WFDS is the first work that addresses the bias issue of FL models from a domain-shifting perspective. That is, each client is treated as an independent data domain, and we aim at obtaining a fair FL model by smoothing all local domains.
2. We propose a novel worst-fair training method in WFDS to train fair local models in FL applications. More importantly, this module enforces the robustness of local fairness against the distribution shift from local distributions to the global distribution.
3. We propose a novel reference domain smoothing module to address the heterogeneity of local fairness. Here, the discrepancy across local domains is implicitly reduced to improve the compatibility of local fairness, and thereby ensure the aggregated global fairness.

[1]We adopt publicly available datasets and will release the implementation of WFDS upon publication.

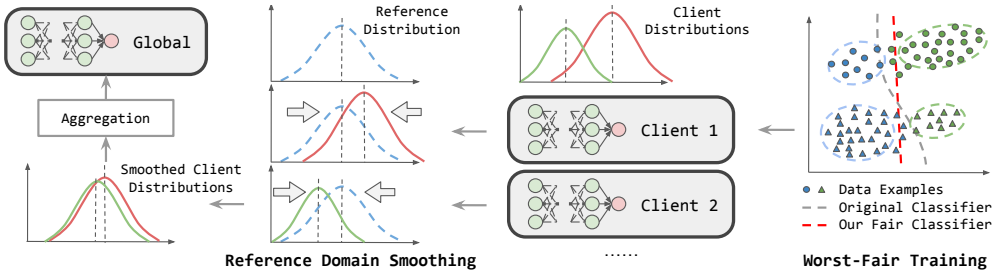


Figure 2: Worst-fair Domain Smoothing (WFDS), which consists of two concurrent modules: local worst-fair training and reference domain smoothing.

4. We demonstrate the effectiveness of the proposed WFDS on multiple real-world datasets, where WFDS consistently outperforms the state-of-the-art baselines by a significant margin.

2 Related Work

Fairness of Machine Learning. A significant amount of algorithms have been proposed to increase the fairness of ML models (Zhang et al., 2018; Beutel et al., 2017; Zhao et al., 2017; Kou et al., 2021; Han et al., 2021; Hashimoto et al., 2018). One stream of the solutions focuses on learning fair data representations by adopting adversarial learning (Zhang et al., 2018; Beutel et al., 2017) or fairness regularization (Zhao et al., 2017). In comparison, Kou et al. (2021); Roh et al. (2020) directly create fair data batches for the training process, addressing the bias issue from the data origin. Another type of methods develop fairness-aware re-weighting strategies to modify the training distribution and adjust the penalization on different training samples to improve model fairness (Han et al., 2021; Hashimoto et al., 2018). These methods have been proved to be effective in addressing demographic bias issue in ML models, but they are tailored for centralized training settings. Moreover, as argue in Ezzeldin et al. (2021), directly applying centralized fairness algorithms in an FL application could not necessarily guarantee the fairness of the aggregated global model.

Federated Learning. Federated Learning allows multiple clients to collaboratively learn a shared machine learning model without exchanging the data of each clients (McMahan et al., 2017). The distributed training process in FL ensures the data privacy, which is extremely important for privacy-sensitive applications such as healthcare or criminal justice. However, FL models might suffer from a non-trivial performance bias against underrepresented demographic groups within the entire population, and only limited solutions have been proposed to address the bias issue in FL (Mohri et al., 2019; Du et al., 2021; Ezzeldin et al., 2021; Zhang et al., 2020; Zeng et al., 2021b; Cui et al., 2021). For instance, in (Mohri et al., 2019), the proposed Agnostic FL (AFL) achieves a weak fairness notion w.r.t. a singe demographic group by optimizing the worst training loss incurred on the protected classes (Zhang et al., 2020). Moreover, Ezzeldin et al. (2021) presents a re-weighting strategy to enable a fair aggregation protocol, eventually leading to a fair global model. Cui et al. (2021) formulate the fair Fl problem as a constrained multi-objective optimization problem to meet the fairness constraints for all clients. Additionally, the optimization-based method in (Cui et al., 2021) requires shared losses and a synchronized update of all local clients, rendering it inapplicable for real-world FL scenarios. In contrast, our WFDS allows local clients to train the local model asynchronously, which is more practical in real-world applications. More importantly, our scheme improves the global fairness from a higher domain-shifting perspective by enforcing the robustness of local fairness against domain shifts and reducing heterogeneity of local fairness.

3 Problem Statement

3.1 Federated Learning

In an FL application, we assume there are a total of K local clients and each client has a local dataset $\boldsymbol{X}_k$ of size N_k. The total number of data points from all clients is $N = \sum_k N_k$. For all clients, each data point is characterized by an input feature $\boldsymbol{x} \sim \mathcal{X}$, a demographic attribute $a \sim \mathcal{A}$ and a

predictive attribute (output) $y \sim \mathcal{Y}$. All clients share the same input space $\mathcal{X}$, the same demographic space $\mathcal{A}$ and the same output space $\mathcal{Y}$. Formally, the local dataset on client k is defined as:

$$\boldsymbol{X}_k = \{(\boldsymbol{x}_1^k, a_1^k, y_1^k), ..., (\boldsymbol{x}_{N_k}^k, a_{N_k}^k, y_{N_k}^k) | (\boldsymbol{x}^k, a^k, y^k) \sim \mathcal{P}_k\}, \tag{1}$$

where $\mathcal{P}_k$ denotes the local data distribution on client k.

The overall goal of federated learning is to collaboratively learn a global model stored in a central server from the local data without requiring the local clients to share data with each other and the central server. To find the optimal global model f parameterized by $\boldsymbol{\theta}^*$, the classic federated learning model aims to minimize the training loss over the samples from all clients:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} l(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k), y_i^k). \tag{2}$$

To minimize Equation 2, McMahan et al. (2017) proposed an algorithm called Federated Averaging (FedAvg). That is, at each round, the local clients firstly receive the same global model $f_{\boldsymbol{\theta}}$ from the central server, perform local training of the model on local data separately, and obtain different local models $(f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2}, ...f_{\boldsymbol{\theta}_K})$. Then, the global model will be updated using a weighted-average of the different local models based on the size of the local datasets. Therefore, in practice, within FedAvg, the optimal global model $\boldsymbol{\theta}^*$ is derived via:

$$\boldsymbol{\theta}^* = \frac{1}{N} \sum_{k=1}^{K} N_k \cdot \boldsymbol{\theta}_k, \quad \text{where} \quad \boldsymbol{\theta}_k = \arg\min_{\boldsymbol{\theta}} \frac{1}{N_k} \sum_{i=1}^{N_k} l(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k), y_i^k), \quad k \in \{1, ..., K\}. \tag{3}$$

Equation 3 is equivalent to equation 2 (McMahan et al., 2017), but after the local training, the local models $f_{\boldsymbol{\theta}_k}$ could be severely biased against a specific demographic group when the local training data is imbalanced (Zhang et al., 2018; Beutel et al., 2017; Zhao et al., 2017; Kou et al., 2021; Han et al., 2021; Hashimoto et al., 2018). Moreover, note that during each local training round, the local model $\boldsymbol{\theta}_k$ on the client k is obtained by modeling local training data distribution $\mathcal{P}_k$:

$$\boldsymbol{\theta}_k = \arg\min_{\boldsymbol{\theta}} \frac{1}{N_k} \sum_{i=1}^{N_k} l(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k), y_i^k) \approx \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}_i^k, y_i^k) \sim \mathcal{P}_k} [l(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^k), y_i^k)]. \tag{4}$$

Since the data on each local client is collected by clients in different regions and the local populations might have drastically different demographic compositions. The potential domain discrepancy among $\mathcal{P}_1$, $\mathcal{P}_2$, ..., $\mathcal{P}_k$ and the global data distribution could lead to a frustrating consequence, where the locally trained fair models could be aggregated into a globally unfair model (e.g., Figure 1d). Additionally, global fairness could not automatically guarantee local fairness on different clients Cui et al. (2021). Therefore, in this work, we propose to reduce the domain discrepancy among local data distributions to reduce the heterogeneity of local fairness for all clients.

3.2 Notion of Demographic Fairness

In this work, we focus on addressing the demographic bias of FL models. To properly measure the bias of the models, we adopt various notions of demographic fairness. Recall a data point $\boldsymbol{x}$ defined in our problem contains a sensitive demographic attribute a (e.g. gender or race) and a predictive attribute y. The first group of fairness notions include the commonly used demographic parity ϕ_D Hardt et al. (2016) and equalized odds Hardt et al. (2016) ϕ_E. These two metrics mainly focus on the model's performance discrepancy on positive predictions.

Definition 3.1 (Demographic Parity Φ_D (binary case)). For a given classifier $f_{\boldsymbol{\theta}}$, demographic parity Φ_D measures the absolute error between the probability of making positive prediction for each demographic group.

$$\Phi_D(f_{\boldsymbol{\theta}}, \boldsymbol{x}, a) = |p(f_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1 | a = 0) - p(f_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1 | a = 1)|. \tag{5}$$

Definition 3.2 (Equalized Odds Φ_E (binary case)). For a given classifier $f_{\boldsymbol{\theta}}$, equalized odds Φ_E measures the absolute error of the true positive rates for each demographic group and the absolute error of the false positive rates for each demographic group.

$$\begin{aligned} \Phi_E(f_{\boldsymbol{\theta}}, \boldsymbol{x}, a, y) = &|p(f_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1 | y = 1, a = 0) - p(f_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1 | y = 1, a = 1)| \\ &+ |p(f_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1 | y = 0, a = 0) - p(f_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1 | y = 0, a = 1)|. \end{aligned} \tag{6}$$

In addition to positive predictions, we also use another group of fairness-aware accuracy metrics to measure the model's demographic fairness as in Yue et al. (2022), namely sub-group accuracy gap and balanced accuracy.

Definition 3.3 (**Sub-group Accuracy** $\mathcal{A}_{sub}$)**.** For a given classifier $f_{\boldsymbol{\theta}}$, the sub-group accuracy $\mathcal{A}_{sub}$ measures the accuracy of the prediction within a specific demographic group characterized by demographic attribute a and predictive attribute y.

$$\mathcal{A}_{sub}(f_{\boldsymbol{\theta}}, a, y) = \frac{\mathbb{E}_{(\boldsymbol{x}_i, y_i)\sim \boldsymbol{X}}\Big[\mathbb{1}(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = y, y_i = y, a_i = a)\Big]}{\mathbb{E}_{(\boldsymbol{x}_i, y_i)\sim \boldsymbol{X}}\Big[\mathbb{1}(y_i = y, a_i = a)\Big]}. \tag{7}$$

Definition 3.4 (**Sub-group Accuracy Gap** Φ_A)**.** For a given classifier $f_{\boldsymbol{\theta}}$, the sub-group accuracy gap Φ_A sums up the absolute error among all demographic groups.

$$\Phi_A(f_{\boldsymbol{\theta}}) = \sum_a \sum_y \sum_{a'} \sum_{y'} |\mathcal{A}_{sub}(a, y) - \mathcal{A}_{sub}(a', y')| \tag{8}$$

Definition 3.5 (**Balanced Accuracy** $\mathcal{A}_B$)**.** For a given classifier $f_{\boldsymbol{\theta}}$, the balanced accuracy computes the averaged sub-group accuracy for all demographic groups.

$$\mathcal{A}_B(f_{\boldsymbol{\theta}}) = \frac{\sum_a \sum_y \mathcal{A}_{sub}(a, y)}{|\mathcal{A}| \cdot |\mathcal{Y}|}. \tag{9}$$

For the fairness notions introduced above, we highlight that all of these metrics are conditioned on the model parameters. Since the model parameter is learned based on the training data distribution, local fairness could be highly heterogeneous across different local clients due to the discrepancy across different local data distributions.

4 Algorithm

Based on the detailed analysis for the classical federated learning framework and various notions of demographic fairness, we note the major challenge of addressing bias issue in FL applications is to overcome the distribution shift from local distribution to the global distribution and the domain discrepancy across all local clients. To this end, we design our fairness-aware FL framework Worst-Fair Domain Smoothing in a way such that robust local fairness is enforced against domain shift, and the domain discrepancy between different local data distributions is implicitly reduced to boost the compatibility of local fairness for all clients.

4.1 Local Worst-Fair Training

The worst-fair training module is designed to achieve local fairness for individual clients and further enhance the robustness of local fairness under distribution shifts. To obtain robust local fairness, this module minimizes a fairness-aware loss over a worst-fair data distribution.

Mathematically, we define this process as a minimax game between a fairness adversary and a fairness booster. The fairness adversary tries to generate fairness-adversarial examples by perturbing the local training samples towards a direction of degraded fairness w.r.t. a selected fairness notion (e.g., demographic parity Φ_D), whereas the fairness booster tries to learn a fair model even if the training data is deliberately perturbed to be biased. The minimax game on client k is formulated as

$$\begin{aligned} &\min_{\boldsymbol{\theta}_k} \frac{1}{C^k} \sum_{j=1}^{C^k} \left[\frac{1}{M_j} \sum_{i=1}^{M_j} \Big[l(f_{\boldsymbol{\theta}_k}(\boldsymbol{x}^k_{j,i} + \boldsymbol{\delta}_{j,i}), y^k_{j,i}) \Big] \right] \\ &s.t. \qquad \boldsymbol{\delta}_{j,i} = \arg\max_{\boldsymbol{\delta}} \hat{\Phi}_*(f_{\boldsymbol{\theta}_k}), \quad \|\boldsymbol{\delta}_{j,i}\| \leq \epsilon, \end{aligned} \tag{10}$$

where C^k denotes the total number of demographic groups within the local data on client k and M_j denotes the number of training samples of the j-th demographic group. $\boldsymbol{\delta}_{j,i}$ is the bias perturbation deliberately crafted by the fairness adversary to bias the training sample $\boldsymbol{x}^k_{j,i}$. $\hat{\Phi}_*$ is the differentiable version of any fairness notion (Φ_D, Φ_E or Φ_A) introduced in Section 3.

The minimization in Equation 10 suggests that the fairness booster obtains local fairness by minimizing a fairness-aware training loss. Herein, the fairness-awareness of the training loss is enabled by re-weighting each training sample inversely proportional to the data frequency (IDF) of the demographic group which this training sample belongs to (Han et al., 2021). That is, the empirical risk is averaged within each demographic group, then the group-level risk is further averaged over the number of demographic groups. Moreover, note that the fairness-aware loss is evaluated over the perturbed training samples $\boldsymbol{x}_{j,i}^k + \boldsymbol{\delta}_{j,i}$. Given the fact that the perturbation $\boldsymbol{\delta}_{j,i}$ is generated in a way such that the selected fairness notion would be maximized, optimizing the fairness-aware loss over the fairness-adversarial examples is essentially performing worst-fair training. This worst-fair training can achieve both local fairness and robustness of the local fairness against domain shift.

In comparison, the maximization in Equation 10 specifies the goal of the fairness adversary is to generate the worst-fair training samples with a specific budget ϵ. The budget ϵ limits the perturbation radius and avoids infinity solutions to the maximization problem. To obtain $\boldsymbol{\delta}_{j,i}$, the maximization is solved using projected gradient descent (PGD) Wang et al. (2021b); Shafahi et al. (2019); Madry et al. (2017); Zeng et al. (2021a); Zhang et al. (2019). Regarding the computation of the gradients for the perturbations, we highlight that $\hat{\Phi}_*$ is differentiable, whereas the original Φ_D, Φ_E or Φ_A is not necessarily differentiable. When computing Φ_D, Φ_E or Φ_A with Equation 5, Equation 6 and Equation 8, the $\arg\max$ operation will be applied to the output logits returned by the classifiers to produce the final discrete predictions, which makes Φ_D, Φ_E or Φ_A non-differentiable. Therefore, in our implementation, we compute soft scores for these metrics by plugging in the normalized output logits instead of the predictions to enable the PGD.

4.2 Reference Domain Smoothing

As motivated in Section 3, another challenge in obtaining a fair global model is due to the domain discrepancy among the local data distributions. Therefore, after obtaining the local fair models using **local worst-fair training**, in this section, we describe how we obtain a global fair model by transforming the aggregation of local fair models to a problem of reducing domain discrepancy of training data distributions across local clients.

To formally define reference domain smoothing, we firstly re-write Equation 10 to be an empirical training loss over a modified fair training distribution $\mathcal{P}_k'$ by absorbing the instance re-weighting from IDF and fairness-adversarial examples into the original local distribution $\mathcal{P}_k$:

$$\begin{aligned}\mathcal{L}_{fair}(f_{\boldsymbol{\theta}_k}) &= \frac{1}{C^k}\sum_{j=1}^{C^k}\left[\frac{1}{M_j}\sum_{i=1}^{M_j}\left[l(f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_{j,i}^k + \boldsymbol{\delta}_{j,i}), y_{j,i}^k)\right]\right] \\ &:\approx \mathbb{E}_{(\boldsymbol{x}_i^{'k}, y_i^k)\sim\mathcal{P}_k'}\left[l(f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_i^{'k}), y_i^k)\right],\end{aligned} \tag{11}$$

where $\boldsymbol{x}^{'k}$ represents the data points sampled from the modified local data distribution $\mathcal{P}_k'$ of client k (namely $\boldsymbol{x}^{'k} = \boldsymbol{x}^k + \boldsymbol{\delta}$). The modification stems from two aspects. On one hand, the IDF re-weighting modifies the importance weights of training loss on individual training samples from i.i.d. to be non-i.i.d. and fairness-aware. Moreover, the worst-fair perturbations generated using $f_{\boldsymbol{\theta}_k}$ also change the original data distribution $\mathcal{P}_k$, making $\mathcal{P}_k'$ a function of $f_{\boldsymbol{\theta}_k}$. Note that since minimizing Equation 11 will lead to a fair model, we define the modified training distribution $\mathcal{P}_k'$ as a fair training distribution on client k.

Remark 1. On the local client k, since $\mathcal{P}_k'$ is a function of the local model $f_{\boldsymbol{\theta}_k}$, and the local model $f_{\boldsymbol{\theta}_k}$ is trained to fit $\mathcal{P}_k'$, $f_{\boldsymbol{\theta}_k}$ is also a function of $\mathcal{P}_k'$. Therefore, the local fairness Φ_*^k is a function of $f_{\boldsymbol{\theta}_k}$ and $\mathcal{P}_k'$:

$$\mathcal{P}_k' := \mathcal{P}_k'(f_{\boldsymbol{\theta}_k}), \quad f_{\boldsymbol{\theta}_k} := f_{\boldsymbol{\theta}_k}(\mathcal{P}_k'), \quad \Phi_*^k = \Phi_*^k(f_{\boldsymbol{\theta}_k}(\mathcal{P}_k')). \tag{12}$$

We note that according to Equation 12, the local models are now trained based on modified fair training distribution $\mathcal{P}_k'$. The discrepancy among $\mathcal{P}_k'$s is defined as the heterogeneity of local fairness across different clients. Therefore, we propose to reduce the heterogeneity of the fairness across different local models by reducing the domain discrepancy of $\mathcal{P}_k'$s among all local clients. However, due to data privacy constraints in the FL setting, it is infeasible to directly measure the domain discrepancy of training data across different clients.

To overcome these challenges, we propose reference domain smoothing, where a reference domain $\mathcal{Q}$ will be simulated for all clients as a domain reference. All data domain on local clients will be smoothed towards the simulated reference domain, so that the domain discrepancy among the modified local data domain $\mathcal{P}'_k$ across clients will be reduced. We aim at reducing the domain discrepancy between $\mathcal{Q}$ and $\mathcal{P}'_k$ rather than $\mathcal{P}_k$, because with Equation 12, the local models are now trained over $\mathcal{P}'_k$ instead of $\mathcal{P}_k$. Within reference domain smoothing, we use the maximum mean discrepancy (MMD) distance to quantify the domain distance between the simulated distribution $\mathcal{Q}$ of the reference domain and $\mathcal{P}'_k$ on local clients. In our implementation of computing MMD, we only use fairness-adversarial examples to approximate $\mathcal{P}'_k$.

MMD estimates the domain distance between two data distributions using samples drawn from them (Gretton et al., 2012). In our problem, given the modified fair local data distribution $\mathcal{P}'_k$ on client k and the simulated global distribution $\mathcal{Q}$, the MMD distance $\mathcal{D}_{\text{MMD}}$ is defined as:

$$\mathcal{D}_{\text{MMD}}(\mathcal{P}'_k, \mathcal{Q}) = \sup_{\mathcal{K} \in \mathcal{H}} \big(\mathbb{E}_{\boldsymbol{x} \sim \mathcal{P}'_k}[\mathcal{K}(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x} \sim \mathcal{Q}}[\mathcal{K}(\boldsymbol{x})] \big), \tag{13}$$

where $\mathcal{K}$ is a function (kernel) in reproducing the kernel Hilbert space $\mathcal{H}$. In practice, we implement the supremum of the expectation in Equation 13 by using the output logits of the samples from two different data domains as in Long et al. (2015); Yue et al. (2021). As for the kernel function, we use Gaussian kernel i.e., $\mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma})$. Using the kernel trick, the squared formulation of MMD distance $\mathcal{L}_{MMD}$ between the $\mathcal{P}'_k$ and $\mathcal{Q}$ could be simplified as:

$$\begin{aligned} \mathcal{L}_{MMD}(f_{\boldsymbol{\theta}_k}, \boldsymbol{x}^{'k}) = & \frac{1}{|\boldsymbol{X}_k||\boldsymbol{X}_k|} \sum_{i=1}^{|\boldsymbol{X}_k|} \sum_{j=1}^{|\boldsymbol{X}_k|} \mathcal{K}(f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_i^{'k}), f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_j^{'k})) \\ & + \frac{1}{|\boldsymbol{X}_Q||\boldsymbol{X}_Q|} \sum_{i=1}^{|\boldsymbol{X}_Q|} \sum_{j=1}^{|\boldsymbol{X}_Q|} \mathcal{K}(f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_i^{Q}), f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_j^{Q})) \\ & - \frac{2}{|\boldsymbol{X}_k||\boldsymbol{X}_Q|} \sum_{i=1}^{|\boldsymbol{X}_k|} \sum_{j=1}^{|\boldsymbol{X}_Q|} \mathcal{K}(f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_i^{'k})), f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_j^{Q})), \end{aligned} \tag{14}$$

where $\boldsymbol{X}_Q$ is a set of training data points $\boldsymbol{x}^Q$ sampled from the simulated reference domain $\mathcal{Q}$, and $\boldsymbol{x}^{'k}$ represents the worst-fair training samples defined in Equation 10. Finally, regarding $\mathcal{Q}$, we explicitly choose a multivariate Gaussian distribution as the simulated distribution, so that the optimization of the MMD loss could be stabilized. Both the mean vector μ_Q and the covariance matrix Σ_Q of $\mathcal{Q}$ are derived using the weighted sum of the mean and the covariance from local data with a secure aggregation scheme (McMahan et al., 2017).

By minimizing Equation 14 on all local clients, each local domain is smoothed towards the simulated reference domain $\mathcal{Q}$. In this way, the domain discrepancy across all local clients will be implicitly reduced as well, indicating that the fair local distributions $\mathcal{P}'_k$s become more similar to each other, and thereby improving the compatibility of local fairness. Eventually, we could reduce the heterogeneity of local fairness across different local clients despite the difference among local training distributions.

4.3 Overall Framework

The overall structure of WFDS consists of two modules introduced above. The local worst-fair training and the reference domain smoothing are optimized simultaneously to obtain robust local fairness and reduce the heterogeneity of local fairness:

$$\begin{aligned} & \boldsymbol{\theta}_k = \arg\min_{\boldsymbol{\theta}_k} \left[\frac{1}{C^k} \sum_{j=1}^{C^k} \Big[\frac{1}{M_j} \sum_{i=1}^{M_j} \big[l(f_{\boldsymbol{\theta}_k}(\boldsymbol{x}_{j,i}^{'k}), y_{j,i}^k) \big] \Big] + \lambda \cdot \mathcal{L}_{MMD}(f_{\boldsymbol{\theta}_k}, \boldsymbol{x}^{'k}) \right] \\ & s.t. \quad \boldsymbol{x}^{'k} = \boldsymbol{x}^k + \boldsymbol{\delta}, \quad \boldsymbol{\delta} = \arg\max_{\boldsymbol{\delta}} \hat{\Phi}_*(f_{\boldsymbol{\theta}_k}), \quad \|\boldsymbol{\delta}\| \leq \epsilon, \quad \boldsymbol{x}^k \in \boldsymbol{X}_k \quad k \in \{1, ..., K\}. \end{aligned} \tag{15}$$

Note that in Equation 15, a non-negative scalar λ is tuned to control the strength of reference domain smoothing. In our experiments, we conducted systematic robustness study to investigate the relation between the efficacy of WFDS and λ. As for the final global aggregation step, it could be any existing FL aggregation protocol. We use FedAvg in our implementation.

Table 1: Main results for improving demographic fairness of FL models with different schemes.

Dataset	Method	$\mathcal{A}_B \uparrow$	$\mathcal{A}_{sub} \downarrow$	$\Phi_D \downarrow$	$\Phi_E \downarrow$
Income	FedAvg	0.759 ± 0.002	1.034 ± 0.019	0.313 ± 0.005	0.513 ± 0.011
	AFL	0.786 ± 0.003	0.759 ± 0.047	0.289 ± 0.008	0.379 ± 0.024
	FairBatch	0.757 ± 0.003	1.032 ± 0.026	0.293 ± 0.004	0.468 ± 0.011
	FairFed	0.760 ± 0.002	1.010 ± 0.018	0.308 ± 0.004	0.498 ± 0.013
	WFDS-KL	$\mathbf{0.811 \pm 0.003}$	$\mathbf{0.283 \pm 0.048}$	$\mathbf{0.218 \pm 0.012}$	$\mathbf{0.130 \pm 0.019}$
	WFDS-MMD	$\mathbf{0.819 \pm 0.002}$	$\mathbf{0.172 \pm 0.021}$	$\mathbf{0.193 \pm 0.010}$	$\mathbf{0.096 \pm 0.006}$
COMPAS	FedAvg	0.580 ± 0.004	1.970 ± 0.028	0.384 ± 0.036	0.769 ± 0.071
	AFL	0.652 ± 0.011	0.856 ± 0.124	0.208 ± 0.017	0.347 ± 0.044
	FairBatch	0.644 ± 0.009	0.944 ± 0.134	0.183 ± 0.037	0.301 ± 0.075
	FairFed	0.647 ± 0.009	0.945 ± 0.140	0.202 ± 0.038	0.335 ± 0.079
	WFSD-KL	$\mathbf{0.666 \pm 0.004}$	$\mathbf{0.315 \pm 0.059}$	$\mathbf{0.119 \pm 0.014}$	$\mathbf{0.158 \pm 0.029}$
	WFDS-MMD	$\mathbf{0.668 \pm 0.004}$	$\mathbf{0.192 \pm 0.058}$	$\mathbf{0.082 \pm 0.009}$	$\mathbf{0.077 \pm 0.019}$

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets In our experiments, we use two real-world datsaets. (1) UCI Census Income dataset (Kohavi, 1996): the task of this dataset is to predict whether the annual income of a person can exceed 50K $ given her/his profile. For this dataset, we select gender as the demographic attribute, and we distribute the training data over 10 different local clients. (2) COMPAS Recidivism Racial Bias dataset (ProPublica, 2022): the desired output for this dataset is the prediction of "recidivism" (positive class) or not given a person's profile. For this dataset, we select race as the demographic attribute (one selected race v.s. all remaining races), and we distribute the training data over 5 different local clients due to data scarcity.

Evaluation Metrics and Baselines To evaluate the fairness of the FL models, we use all fairness notions defined in Section 3. Demographic parity Φ_D and Equalized odds Φ_E are commonly used fairness metrics as in Zhang et al. (2018); Beutel et al. (2017); Zhao et al. (2017); Kou et al. (2021). Inspired by Yue et al. (2022), we also incorporate the evaluation of the models' performance on negative predictions by computing the balanced accuracy $\mathcal{A}_B$ and the sub-group accuracy gap Φ_A. Considering the code availability and whether the experimental setting is comparable with ours, we select FedAvg (McMahan et al., 2017), AFL (Mohri et al., 2019), FairBatch (Roh et al., 2020), and FairFed (Ezzeldin et al., 2021) as the baseline algorithms for comparison. Moreover, we use multi-layer perceptron (MLP) as the model architecture for all experiments. Finally, all experiments are repeated for 10 times.

5.2 EXPERIMENTAL RESULTS

Efficiency Evaluation In the first set of experiments, we compare our WFDS scheme with the baselines methods and present evaluation results for all datasets in Table 1. Each row of Table 1 represents an FL training scheme, and each column includes metric scores with mean and standard deviation. We report balanced accuracy $\mathcal{A}_B$, sub-group accuracy gap $\mathcal{A}_{sub}$, demographic parity Φ_D and equalized odds Φ_D. The best results are marked in bold. We observe: (1) the proposed WFDS framework significantly reduces the demographic bias of the classifiers in FL applications, while outperforming all baseline methodds. For instance, on Income dataset, $\mathcal{A}_{sub}$, Φ_D and ϕ_D of the trained FL model using WFDS-MMD are reduced to 0.172, 0.193, 0.096 from 1.034, 0.313, 0.513 of the FedAvg. (2) In addition, WFDS also achieves a higher balanced accuracy $\mathcal{A}_B$ compared to other methods. For instance, on Income dataset, WFDS-MMD achieves a balanced accuracy of 0.819, whereas $\mathcal{A}_B$ of all other methods including FedAvg is lower than 0.8. (3) Finally, on both datasets, WFDS-MMD consistently outperforms WFDS-KL on all evaluation metrics, indicating that MMD is indeed the best scheme in terms of reducing the domain discrepancy between local training data

distributions and mitigating the heterogeneity of local fairness across all clients. Similar trend is also observed for demographic parity and equalized odds.

Robustness Study We study the robustness of WFDS-MMD w.r.t. the hyperparameter λ. In particular, we vary the λ but fix other configuration of our framework for the robustness study. The results on Income dataset are reported in Table 2. We observe that in general with larger λ, the FL model becomes more fair in terms of the sub-group accuracy gap $\mathcal{A}_{sub}$, demographic parity Φ_D and equalized odds Φ_E. Regarding the balanced accuracy, WFDS becomes less sensitive to λ. Due to space limit, additional results on COMPAS dataset are in Appendix A.

Table 2: Robustness study for WFDS with different global domain smoothing strength on Income.

Dataset	$\boldsymbol{\lambda}$	$\mathcal{A}_B \uparrow$	$\mathcal{A}_{sub} \downarrow$	$\Phi_D \downarrow$	$\Phi_E \downarrow$
Income	0.1	0.817 ± 0.003	0.207 ± 0.045	0.201 ± 0.011	0.107 ± 0.020
	0.3	0.816 ± 0.002	0.191 ± 0.035	0.197 ± 0.011	0.100 ± 0.014
	0.5	0.818 ± 0.004	0.181 ± 0.031	0.194 ± 0.011	0.098 ± 0.010
	0.7	0.818 ± 0.003	0.174 ± 0.031	$\mathbf{0.191 \pm 0.012}$	0.099 ± 0.009
	0.9	$\mathbf{0.819 \pm 0.002}$	$\mathbf{0.172 \pm 0.021}$	0.193 ± 0.010	$\mathbf{0.096 \pm 0.006}$

Ablation Study We evaluate the contribution of the worst-fair training module and the reference domain smoothing module by comparing our results from WFDS-MMD to the results trained without worst-fair training and reference domain smoothing. The results on Income dataset are reported in Table 3. Note that removing the worst-fair training module only masks out the worst-fair training samples but preserves the inversely proportional frequency loss during the training. In comparison, removing the reference domain smoothing module implies that the global fairness is achieved only with robust local fairness. As expected, we observe that the model's performance in regards to all evaluation metrics becomes worse by removing either module in the WFDS framework. For instance, the sub-group accuracy gap increases by 151.7% and 26.7% when removing the worst-fair training module and the reference domain smoothing module. Similarly, $\mathcal{A}_{sub}$, Φ_D and Φ_E becomes larger when removing either module. Due to space limit, additional results on COMPAS dataset are reported in Appendix A.

Table 3: Ablation study of WFDS on Income.

Income	$\mathcal{A}_B \uparrow$	$\mathcal{A}_{sub} \downarrow$	$\Phi_D \downarrow$	$\Phi_E \downarrow$
WFDS-MMD	$\mathbf{0.819 \pm 0.002}$	$\mathbf{0.172 \pm 0.021}$	$\mathbf{0.193 \pm 0.010}$	$\mathbf{0.096 \pm 0.006}$
w/o Worst-Fair Training	0.805 ± 0.002	0.433 ± 0.032	0.247 ± 0.009	0.217 ± 0.016
w/o Ref. Domain Smoothing	0.817 ± 0.003	0.218 ± 0.055	0.204 ± 0.013	0.114 ± 0.024

6 CONCLUSION

In this paper, we propose a novel worst-fair domain smoothing framework for addressing demographic bias issue in federated learning applications. We design our framework by jointly considering the robustness of local fairness and the domain discrepancy among the training data across all local clients. To the best of our knowledge, WFDS is the first work that addresses the demographic bias issue of FL models from a domain-shifting perspective. Experimental results on real-world datasets demonstrate that our method significantly improves the fairness of FL models by outperforming state-of-the-art baseline methods.

REFERENCES

Fabio Bacchini and Ludovica Lorusso. Race, again: how face recognition technology reinforces racial discrimination. *Journal of information, communication and ethics in society*, 2019.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems*, 34:26091–26102, 2021.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 181–189. SIAM, 2021.

Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. *arXiv preprint arXiv:2110.00857*, 2021.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Balancing out bias: Achieving fairness through training reweighting. *arXiv preprint arXiv:2109.08253*, 2021.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

Ziyi Kou, Lanyu Shang, Huimin Zeng, Yang Zhang, and Dong Wang. Exgfair: A crowdsourcing data exchange approach to fair human face datasets augmentation. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 1285–1290. IEEE, 2021.

SM Angwin J Larson and Lauren Kirchner. There's software used across the country to predict future criminals. and it's biased against blacks, 2016.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pp. 4615–4625. PMLR, 2019.

ProPublica. Compas recidivism risk score data and analysis. 2022. URL `URLhttps://www.propublica.org/datastore/results?q=compas`.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021a.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. *arXiv preprint arXiv:2112.08304*, 2021b.

Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. Contrastive domain adaptation for question answering using limited text corpora. *arXiv preprint arXiv:2108.13854*, 2021.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. *arXiv preprint arXiv:2208.09578*, 2022.

Huimin Zeng, Chen Zhu, Tom Goldstein, and Furong Huang. Are adversarial examples created equal? a learnable weighted minimax risk for robustness under non-uniform attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10815–10823, 2021a.

Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021b.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Daniel Yue Zhang, Ziyi Kou, and Dong Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1051–1060. IEEE, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.