

GLOCALFAIR: Jointly Improving Global and Local Group Fairness in Federated Learning

Syed Irfan Ali Meerza¹, Luyang Liu², Jiaxin Zhang³, Jian Liu¹

¹University of Tennessee Knoxville

²Google Research

³Intuit AI Research

smeerza@vols.utk.edu, luyangliu@google.com, jxzhangai@gmail.com, jliu@utk.edu

Abstract

Federated learning (FL) has emerged as a prospective solution for collaboratively learning a shared model across clients without sacrificing their data privacy. However, the federated learned model tends to be biased against certain demographic groups (e.g., racial and gender groups) due to the inherent FL properties, such as data heterogeneity and party selection.

Unlike centralized learning, mitigating bias in FL is particularly challenging as private training datasets and their sensitive attributes are typically not directly accessible. Most prior research in this field only focuses on global fairness while overlooking the local fairness of individual clients. Moreover, existing methods often require sensitive information about the client’s local datasets to be shared, which is not desirable. To address these issues, we propose GLOCALFAIR, a client-server co-design fairness framework that can jointly improve global and local group fairness in FL without the need for sensitive statistics about the client’s private datasets. Specifically, we utilize constrained optimization to enforce local fairness on the client side and adopt a fairness-aware clustering-based aggregation on the server to further ensure the global model fairness across different sensitive groups while maintaining high utility. Experiments on two image datasets and one tabular dataset with various state-of-the-art fairness baselines show that GLOCALFAIR can achieve enhanced fairness under both global and local data distributions while maintaining a good level of utility and client fairness.

1 Introduction

Federated Learning (FL) [McMahan *et al.*, 2017] allows multiple data holders (i.e., *clients*) to collaboratively train a global model under the coordination of a central *server* without directly sharing their private datasets. Its strong emphasis on client privacy has made it particularly attractive for many privacy-critical applications, such as speaker verification [Granqvist *et al.*, 2020], predictive typing [Hard *et al.*, 2018], image classification [Chen and Chao, 2021] and healthcare [Rieke *et al.*, 2020; Xu *et al.*, 2021]. Although

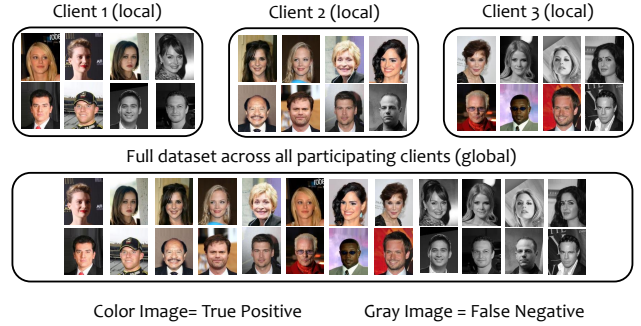


Figure 1: An illustrative example of bias in classification when considering two sensitive gender groups (i.e., male vs. female) in FL with three participating clients. At the global scale, the model is fair due to the equal true-positive rate (or false-negative rate). However, when considering the model’s fairness on each client, the classification result is biased. Note that the model performs differently between clients mainly due to the client non-IIDness in FL, such as different distributions in racial and age groups.

providing great privacy benefits, the inherent characteristics of FL, such as data heterogeneity, party selection, and client dropping out, will inevitably make the resulting FL model biased against specific demographic groups (e.g., racial and gender groups), causing specific populations to receive undesirable treatments. Additionally, to protect clients’ privacy, in FL the server is prohibited from directly accessing the clients’ datasets, rendering most centralized bias mitigation strategies inapplicable [Grgić-Hlača *et al.*, 2018; Zhang *et al.*, 2018; Lohia *et al.*, 2019]. Therefore, how to improve fairness has become an open problem in this field and has gained considerable research attention recently [Kairouz *et al.*, 2021].

Existing Efforts in Improving Fairness in FL. Initial studies [Li *et al.*, 2020; Mohri *et al.*, 2019; Lyu *et al.*, 2020; Li *et al.*, 2021; Zhang *et al.*, 2020b] in this field mainly focus on *client fairness*, which aims to promote equalized accuracies across all clients. This is typically achieved through reducing the client’s performance disparity [Li *et al.*, 2020] or maximizing the performance of the worst client [Mohri *et al.*, 2019]. However, under such frameworks, the FL model can show similar accuracies on all clients while still being biased against certain under-represented groups. To address this, a few recent studies [Abay *et al.*, 2020; Zhang *et al.*, 2020a; Rodríguez-Gálvez *et al.*, 2021; Chu *et al.*, 2021; Ezzeldin *et al.*, 2023; Du *et al.*, 2021] have pro-

posed *group fairness* solutions for improving fairness across different sensitive groups. These studies mainly utilize deep multi-agent reinforcement learning [Zhang *et al.*, 2020a], re-weighting mechanisms [Abay *et al.*, 2020; Ezzeldin *et al.*, 2023; Du *et al.*, 2021], optimization with fairness constraints [Chu *et al.*, 2021] or the modified method of differential multipliers [Rodríguez-Gálvez *et al.*, 2021] to achieve group fairness in FL.

Limitations of Existing Group Fairness Research in FL

While the aforementioned group fairness approaches can improve FL fairness in certain cases, most of them [Zhang *et al.*, 2020a; Rodríguez-Gálvez *et al.*, 2021; Chu *et al.*, 2021; Du *et al.*, 2021] only focus on achieving *global fairness* under the overall data distribution, without considering *local fairness* on individual clients. Due to the non-IID¹ nature of the client’s data in FL, local distribution at a specific client may not reflect the exact global data distribution. Thus, merely achieving global fairness cannot provide guarantees of local fairness. In practice, this could result in emergent group unfairness at individual clients due to the client’s “non-IIDness” in FL. Figure 1 provides an illustrative example of the difference between global group fairness and local group fairness. To further illustrate, let’s consider a cross-silo FL scenario where image classifiers for brain tumor detection are trained across several hospitals with distinct patient demographic compositions. In such a case, local fairness is more important since the model will eventually be deployed at each client and thus would directly affect the model’s diagnosis accuracy [Cui *et al.*, 2021]. To the best of our knowledge, only a few studies [Abay *et al.*, 2020; Ezzeldin *et al.*, 2023] attempt to address the local group fairness issue. However, they all require clients to share sensitive statistics (e.g., data sample counts per sensitive group or group-specific positive prediction rates), which would raise privacy concerns. Additionally, they rely on either reweighting or debiasing-based mechanisms to ensure local fairness, which will be heavily influenced by the distribution of local datasets, thus they may only achieve suboptimal fairness with highly heterogeneous datasets in FL.

To circumvent the aforementioned limitations, in this paper, we propose GLOCALFAIR, the first client-server co-design fairness framework that can jointly improve global and local group fairness to learn a fair global model on both local datasets of individual clients and the global dataset. Specifically, to ensure local fairness, we pose the classification task on each client as a constrained optimization problem with a constraint that maximizes the true-positive rate (TPR) of sensitive groups and minimizes the false-positive rate (FPR). We then transform the constrained optimization problem into a two-player zero-sum game. To ensure global group fairness, we incorporate a clustering-based aggregation method on the server side. Specifically, we cluster the client updates based on their fairness level measure using the Gini Coefficient [Gini, 1912] and calculate a weighted mean within each cluster, based on their dataset size, to update the global model. We believe such two-layer fairness strategies can ensure enhanced model fairness at both global and local levels.

Our framework brings several benefits over existing meth-

ods: (1) *Improved Data Privacy*: In GLOCALFAIR, the server assigns weight to client updates using a clustering algorithm. This way does not require the server to have access to any information about each client’s sensitive local dataset; (2) *Customizable Fairness Configuration on Client Model*: By adopting constrained optimization at the client’s side, our framework provides the client with the freedom to choose which fairness metric (FPR or TPR) to emphasize locally. For instance, in the brain tumor diagnosis application, it is more expensive to falsely classify a tumor case as healthy (false negative) than to diagnose a healthy person as malignant (false positive); and (3) *Multiple Sensitive Attributes*: Leveraging constrained optimization, our framework can be naturally generalized to the scenario with multiple sensitive attributes by incorporating additional constraints.

We conduct extensive experiments involving multiple federated datasets, including image, and tabular data formats, and various state-of-the-art FL fairness baselines, such as group-fairness studies FAIRFED [Ezzeldin *et al.*, 2023] and FEDFB [Zeng *et al.*, 2022], and client-fairness studies q-FFL [Li *et al.*, 2020] and GIFAIR [Yue *et al.*, 2023]. The results demonstrate that GLOCALFAIR can achieve enhanced fairness under both global and local data distributions while maintaining a good level of model utility and client fairness.

2 Related Work

Client Fairness in Federated Learning. The goal of client fairness is to diminish the performance difference of models across different clients (known as the *client parity*). To achieve this, Li *et al.* [Li *et al.*, 2020] propose q-Fair Federated Learning (q-FFL), an algorithm that aims to encourage a uniform accuracy distribution across all clients by re-weighting the clients according to their local empirical loss. Mohri *et al.* [Mohri *et al.*, 2019] propose an agnostic federated learning framework to maximize the performance of the worst-performing client. GIFAIR [Yue *et al.*, 2023] leverages a regularization to penalize spread in client loss for uniform performance. Differently, Ditto [Li *et al.*, 2021] seeks to reduce a client’s performance variance through personalization. In addition, there are some other studies [Lyu *et al.*, 2020; Zhang *et al.*, 2020b] aiming to achieve proportional fairness that allows clients with higher contributions to be rewarded with a better-performing local model. Despite improving client disparity, these methods still fail to address the discrimination against certain demographic groups in the FL scenario due to the high data heterogeneity across clients.

Group Fairness in Federated Learning. A better practice of responsible AI is group fairness, which requires the model to treat different demographic groups defined by the sensitive attribute (e.g., race and gender) equally. Several recent studies have explored group fairness in the FL setting from the global fairness perspective. In particular, FairFL [Zhang *et al.*, 2020a], AgnosticFair [Du *et al.*, 2021], FCFL [Cui *et al.*, 2021], and FEDFB [Zeng *et al.*, 2022] propose to achieve global group fairness by incorporating fairness constraints into objective functions. However, this requires clients to share sensitive statistical information with the server. Beyond global fairness, only a few studies have tackled the local fairness issue by designing fairness mechanisms on the client side or the

¹IID: identically and independently distributed.

server side. For instance, Abay *et al.* [Abay *et al.*, 2020] propose to adopt centralized fair learning strategies by applying debiasing mechanisms locally on each client. Unfortunately, this method would be less effective if the client’s sensitive attribute distributions are highly heterogeneous, making it unfavorable in practical scenarios. FAIRFED [Ezzeldin *et al.*, 2023] can adaptively modify the aggregation weights on the server side according to the mismatch between the server’s global fairness measure and the client’s local fairness measure. However, this method requires the statistics from the client’s private datasets to be shared at the start of the training. Differently, in this paper, we propose a client-server co-design framework in FL for jointly improving global and local group fairness regarding multiple sensitive attributes without requiring clients to share any sensitive information about their local datasets.

3 Primer on Federated Learning and Fairness

3.1 Federated Learning

The goal of FL is to collaboratively train a global machine-learning model among a large number of distributed clients under the coordination of a central server without accessing client data. Let θ denote the learned parameters of the global model, then the main objective of FL is to solve:

$$\min_{\theta} f(\theta) = \sum_{k=1}^n p_k L_k(\theta), \quad L_k = \frac{1}{d_k} \sum_{j_k=1}^{d_k} l_{j_k}(\theta), \quad (1)$$

where n is the total number of clients. For the k -th client, p_k denotes the probability of its participation in the current training round ($\sum_{k=1}^n p_k = 1$). L_k is the empirical loss over its local data where l_{j_k} is the loss on the j -th sample of the k -th client, and d_k is the number of local data samples. The majority of studies solve the optimization problem formulated in Equation 1 by choosing a selection of clients, each with probability p_k at each round and then executing an optimizer such as stochastic gradient descent (SGD) for several iterations locally at each client. FEDAVG [McMahan *et al.*, 2017] is a commonly used approach for solving Equation 1 in non-convex settings. The approach works simply by applying E epochs of SGD locally at each chosen client and then averaging the resulting local models. Unfortunately, this approach introduces highly inconsistent performance between different clients. Specifically, the resulting global model can be biased toward the clients with a greater volume of data points or those who frequently engage in the training process. Additionally, naive averaging can result in a global model that is biased against certain demographic groups presented in the dataset.

3.2 Fairness Notion

In client fairness, if the performance of *model-a* on the n clients is more uniform than the performance of *model-b* on the n clients, we informally state that *model-a* provides a fairer solution than *model-b*. For group fairness, the model performance is measured with respect to how it performs compared to the underlying groups defined by sensitive attribute G . If a model performs equally for the privileged group $G = 1$ and the underprivileged group $G = 0$, then the model is fair through the group fairness lens. Formally, given data sample X and its corresponding label Y , for a model with binary pre-

diction $\hat{Y}$, we consider the following two metrics to quantify its group fairness:

Demographic Parity [Dwork *et al.*, 2012]: If a classifier’s prediction $\hat{Y}$ is statistically independent of G , it meets demographic parity under a distribution (X, G, Y) . This is equivalent to $\mathbb{E}[\hat{Y}|G = a] = \mathbb{E}[\hat{Y}]$, where $a = 0$ or 1 for a binary group. Demographic parity can also be measured by the Demographic Parity Difference (DPD), with $Pr(\cdot)$ denoting the probability:

$$DPD = \left| Pr(\hat{Y} = 1|G = 1) - Pr(\hat{Y} = 1|G = 0) \right|. \quad (2)$$

Equalized Odds [Hardt *et al.*, 2016]: If a classifier satisfies equalized odds, its predictions $\hat{Y}$ are conditionally independent of G given the true label Y under a distribution (X, G, Y) . Mathematically, this can be expressed as $\mathbb{E}[\hat{Y}|G = a, Y = 1] = \mathbb{E}[\hat{Y}|Y = 1]$, where a represents a sensitive group. In other words, a model satisfies equalized odds if the probability of a true positive and the probability of a false positive are the same across different sensitive groups. Equalized odds can be formulated as:

$$Pr(\hat{Y} = 1|Y = y, G = 1) = Pr(\hat{Y} = 1|Y = y, G = 0). \quad (3)$$

3.3 Global and Local Group Fairness

Clients in FL generally have non-IID data distributions. To account for this data heterogeneity, we further consider two levels of fairness: global and local group fairness. The global fairness performance of a given model considers the entire dataset across the n clients who participate in FL, whereas the local fairness performance is determined by only applying the fairness metrics described in Equations 2 and 3 to the local dataset D_k at client k . Because of the non-IID characteristics of the local data distributions, the overall global data distribution may not be well represented by a single local distribution at any of the clients. As a result, only mitigating the fairness issue on each client may not improve global fairness. And vice versa, improving global fairness may not ensure local fairness for each client. To balance local and global fairness, we propose a method that can jointly mitigate the bias in the trained model, addressing both the individual client’s local datasets and the full datasets spanning all participating clients.

4 Design of GLOCALFAIR

4.1 Design Overview

Due to the inherent data heterogeneity in FL, the distribution of local datasets on individual clients may be highly distinct from the distribution of the full dataset across all participating clients. Thus, merely achieving global fairness cannot provide a guarantee of local fairness. To solve this challenge, we propose a client-server co-design fairness framework, GLOCALFAIR, to jointly improve both global and local group fairness. As shown in Figure 2, we consider n clients that collaboratively train a global model. On the client side, we use constrained optimization to ensure equitable performance of the shared global model across each sensitive group present in the local datasets. Specifically, in communication round t , all participating clients first download the current global model (θ_{global}^t) from the central server and then compute local model updates on their local datasets, with constraints to minimize

False Positive Rate (FPR) and False Negative Rate (FNR) to values below preset thresholds set by τ_{FNR} and τ_{FPR} . Next, the clients upload their model updates (θ_k^t , $k \in \{1, \dots, n\}$) to the server to participate in the federated training. On the server side, we employ K-means clustering to group the client updates based on their Gini coefficients [Gini, 1912], which serve as a fairness proxy. Subsequently, the updates of each cluster are aggregated, where the weight assigned to each update is proportional to the amount of data utilized to train the local model. Then, the aggregated updates undergo another round of aggregation, where a set of weights $\hat{w}_1, \dots, \hat{w}_n$ is assigned based on their cumulative Gini coefficient value, favoring updates that exhibit greater fairness. This results in a more balanced and fair distribution of updates. The resulting aggregated update is then used to update the global model, leading to a more equitable and unbiased model ($\theta_{\text{global}}^{t+1}$). Finally, this updated model is sent back to the clients in the next round, ensuring that all participants benefit from the fairness improvements gained on both the server and client sides.

4.2 Local Fairness by Imposing Constraints

Local Fairness Constraints

To ensure local fairness (e.g., group fairness of the learned model θ_k^t on the k -th client's local dataset), we propose a constrained optimization-based approach that is employed on each client. Specifically, in GLOCALFAIR, the fairness problem is recast as a problem of minimizing an empirical loss while adhering to one or more fairness constraints. We constrain the model with two fairness metrics, namely, False Negative Equality Difference (FNEDE) and False Positive Equality Difference (FPEDE),

Thus, the binary classification task, incorporating fairness constraints to mitigate unintended bias, can be formulated as follows:

$$\begin{aligned} \min_{\theta_k^t \in \Theta} \quad & \frac{1}{d_k} \sum_{X \in D_k} f_L(X; \theta_k^t), \\ \text{s.t.} \quad & \max\{\text{FNR}_{G_1}, \dots\} - \text{FNR}_{\text{overall}} \\ & < \tau_{\text{FNR}} \cdot \text{FNR}_{\text{overall}} - \min\{\text{FNR}_{G_1}, \dots\} \\ & < \tau_{\text{FNR}} \cdot \max\{\text{FPR}_{G_1}, \dots\} - \text{FPR}_{\text{overall}} \\ & < \tau_{\text{FPR}} \cdot \text{FPR}_{\text{overall}} - \min\{\text{FPR}_{G_1}, \dots\} < \tau_{\text{FPR}}, \end{aligned} \quad (4)$$

where f_L is the binary cross-entropy loss that we want to minimize, D_k is the local dataset, and θ_k^t is the local model of client k at t th communication round and τ_{FNR} and τ_{FPR} are the allowed deviations from overall FNRs and FPRs, respectively, $\text{FNR}_{\text{overall}}$ and $\text{FPR}_{\text{overall}}$ are the false negative and false positive rates on the entire local dataset, respectively, and FNR_{G_p} and FPR_{G_p} represent the corresponding rates in the group G_p .

Constraint Optimization

Solving the constrained problem in Equation 4 presents significant challenges due to the data-dependent nature of the constraints, which leads to high computational costs for verification. Furthermore, the constraints represent a linear combination of indicators, rendering them non-sub-differentiable with respect to θ_k^t . The problem with constraints outlined in Equation 4 can be resolved through the following alternative

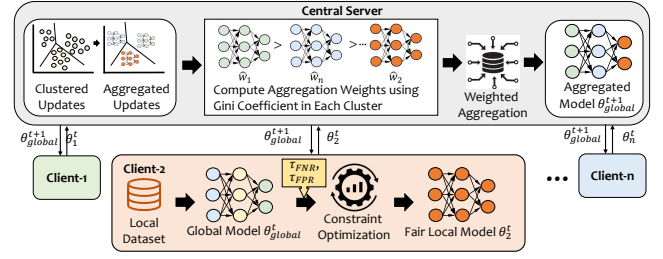


Figure 2: Illustration of the proposed GLOCALFAIR approach:

$$\min_{\theta_k^t \in \Theta} \frac{1}{d_k} \sum_{X \in D_k} f_L(X; \theta_k^t) + \sum_{i=1}^m \lambda_i g_i(\theta_k^t), \quad (5)$$

where $g_i(\theta_k^t)$, $i \in \{1, \dots, m\}$ represents the i th fairness constraint and λ_i is a dual term that maximizes the objective function when the constraints are violated and relaxes the constraints when fairness goals are achieved. To solve this problem, a gradient-based approach is not viable due to the piecewise constant nature of the constraint functions in our scenario, resulting in near-zero gradients across most points. Utilizing a surrogate is an alternative, where indicators in Equation 4 are replaced with sigmoids for gradient-based optimization. While it resolves differentiability, a new issue emerges: Nash equilibrium aligns with solutions satisfying the sigmoid-relaxed constraint rather than the real one. To solve these issues, we approximate the penalty function through a maximization task. Specifically, we convert the Equation 5 into a min-max optimization problem:

$$\max_{\lambda_i \geq 0} \min_{\theta_k^t \in \Theta} \frac{1}{d_k} \sum_{X \in D_k} f_L(X; \theta_k^t) + \sum_{i=1}^m \lambda_i g_i(\theta_k^t), \quad (6)$$

which converts the problem into a two-player zero-sum game. One player optimizes λ_i to maximize the objective, while the other optimizes θ_k^t to minimize it. We optimize the inner minimization by taking the gradient to update the model weights:

$$\theta_k^{t*} \leftarrow \theta_k^t - \eta \left(\frac{1}{d_k} \sum_{X \in D_k} \nabla f_L(X; \theta_k^t) + \sum_{i=1}^m \lambda_i \nabla g_i(\theta_k^t) \right). \quad (7)$$

Similarly, we update the dual term λ_i through gradient ascent to penalize the constraint violations and promote constraint satisfaction:

$$\lambda_i^* \leftarrow \lambda_i + \eta (\nabla g_i(\theta_k^{t*})). \quad (8)$$

4.3 Global Fairness by Fairness-aware Clustering

Regular averaging aggregation methods, such as FEDAVG [McMahan *et al.*, 2017], favor the clients that have more data points and frequently participate in the training, which potentially leads to a biased global model if these clients contribute biased local updates. In GLOCALFAIR, we address this problem by clustering client updates based on their Gini coefficients [Gini, 1912], which are used for measuring inequality in the distribution of weights in the client update. We observe that biased models often exhibit skewed or imbalanced weight distributions due to imbalanced datasets or scenarios where specific parameters or features hold disproportionate significance, particularly under our local fairness constraints. As

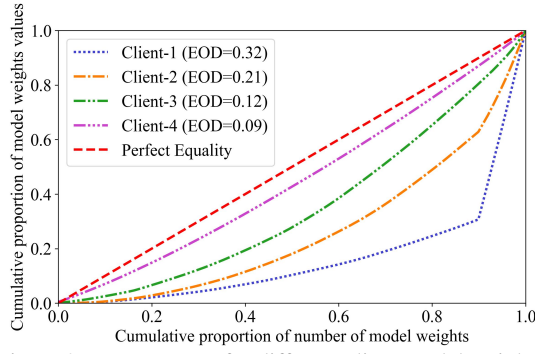


Figure 3: Lorenz curve for different client model weights.

shown in Equations 7 and 8, the introduction of fairness constraints leads to penalties when violations occur, prompting adjustments in model weights to meet these constraints. A higher bias in the model results in more pronounced changes in λ_i , consequently leading to more significant adjustments in model weights θ_k^t and thereby disrupting its weight distribution. Hence, gauging model weight distributions could serve as a useful proxy for evaluating fairness within our proposed framework.

To better understand this, we plot Lorenz curves for four local model updates with corresponding Equalized Odds Difference (EOD) in Figure 3. Within biased models (clients 1 and 2), the Lorenz curves markedly diverge from the line of perfect equality, indicating the dominance of a limited subset of parameters in their weight distributions, while the Lorenz curve of relatively fair models (clients 3 and 4 with relatively lower EOD) would closely follow the line of perfect equality (empirical analysis is presented in Supplemental Material 12.1). We employ the Gini coefficient to provide an approximate measure of the fairness level within model updates. Subsequently, we employ K-means clustering to group client updates based on their Gini coefficients:

$$\min \sum_{i=1}^p \sum_{j=1}^{n_i} \left\| \mathcal{G}_{ij} - \frac{1}{n_i} \sum_{j=1}^{n_i} \mathcal{G}_{ij} \right\|^2, \quad (9)$$

where $\mathcal{G}_{ij} = \frac{\sum_{a=1}^w \sum_{b=1}^w |W_{i,j,a}^t - W_{i,j,b}^t|}{2w^2 \hat{W}_{i,j}^t}$ is the Gini coefficient of the j th client in the i th cluster. $W_{i,j,a}^t$ and $W_{i,j,b}^t$ represents the weights in j th client's model, and $\hat{W}_{i,j}^t$ is the mean of all the weights while w is the number of weights in the j th client's model. p is the number of clusters determined by the elbow method, and n_i is the number of clients in the i th cluster. We then aggregate the updates using FEDAVG with proportional weights at each cluster. The weight assigned to each client's update is proportional to the number of samples they contributed during that round of training. Those aggregated updates are further aggregated based on the cumulative Gini coefficient of each cluster, favoring the fairer cluster. The resulting global model θ_{global}^{t+1} can be obtained by:

$$\theta_{global}^{t+1} = \theta_{global}^t + \sum_{i=1}^p \exp\left(-\gamma \frac{1}{n_i} \sum_{j=1}^{n_i} \mathcal{G}_{ij}\right) \sum_{j=1}^{n_i} \frac{d_{ij}}{\sum_{k=1}^{n_i} d_{ik}} \theta_{ij}^t, \quad (10)$$

where γ is a parameter that controls the degree of fairness. Note that when $\gamma = 0$, the global model computed in Equation 10 is equivalent to that computed by FEDAVG. d_{ij} denotes

the number of data samples contributed by the j th client in the i th cluster. By reducing the influence of variations in updates across clients and prioritizing updates from clients with lower Gini coefficients (i.e., more fairness), our approach can promote fairness and reduce bias in the global model.

5 Experimental Evaluation

5.1 Experimental Setup

Federated Datasets

(1) *CelebA* [Liu *et al.*, 2018]: A collection of 200k celebrity face images from that have been manually annotated. The dataset has up to 40 labels, each of which is binary-valued. We resize each image to 128×128px and create a classification task to classify whether the celebrity in the image is smiling or not. For CelebA, each subject's gender (male or female) and age (young or adult) are the sensitive attributes.

(2) *Adult Income* [Dua and Graff, 2017]: A tabular dataset that is widely investigated in ML fairness literature. It contains 48,842 samples with 14 attributes. We create a binary classifier to classify whether a subject's income is above 50k or not. In this dataset, race (white or non-white) and gender (male or female) are used as the sensitive attributes.

(3) *UTK Faces* [Zhang *et al.*, 2017]: A large-scale face dataset with more than 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variations in pose, facial expression, illumination, occlusion, resolution, etc. We create a binary classification task of classifying young (less than 40 years) and old (older than 40 years). Here, race (white or non-white) and gender (male or female) are used as the sensitive attributes.

Evaluation Metric

(1) *Equalized Odds Difference (EOD)*: We use EOD of each sensitive group to measure group fairness:

$$\text{EOD} = |Pr(\hat{Y} = 1|G_g = 0) - Pr(\hat{Y} = 1|G_g = 1)|_{g \in N}. \quad (11)$$

(2) *Demographic Parity Disparity (DP-Dis)*: DP-Dis is another metric used for measuring group fairness, which is calculated as:

$$\text{DP-Dis} = \max_{g \in N} |Pr(\hat{Y} = 1|G_g = a) - Pr(\hat{Y} = 1)|. \quad (12)$$

(3) *Utility*: In our experiments, we use the model's prediction accuracy on the testing dataset as the metric to quantify global model utility.

(4) *Discrepancy (DIS)*: To evaluate client fairness, we use the performance discrepancy among clients, which is measured as the difference between the highest and lowest accuracies on the clients.

Note that EOD and DP-Dis are common fairness metrics that have been widely used in the literature [Zeng *et al.*, 2022; Ezzeldin *et al.*, 2023].

Parameter Selection

GLOCALFAIR uses γ to control the balance between fairness and utility of the global model and (τ_{FNR}, τ_{FPR}) to adjust the local fairness constraint at the client's side. Although a lower value for (τ_{FNR}, τ_{FPR}) can result in better FPR and FNR, it would also increase the risk of overconstraining the model and potentially converge to a trivial solution (e.g., a model that outputs same prediction for all inputs). To address this,

Table 1: Global group fairness comparison of GLOCALFAIR with various baselines on three federated datasets.

Method	CelebA				Adult Income				UTK Faces			
	EOD ($\downarrow$)		DP-Dis ($\downarrow$)	Utility ($\uparrow$)	EOD ($\downarrow$)		DP-Dis ($\downarrow$)	Utility ($\uparrow$)	EOD ($\downarrow$)		DP-Dis ($\downarrow$)	Utility ($\uparrow$)
	Gender	Age			Gender	Race			Gender	Age		
FEDAVG [McMahan <i>et al.</i> , 2017]	0.45	0.50	0.39	83.0%	0.30	0.32	0.35	83.8%	0.41	0.48	0.31	87.0%
q-FFL [Li <i>et al.</i> , 2020]	0.36	0.39	0.34	79.3%	0.28	0.27	0.29	81.5%	0.37	0.41	0.39	86.0%
GIFAIR [Yue <i>et al.</i> , 2023]	0.30	0.27	0.25	79.9%	0.23	0.24	0.20	82.7%	0.31	0.29	0.27	85.9%
FAIRFED [Ezzeldin <i>et al.</i> , 2023]	0.16	0.20	0.17	78.6%	0.12	0.12	0.11	77.5%	0.18	0.22	0.19	85.1%
FEDFB [Zeng <i>et al.</i> , 2022]	0.21	0.17	0.15	79.1%	0.10	0.13	0.12	79.0%	0.23	0.19	0.17	85.3%
GLOCALFAIR	0.09	0.10	0.10	79.2%	0.08	0.11	0.09	81.2%	0.09	0.12	0.14	86.4%

Table 2: EOD and DP-Dis of individual clients and Discrepancy among the client performance to measure client fairness. Lower mean/std values of EOD & DP-Dis mean greater local group fairness. Lower DIS means higher client fairness.

Method	CelebA					Adult Income					UTK Faces				
	EOD (Gender Group)		DP-Dis		DIS (↓)	EOD (Gender Group)		DP-Dis		DIS (↓)	EOD (Gender Group)		DP-Dis		DIS (↓)
	Mean(↓)	Std(↓)	Mean(↓)	Std(↓)		Mean(↓)	Std(↓)	Mean(↓)	Std(↓)		Mean(↓)	Std(↓)	Mean(↓)	Std(↓)	
	FEDAVG [McMahan <i>et al.</i> , 2017]	0.43	0.25	0.39	0.27	13.5%	0.21	0.15	0.33	0.26	34.6%	0.32	0.21	0.35	0.29
q-FFL [Li <i>et al.</i> , 2020]	0.40	0.25	0.39	0.25	8.3%	0.18	0.12	0.32	0.28	28.0%	0.29	0.18	0.32	0.29	11.2%
GIFAIR [Yue <i>et al.</i> , 2023]	0.37	0.23	0.38	0.24	7.7%	0.17	0.09	0.30	0.27	11.4%	0.27	0.18	0.31	0.29	10.1%
FAIRFED [Ezzeldin <i>et al.</i> , 2023]	0.33	0.21	0.33	0.20	9.8%	0.15	0.07	0.26	0.23	25.0%	0.22	0.15	0.27	0.25	11.2%
FEDFB [Zeng <i>et al.</i> , 2022]	0.29	0.20	0.26	0.20	9.8%	0.14	0.08	0.23	0.21	27.3%	0.23	0.12	0.13	0.18	11.5%
GLOCALFAIR	0.17	0.17	0.21	0.17	9.5%	0.11	0.05	0.20	0.15	17.3%	0.14	0.16	0.11	0.15	10.1%

we conducted preliminary experiments to explore different parameter values (presented in Supplemental Material 12.3) and empirically set γ to 0.6 for all datasets and $(\tau_{\text{FNR}}, \tau_{\text{FPR}})$ to $(0.1, 0.08)$, $(0.1, 0.08)$, and $(0.06, 0.06)$ for the CelebA, Adult Income, and UTK Faces datasets, respectively, to maintain good model utility.

Baselines

We compare GLOCALFAIR with a set of state-of-the-art solutions, including a non-fairness method (FEDAVG [McMahan *et al.*, 2017]), client-fairness methods (q-FFL [Li *et al.*, 2020] and GIFAIR [Yue *et al.*, 2023]), and group-fairness methods (FEDFB [Zeng *et al.*, 2022] and FAIRFED [Ezzeldin *et al.*, 2023]).

(1) FEDAVG [McMahan *et al.*, 2017]: A common aggregation method for FL, where the client updates are weighted based on the number of the data samples in their local datasets. It does not account for any type of fairness.

(2) q-FFL [Li *et al.*, 2020]: q-FFL is one of the client-fairness-based methods, aiming to equalize the accuracies of all the clients by dynamically reweighting the aggregation, favoring the clients with poor performance.

(3) GIFAIR [Yue *et al.*, 2023]: GIFAIR aims to achieve client fairness using regularization term to penalize the spread in the loss.

(4) FEDFB [Zeng *et al.*, 2022]: FEDFB is a group fairness method, where they have fitted the concept of fair batch from centralized learning into FL by leveraging the shared group-specific positive prediction rate for each client.

(5) FAIRFED [Ezzeldin *et al.*, 2023]: FAIRFED is a group fairness method that can improve both local and group fairness. It employs FEDAVG and a fairness-based reweighting mechanism to account for the mismatch between global fairness measure and local fairness measure.

5.2 Global Group Fairness

Table 1 shows the results of global group fairness (EOD and DP-Dis on the global test datasets) on the three datasets. As FAIRFED is limited to only a single binary sensitive group, to make a fair comparison, we run the federated training twice for all three datasets (once for each sensitive group) and report

the EOD and the average DP-Dis of these training rounds. As we can see from the results, FEDAVG, q-FFL, and GIFAIR perform poorly as their objective is not group fairness. As for the state-of-the-art group fairness solutions, FAIRFED and FEDFB, we find that GLOCALFAIR can outperform these solutions in all three datasets, with a significant global group fairness enhancement. Specifically, GLOCALFAIR can improve EOD of gender and age by 44% and 41%, respectively, and improve DP-Dis by 34% compared to FEDFB in CelebA. We observe a similar fairness improvement for the Adult Income and UTK Faces datasets. By jointly considering all the fairness metrics, GLOCALFAIR achieves an overall 21% improvement compared to FEDFB and 18% compared to FAIRFED in the Adult Income dataset. For the UTK Faces dataset, GLOCALFAIR achieves 37% and 40% fairness improvements compared to FEDFB and FAIRFED, respectively.

5.3 Local Group Fairness

To compare the local group fairness, we measure the mean and standard deviation of EOD and DP-Dis on individual clients. As shown in Table 2, we observe that the proposed GLOCALFAIR is able to achieve a higher level of local group fairness compared to FAIRFED and FEDFB with lower EOD and DP-Disparity values (in terms of both mean and standard deviation). Most noticeably, in the CelebA dataset, GLOCALFAIR can achieve an average EOD of 0.15 for the gender group, which is 48% lower than FEDFB with the second lowest measured EOD. To further visualize the distribution of local fairness measures across all clients, we compare the distribution of EOD and DP-Disparity with other group fairness methods in Figure 4. Note that for a completely fair model, the measured fairness metrics of clients should be of the same value and approach 0. We observe that the result of GLOCALFAIR is more centered with lower variance, indicating better local group fairness. From the figure, we can observe that for gender groups in CelebA, almost 20 out of 100 clients have near-zero EOD. Similarly, for DP-Dis in the CelebA dataset, most of the clients have less than a 0.2 disparity. These results demonstrate that by utilizing the proposed server-client co-design framework, GLOCALFAIR can achieve greater local

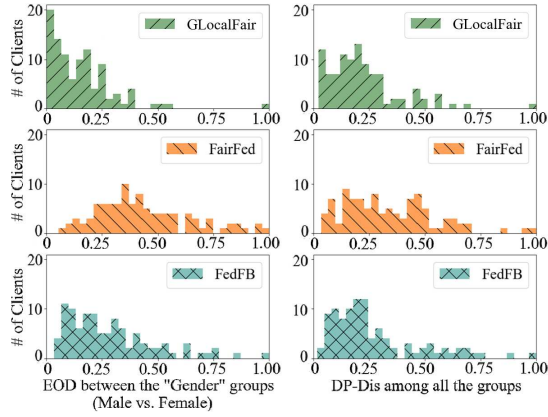


Figure 4: Distribution of EOD and DP-Dis on individual clients for the CelebA datasets (Results on other datasets can be found in Supplementary Material 12.4).

group fairness among the clients compared to other group fairness methods.

5.4 Utility and Client Fairness

From Table 1 and 2, we observe that the proposed GLOCALFAIR offers competitive and stable performance in terms of maximizing the utility and mitigating discrepancy among the clients compared to other group-fairness-based methods. FEDAVG achieves relatively higher model utility (e.g., the highest utility in the CelebA dataset) compared to other methods, which is expected as it is only optimized for utility and does not take fairness into account. Although the proposed GLOCALFAIR is not optimized for client fairness, it can still outperform other group-fairness baselines (FEDFB and FAIRFED). Specifically, as can be seen from Table 2, on the CelebA dataset, GLOCALFAIR achieves an average performance improvement of 5% and 7% in discrepancy compared to FEDFB and FAIRFED, respectively. Moreover, GLOCALFAIR achieves 18% and 15% improvement on the Adult Income dataset and 6% and 9.5% on the UTK Faces dataset. These results show that GLOCALFAIR can maintain a good level of model utility and client fairness while providing enhanced group fairness compared to other baselines.

Table 3: Performance under different FL data settings.

Metric	Method	CelebA			UTK Faces		
		α			α		
		0.1	10	100	0.1	10	100
EOD ($\downarrow$)	FEDAVG	0.50	0.45	0.32	0.48	0.41	0.40
	q-FFL	0.42	0.36	0.28	0.44	0.37	0.36
	GIFAIR	0.33	0.27	0.25	0.33	0.31	0.29
	FAIRFED	0.25	0.18	0.17	0.25	0.18	0.16
	FEDFB	0.28	0.21	0.18	0.27	0.19	0.16
	GLOCALFAIR	0.11	0.09	0.08	0.12	0.09	0.08
DP-Dis ($\downarrow$)	FEDAVG	0.50	0.47	0.30	0.35	0.31	0.29
	q-FFL	0.44	0.35	0.27	0.39	0.39	0.34
	GIFAIR	0.32	0.27	0.22	0.31	0.27	0.26
	FAIRFED	0.21	0.16	0.15	0.25	0.19	0.17
	FEDFB	0.22	0.15	0.12	0.18	0.21	0.16
	GLOCALFAIR	0.11	0.07	0.07	0.15	0.14	0.12
Utility ($\uparrow$)	FEDAVG	81.7%	83.0%	85.0%	84.2%	87.0%	88.2%
	q-FFL	79.7%	79.3%	81.2%	83.3%	86.0%	87.1%
	GIFAIR	78.7%	79.9%	81.7%	82.8%	85.9%	86.4%
	FAIRFED	76.7%	78.6%	80.1%	82.1%	85.1%	86.9%
	FEDFB	76.2%	79.1%	79.9%	81.1%	85.3%	86.0%
	GLOCALFAIR	78.5%	79.1%	80.2%	83.9%	86.4%	86.6%

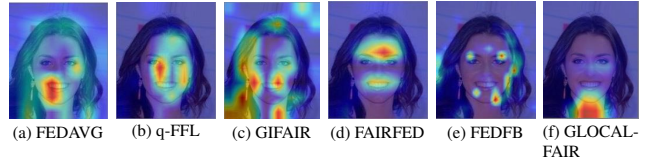


Figure 5: Comparison of feature heat maps produced by different baselines and GLOCALFAIR (More results can be found in Supplementary Material 12.7).

5.5 Impact of Different FL Data Settings

To assess the robustness of our method to data heterogeneity, we perform experiments under different data distributions, including both IID and non-IID scenarios. We distribute the dataset among n clients such that the distribution of sensitive attributes (G) is non-IID and can be controlled using a heterogeneity controller α , where $\alpha \rightarrow \infty$ corresponds to IID distributions. To create this heterogeneity, we use a power law distribution [Clusset *et al.*, 2009]. The results in Table 3 show that while data heterogeneity has a significant impact on baseline methods, our method is immune to it due to the use of a clustering algorithm that helps to mitigate variance among client updates.

5.6 Interpretability

To examine model interpretability, we consider a binary classification task (i.e., smiling or not) on the CelebA dataset and compare the heat maps of the models produced by different FL framework baselines. We tested 20 randomly selected images from the CelebA test set and generated the feature heat map using the Grad-CAM method [Selvaraju *et al.*, 2017]. A visual comparison is provided in Figure 5, in which we observe that the heat maps of models from FEDAVG, q-FFL, and GIFAIR mostly concentrate on regions of the human face (e.g., hair and eyes) that are closely related to sensitive demographic information such as gender or identity, which may introduce a bias towards certain groups. As FAIRFED and FEDFB focus on removing bias on sensitive attributes in the model, we can see from the heat maps that they only concentrate on specific regions that are highly related to smiling, such as lips. However, FAIRFED concentrates on the forehead and FEDFB concentrates on the hair. Emphasizing features from these areas can introduce a bias toward the age and gender of the subject. In the case of GLOCALFAIR, we found that the heat map concentrates on less-biased areas of the human face, such as the chin, which indicates better model fairness compared to other methods.

6 Conclusion

We propose GLOCALFAIR, a client-server co-design framework that mitigates the demographic biases while ensuring group fairness in FL at both global and local levels. We utilize constrained optimization to counter algorithmic biases on the client side and apply clustering-based aggregation on the server side for equitable global model performance among sensitive groups while preserving high utility. Our experiments across diverse datasets demonstrate that our approach achieves strong group fairness results globally and locally while also preserving utility similar to state-of-the-art fair FL frameworks.

References

- [Abay *et al.*, 2020] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.
- [Caldas *et al.*, 2018] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.
- [Chen and Chao, 2021] Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. *International Conference on Learning Representations (ICLR)*, 2021.
- [Chu *et al.*, 2021] Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021.
- [Clauset *et al.*, 2009] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
- [Cui *et al.*, 2021] 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, 2021.
- [Du *et al.*, 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)*, pages 181–189. SIAM, 2021.
- [Dua and Graff, 2017] Dheeru Dua and Casey Graff. Uci machine learning repository. <https://archive.ics.uci.edu/ml/datasets/adult>, 2017. [Online; accessed 17-December-2021].
- [Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
- [Ezzeldin *et al.*, 2023] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7494–7502, 2023.
- [Gini, 1912] Corrado Gini. Variabilità e mutabilità. reprinted in e. pizetti & t. salvemini. *Memorie di metodologica statistica*, 1912.
- [Granqvist *et al.*, 2020] Filip Granqvist, Matt Seigel, Rogier van Dalen, Áine Cahill, Stephen Shum, and Matthias Paulik. Improving On-Device Speaker Verification Using Federated Learning with Privacy. In *Proc. Interspeech 2020*, pages 4328–4332, 2020.
- [Grgić-Hlača *et al.*, 2018] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [Hard *et al.*, 2018] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.
- [Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [Kairouz *et al.*, 2021] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.
- [Li *et al.*, 2020] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *International Conference on Learning Representations (ICLR)*, 2020.
- [Li *et al.*, 2021] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.
- [Liu *et al.*, 2018] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.
- [Lohia *et al.*, 2019] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 IEEE international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.
- [Lyu *et al.*, 2020] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. In *Federated Learning*, pages 189–204. Springer, 2020.
- [McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [Mohri *et al.*, 2019] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.
- [Rieke *et al.*, 2020] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

- [Rodríguez-Gálvez *et al.*, 2021] Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS Workshop*, 2021.
- [Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [Xu *et al.*, 2021] 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.
- [Yue *et al.*, 2023] Xubo Yue, Maher Nouiehed, and Raed Al Kontar. Gifair-fl: A framework for group and individual fairness in federated learning. *INFORMS Journal on Data Science*, 2(1):10–23, 2023.
- [Zeng *et al.*, 2022] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2022.
- [Zhang *et al.*, 2017] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.
- [Zhang *et al.*, 2018] 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*, pages 335–340, 2018.
- [Zhang *et al.*, 2020a] 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)*, pages 1051–1060. IEEE, 2020.
- [Zhang *et al.*, 2020b] Jingfeng Zhang, Cheng Li, Antonio Robles-Kelly, and Mohan Kankanhalli. Hierarchically fair federated learning. *arXiv preprint arXiv:2004.10386*, 2020.

7 GLOCALFAIR Algorithm

The detailed steps performed on the server and clients in GLOCALFAIR are summarized in Algorithm 1.

Algorithm 1 GLOCALFAIR Algorithm

Server:

Initialize global model parameter θ_{global}^0

for each round $t = 0, 1, \dots$ **do**

 Server randomly selects a subset of n clients

 Server sends the latest global model θ_{global}^t to the selected clients

for each client $k = 0, 1, \dots, n$ **do**

 Gather local update (θ_k^t)

 Cluster the updates upon their Gini coefficients ($\mathcal{G}$)

$\min \sum_{i=1}^p \sum_{j=1}^{n_i} \|\mathcal{G}_{ij} - \frac{1}{n_i} \sum_{j=1}^{n_i} \mathcal{G}_{ij}\|^2$

end for

 Update the global model parameter

$\theta_{global}^{t+1} = \theta_{global}^t + \sum_{i=1}^p \exp(-\gamma \frac{1}{n_i} \sum_{j=1}^{n_i} \mathcal{G}_{ij})$

$\sum_{j=1}^{n_i} \frac{d_{ij}}{\sum_{k=1}^{n_i} d_{ik}} \theta_{ij}^t$

end for

Client (the k -th client):

Receive global model parameter θ_{global}^t from the server at the round t ;

Train the local model on the local dataset to obtain the updated model parameter θ_k^t

Initialize $\theta_k^{t,0} = \theta_k^t$ for local constraint optimization

for $j = 0, 1, \dots, J-1$ **do**

 Let $\tilde{\Delta}_\theta^j$ be stochastic gradient of Equation 6

 Let $\tilde{\Delta}_\lambda^j$ be stochastic subgradient of Equation 6

 Update $\theta_k^{t,j+1} = \Pi_\Theta(\theta_k^{t,j} - \eta_\theta \tilde{\Delta}_\theta^j)$

 Update $\lambda^{(j+1)} = \Pi_\Lambda(\lambda^j + \eta_\lambda \tilde{\Delta}_\lambda^j)$

end for

Update θ_k^t to be the best local update $\theta_k^{t,j}$, $j \in [0, \dots, J-1]$, with the lowest FPR/FNR

Send θ_k^t to the server

8 Convergence Rates

Theorem 1. Define $\mathcal{M}$ as the set of all stochastic $(m+1) * (m+1)$ matrices, $\Lambda := \Delta^{m+1}$ as the $(m+1)$ dimensional simplex, where m is the number of functional constraints, and assume that each $\hat{g}_i$ upper bounds the corresponding g_i (g_i represents the i^{th} constraint). The max-min formulation of Equation 6 in the paper uses a non-zero-sum two-player game where the λ -player chooses how much the θ -player should penalize the (differentiable) sigmoid relaxed constraints but does so in such a way as to satisfy the original constraints. Let simplify the Equation 6 from the main paper as:

$$\max_{\lambda_i \geq 0} \min_{\theta \in \Theta} f_L(\theta) + \sum_{i=1}^m \lambda_i g_i(\theta), \quad (13)$$

where $f_L(\theta)$ is the binary cross-entropy loss that we want to minimize. Let denote the objective of two players (λ and θ)

as:

$$\begin{aligned}\mathcal{L}_\theta(\theta, \lambda) &= f_L(\theta) + \sum_{i=1}^m \lambda_i g_i(\theta), \\ \mathcal{L}_\lambda(\theta, \lambda) &= \sum_{i=1}^m \lambda_i g_i(\theta),\end{aligned}\quad (14)$$

where $\theta \in \Theta$ and $\lambda \in \Lambda$ jointly distributed random variables such that:

$$\begin{aligned}\mathbb{E}_{\theta, \lambda}[\mathcal{L}_\theta(\theta, \lambda)] - \inf_{\theta^* \in \Theta} \mathbb{E}_\lambda[\mathcal{L}_\theta(\theta^*, \lambda)] &\leq \epsilon_\theta, \\ \max_{M^* \in \mathcal{M}} \mathbb{E}_{\theta, \lambda}[\mathcal{L}_\lambda(\theta, M^* \lambda)] - \mathbb{E}_{\theta, \lambda}[\mathcal{L}_\lambda(\theta, \lambda)] &\leq \epsilon_\lambda.\end{aligned}\quad (15)$$

Suppose $\bar{\theta}$ is a random variable for which $\bar{\theta} = \theta^t$ with probability $\lambda_1^t / \sum_{s=1}^T \lambda_1^s$ and let $\bar{\lambda} := (\sum_{t=1}^T \lambda^t) / T$, then $\bar{\theta}$ represents a sub-optimal solution when:

$$\mathbb{E}_{\bar{\theta}}[f_L(\bar{\theta})] \leq \inf_{\theta \in \Theta: \forall i, \hat{g}_i(\theta^*) \leq 0} f_L(\theta^*) + \frac{\epsilon_\theta + \epsilon_\lambda}{\bar{\lambda}_1}, \quad (16)$$

and it is a nearly-feasible solution when:

$$\max_{i \in [m]} \mathbb{E}_{\bar{\theta}}[g_i(\bar{\theta})] \leq \frac{\epsilon_\lambda}{\bar{\lambda}_1}. \quad (17)$$

Note that the optimality inequality is weaker than it may appear since the comparator in this equation is not the optimal solution w.r.t. the constraints g_i , but rather w.r.t. the sigmoid constraints $\hat{g}_i$.

Proof. Optimality: If we choose M^* to be a matrix with its first row being all-one and all other rows being all-zero, then $\mathcal{L}(\theta, M^* \lambda) = 0$, which shows that the first term in the LHS of the second line of Equation 15 is non-negative. Hence $-\mathbb{E}_{\theta, \lambda}[\mathcal{L}_\lambda(\theta^t, \lambda^t)] \leq \epsilon_\lambda$, and thus by the definition of $\mathcal{L}_\lambda$ and the fact that $\hat{g}_i \geq g_i$, we have:

$$\mathbb{E}_{\theta, \lambda} \left(\sum_{i=1}^m \lambda_{i+1} \hat{g}_i(\theta) \right) \geq -\epsilon_\lambda. \quad (18)$$

Here, $\mathcal{L}_\theta$ is linear in λ , so the first line of the Equation 15, combined with the above results and the definition of $\mathcal{L}_\theta$, becomes:

$$\mathbb{E}_{\theta, \lambda}[\lambda_1 f_L(\theta)] - \inf_{\theta^* \in \Theta} \left(\bar{\lambda}_1 f_L(\theta^*) + \sum_{i=1}^m \lambda_{i+1} \hat{g}_i(\theta^*) \right) \leq \epsilon_\theta + \epsilon_\lambda. \quad (19)$$

Suppose θ^* is the optimal solution that satisfies the sigmoid constraints $\hat{g}$, so that $\hat{g}_i(\theta^*) \leq 0$ for all $i \in [m]$, then we can have:

$$\mathbb{E}_{\theta, \lambda}[\lambda_1 f_L(\theta)] - \bar{\lambda}_1 f_L(\theta^*) \leq \epsilon_\theta + \epsilon_\lambda, \quad (20)$$

which is the optimality claim.

Feasibility: To simplify the notation, suppose $l_1(\theta) := 0$ and $l_{i+1}(\theta) := g_i(\theta)$ for $i \in [m]$, then $\mathcal{L}_\lambda(\theta, \lambda) = \langle \lambda, l(\theta) \rangle$. Consider the first term in LHS of the second line of Equation 15:

$$\begin{aligned}\max_{M^* \in \mathcal{M}} \mathbb{E}_{\theta, \lambda}[\mathcal{L}_\lambda(\theta, M^* \lambda)] &= \max_{M^* \in \mathcal{M}} \mathbb{E}_{\theta, \lambda}[\langle M^* \lambda, l(\theta) \rangle], \\ &= \max_{M^* \in \mathcal{M}} \mathbb{E}_{\theta, \lambda} \left[\sum_{i=1}^{m+1} \sum_{j=1}^{m+1} M_{j,i}^* \lambda_i l_j(\theta) \right], \\ &= \sum_{i=1}^{m+1} \max_{j \in [m+1]} \mathbb{E}_{\theta, \lambda}[\lambda_i l_j(\theta)],\end{aligned}\quad (21)$$

where we used the fact that since M^* is left-stochastic, each of its columns is an $(m+1)$ -dimensional multinoulli distribution. For the second term in the LHS of the second line of Equation 15, we can use the fact that $l_1(\theta) = 0$:

$$\mathbb{E}_{\theta, \lambda} \left[\sum_{i=2}^{m+1} \lambda_i l_i(\theta) \right] \leq \sum_{i=2}^{m+1} \max_{j \in [m+1]} \mathbb{E}_{\theta, \lambda}[\lambda_i l_j(\theta)]. \quad (22)$$

Plugging these two results into the second line of Equation 15, the two sums collapse, leaving:

$$\max_{i \in [m+1]} \mathbb{E}_{\theta, \lambda}[\lambda_1 l_i(\theta)] \leq \epsilon_\lambda. \quad (23)$$

Given the definition of l_i , we can yield the feasibility claim. $\square$

Corollary 1. Let $f_1, f_2, \dots : \Theta \rightarrow \mathbb{R}$ be a sequence of convex functions that we wish to minimize on a compact convex set Θ . Let the step size $\eta = \frac{B_\Theta}{B_\Delta \sqrt{2T}}$ where $B_\Theta \geq \max_{\theta \in \Theta} \|\theta\|_2$

and $B_\Delta \geq \|\hat{\Delta}^t\|_2$ is a uniform upper bound on the norms of stochastic subgradients. Suppose that we perform T iterations of the following stochastic update starting from $\theta^1 = \arg \min_{\theta \in \Theta} \|\theta\|_2$:

$$\theta^{t+1} = \Pi_\Theta \left(\theta^t - \eta \hat{\Delta}^t \right), \quad (24)$$

where $\mathbb{E}[\hat{\Delta}^t | \theta^t] \in_t(\theta^t)$, i.e. $\hat{\Delta}^t$ is a stochastic subgradient of f_t at θ^t and Π_Θ projects its arguments onto Θ w.r.t. the Euclidean norm. Then with probability $1 - \delta$ over the draws of the stochastic subgradients:

$$\frac{1}{T} \sum_{t=1}^T f_t(\theta^t) - \frac{1}{T} \sum_{t=1}^T f_t(\theta^*) \leq 2B_\Theta B_\Delta \sqrt{\frac{1 + 16\ln \frac{1}{\delta}}{T}}, \quad (25)$$

where $\theta^* \in \Theta$ is an arbitrary reference vector.

Lemma 1. Suppose that Θ is a compact convex set and the objective, and sigmoid constraint functions $\hat{g}_1, \dots, \hat{g}_m$ are convex (but not $g_1, \dots, g_m$). The three upper bounds are $B_\Theta \geq \max_{\theta \in \Theta} \|\theta\|_2$, $B_\Delta \geq \max_{t \in T} \|\hat{\Delta}^t\|_2$ and $B_\Delta \geq \max_{t \in T} \|\Delta_\lambda^t\|_\infty$. If we run the client side optimization with step size $\eta_\theta := B_\Theta / B_\Delta \sqrt{2T}$ and $\eta_\lambda := \sqrt{(m+1)\ln(m+1)/TB_\Delta^2}$, then the results would satisfy the conditions of **Theorem 1** for:

$$\begin{aligned}\epsilon_\theta &= 2B_\Theta B_\Delta \sqrt{\frac{1 + 16\ln \frac{2}{\delta}}{T}}, \\ \epsilon_\lambda &= 2B_\Delta \sqrt{\frac{2(m+1)\ln(m+1)(1 + 16\ln \frac{2}{\delta})}{T}},\end{aligned}\quad (26)$$

with probability $1 - \delta$ over the draws of stochastic subgradients.

Proof. By applying **Corollary 1** to the optimization over θ and with probability $1 - 2\delta'$ over the stochastic subgradients, we can get:

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathcal{L}_\theta(\theta^t, \lambda^t) - \frac{1}{T} \sum_{t=1}^T \mathcal{L}_\theta(\theta^*, \lambda^t) &\leq \\ 2B_\Theta B_\Delta \sqrt{\frac{1 + 16\ln \frac{1}{\delta'}}{T}}, \\ \frac{1}{T} \sum_{t=1}^T \mathcal{L}_\lambda(\theta^t, M^* \lambda^t) - \frac{1}{T} \sum_{t=1}^T \mathcal{L}_\lambda(\theta^*, \lambda^t) &\leq \\ 2B_\Delta \sqrt{\frac{2(m+1)\ln(m+1)(1 + 16\ln \frac{2}{\delta})}{T}}. \end{aligned} \quad (27)$$

By taking $\delta = 2\delta'$ and using the definitions of $\bar{\theta}$ and $\bar{\lambda}$, we can yield the claimed result. $\square$

9 Federated Datasets

We evaluate the proposed GLOCALFAIR using the following three datasets:

(1) *CelebA* [Liu *et al.*, 2018]: A collection of 200k celebrity face images from the Internet that have been manually annotated. The dataset has up to 40 labels, each of which is binary-valued. We resize each image to 128x128px and create a classification task to classify whether the celebrity in the image is smiling or not. For CelebA, each subject’s gender (male or female) and age (young or adult) are considered sensitive attributes. We sort the images by the celebrity ID according to Caldas *et al.* [Caldas *et al.*, 2018]. This results in a dataset with a total number of 200,288 images from 9,343 celebrities. The dataset is then randomly partitioned for training and testing with no overlapping celebrity ID, each containing 180,429 images from 8,408 celebrities and 19,859 images from 935 celebrities, respectively. Finally, the training dataset is randomly distributed into 100 silos with no overlapping celebrity ID to emulate the cross-silo FL setup. In addition, in each silo, we randomly split the image data in a 70-10-20 manner, where 70% of the data are used for training, 10% are used for validating constrained optimization, and 20% are reserved for testing client fairness.

(2) *Adult Income* [Dua and Graff, 2017]: A tabular dataset that is widely investigated in machine learning fairness literature. It contains 48,842 samples with 14 attributes. We create a binary classifier to classify whether a subject’s income is above 50k or not. In this dataset, race (white or non-white) and gender (male or female) are used as the sensitive attributes. We keep 10% of the remaining dataset for testing the global model. We distribute the rest of the dataset among four clients where each client holds one of the four possible combinations (i.e., white male, nonwhite male, white female, and non-white female) of two sensitive groups. Clients use the same 70-10-20 data split as the previous dataset.

(3) *UTK Faces* [Zhang *et al.*, 2017]: A large-scale face dataset with more than 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variations in pose, facial expression, illumination, occlusion, resolution, etc. We resize each image to 128x128px and create a binary classification task of classifying young (less than 40 years) and old (older than 40 years). Here, race (white or non-white) and gender (male or female) are used as the sensitive attributes. We divide the dataset into 90% training and 10% testing data. We randomly distributed the training dataset to 20 clients where each client’s data quantity has a mean of 900 and a variance of 250.

10 Model Description

We evaluate GLOCALFAIR on three datasets, namely, CelebA, Adult Income, and UTK Faces. The details of the FL model used for each dataset are presented below:

- (1) *CelebA*: We train a Resnet34 [He *et al.*, 2016] model with a single fully connected layer, equipped with a sigmoid activation function for binary classification (i.e., classifying whether the celebrity in the image is smiling or not). The input layer is modified to accommodate the image dimensions of 128x128 pixels, characteristic of the CelebA dataset to reduce the computation cost at the clients. We use Adam optimizer with a learning rate of 0.01 with 10% decay rate.
- (2) *Adult Income*: We use a Multi-layer Perceptron (MLP) model containing two hidden layers, each with 32 and 16 units, respectively, and a single unit output layer for binary classification (i.e., classifying whether a subject’s income is above 50k or not). We use SGD with a learning rate of 0.01 and momentum of 0.9.
- (3) *UTK Faces*: We train a MobileNetV2 [Sandler *et al.*, 2018] model with an input size of 128x128. We use Adam optimizer with a learning rate of 0.05 with 10% decay rate.

11 Federated Learning Settings

In our evaluation, we adjust the FL setup in each dataset to accommodate data from different genres. For datasets with fewer clients, we choose to use a higher client participation rate.

- (1) *CelebA*: We use 100 clients to train the model for 150 communication rounds with a client participation rate of 0.3 and a γ value of 0.6.
- (2) *Adult Income*: We use 4 clients to train the model for 30 communication rounds with a client participation rate of 0.75 and a γ value of 0.6.
- (3) *UTK Faces*: We use 20 clients to train the model for 150 communication rounds with a client participation rate of 0.4 and a γ value of 0.6.

12 Additional Evaluation

12.1 Group Fairness vs. Gini Coefficients

We conduct experiments on the CelebA dataset to investigate the relationship between the fairness of client model updates

Table 4: Global and Local group fairness Evaluation for Dual-fold Fairness Strategy

Method	Global Fairness			Local Fairness				
	EOD ($\downarrow$) Gender	DP-Dis ($\downarrow$)	Utility ($\uparrow$)	EOD (Gender Group)		DP-Dis		DIS ($\downarrow$)
				Mean($\downarrow$)	Std($\downarrow$)	Mean($\downarrow$)	Std($\downarrow$)	
Local Fairness Constraint	0.15	0.16	81.6%	0.21	0.22	0.22	0.20	13.1%
Fairness-aware Clustering	0.13	0.12	77.9%	0.25	0.21	0.27	0.23	8.9%
GLOCALFAIR	0.09	0.10	79.2%	0.17	0.17	0.21	0.17	9.5%

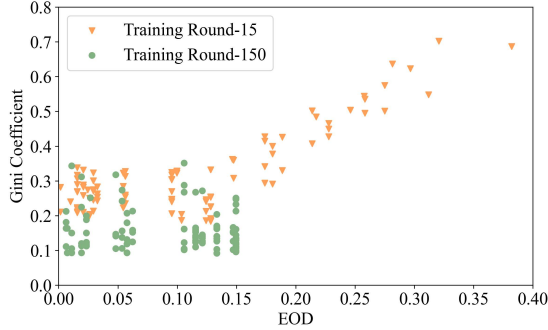


Figure 6: Relationship between Model Fairness and Gini Coefficient Calculated on Model Weights

and their Gini Coefficients, which are calculated to measure the distribution inequality of their model weights and are used as a proxy of group fairness in this paper. To empirically show the relationship between group fairness and Gini Coefficients of model updates, we analyze and visualize the Gini Coefficient values for each client’s update during the 15th training round and the final training round (150th) in relation to their EOD values (i.e., one of the group fairness metrics we use in the paper) of each client’s model in Figure 6. Our observations reveal a strong correlation between the Gini Coefficient and EOD values. Specifically, clients whose models exhibited higher EOD values also displayed higher Gini Coefficients. This alignment suggests that biased models are typically associated with unevenly distributed model weights under our proposed local fairness constraints. Conversely, clients whose models boasted lower EOD values showcased lower Gini Coefficients, indicating a more equitable distribution of model weights. As the fairness of the models improved at the end of the training, as evidenced by lower EOD values, the Gini coefficients of these models also reduced.

12.2 Benefits of Two Fold Fairness Optimization in GLOCALFAIR

We embrace a dual-fold strategy for fairness optimization. This strategy encompasses both local and global dimensions of fairness. The first facet of our approach focuses on attaining local fairness by integrating a local fairness constraint optimization mechanism at the client level. Complementing this, the second facet of our approach addresses global fairness through a distinctive methodology that employs fairness-aware clustering-based aggregation on the server side. This server-level operation aspires to harmonize the contributions from diverse clients, ensuring that the aggregation process mitigates the amplification of any inherent disparities. To rigorously assess the efficacy of these individual fairness en-

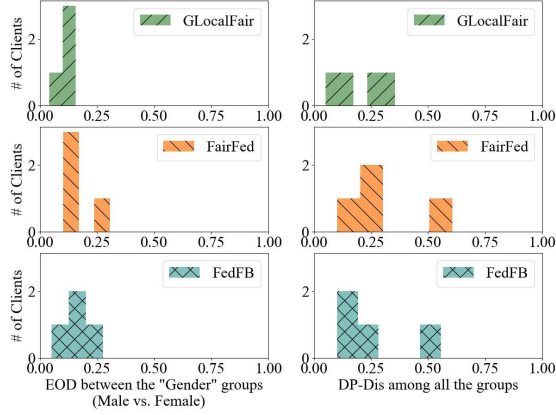
Table 5: Impact of the fairness constraint τ_{FPR} and τ_{FNR} .

False Positive Rate Limit (τ_{FPR})	False Negative Rate Limit (τ_{FNR})	EOD ($\downarrow$)		DP-Dis ($\downarrow$)	Utility ($\uparrow$)
		Gender	Age		
0.08	0.10	0.09	0.10	0.10	79.1%
0.10	0.10	0.13	0.13	0.15	78.2%
0.15	0.10	0.16	0.21	0.16	74.9%
0.20	0.10	0.18	0.19	0.18	74.3%
0.08	0.08	0.55	0.45	0.55	67.3%
0.08	0.12	0.10	0.11	0.13	76.9%
0.10	0.12	0.11	0.13	0.15	77.3%
0.10	0.15	0.15	0.15	0.18	75.1%

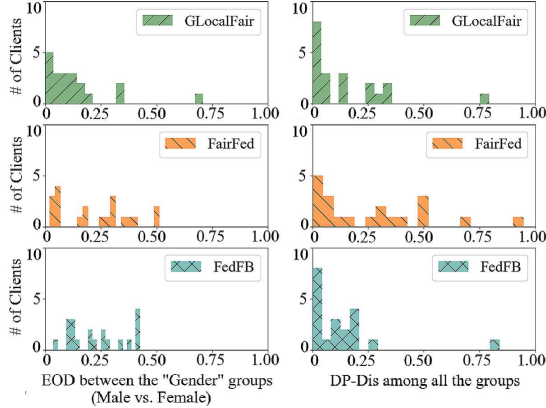
hancement methods, we conducted an additional experiment using the CelebA dataset. Specifically, the experiment aims to evaluate the unique contributions of each strategy (client-level or server-level) to the overall fairness improvement. We use FEDAVG on the server for aggregation when we use local fairness constraint optimization only. In Table 4, we show the results of local fairness constraint optimization and the fairness-aware clustering-based aggregation individually. As we can observe from the table, the global fairness measured on the global model is lower when we employ only the local fairness constraints on the client side, but it gets improved when fairness-aware clustering is used, showing the local fairness improvement does not fully improve the global fairness. Similarly, when the local fairness constraints are used, the local fairness is improved, but the discrepancy among the client’s performance is degraded due to the use of FEDAVG on the server side. Local fairness is degraded when only the fairness-aware clustering is used on the server side, but the discrepancy among the client’s performance is improved. When we combine both the fairness strategies in GLOCALFAIR, they complement each other and improve both local and global group fairness.

12.3 Impact of Client Fairness Constraint

At each client, we use constraint optimization on top of the learning algorithm to enforce local fairness. In each round of training, we constrain the false positive rate (FPR) and false negative rate (FNR) and use a misclassification penalty for constraint violations. We evaluate GLOCALFAIR with different τ_{FPR} values (i.e., 0.08, 0.10, 0.15, and 0.20) and a fixed τ_{FNR} value (i.e., 0.10) on the CelebA dataset in Table 5. We observe that it achieves the best performance when $\tau_{FPR} = 0.08$. Although further decreasing τ_{FPR} may seemingly increase performance, in our experiments, we find that any τ_{FPR} values smaller than 0.08 would result in a trivial model that outputs the same prediction for all data samples.



(a) Adult Income



(b) UTK Faces

Figure 7: Distribution of EOD and DP-Dis on individual clients for the Adult Income and UTK Faces datasets.

On the other hand, if we increase τ_{FPR} to be larger than 0.15, there would be almost no effect of constraint as the model will converge with an FPR value lower than the limit. We also test a few combinations with varying τ_{FNR} . We find that with τ_{FPR} fixed, using a smaller τ_{FNR} value can help improve model fairness. However, if we set both τ_{FNR} and τ_{FPR} to 0.08, it over-constrains the model and makes it unstable. Also, any value larger than 0.15 for τ_{FNR} has almost no effect on the model performance.

12.4 Local Group Fairness for the Adult Income and UTK Faces Datasets

Figure 7 compares the local fairness (i.e., distribution of EOD and DP-Dis among individual clients) of GLOCALFAIR with FEDFB and FAIRFED on the Adult Income and UTK Faces dataset. Note that for a completely fair model, the measured fairness metrics of clients should be of the same value and approach 0. We observe that the result of GLOCALFAIR is more centered with lower variance, indicating better local group fairness. Especially for the UTK Faces gender groups (Figure 7b), almost 17 out of 20 clients have near-zero EOD. Similar results can be observed for DP-Dis in the UTK Faces dataset as most of the clients have less than 0.3 disparity. For the Adult Income dataset, we also observe that GLOCALFAIR has the

Table 6: Global group fairness vs. model utility with different fairness level γ .

Fairness Indicator (γ)	EOD ($\downarrow$)		DP-Dis ($\downarrow$)	Utility ($\uparrow$)
	Gender	Age		
0.2	0.41	0.43	0.36	82.5%
0.4	0.23	0.25	0.15	80.6%
0.6	0.09	0.10	0.10	79.1%
0.8	0.09	0.11	0.10	78.9%

Table 7: Customizable fairness configuration on client models.

τ_{FPR}	τ_{FNR}	Utility ($\uparrow$)	Precision	Recall
0.06	-	80.9%	0.95	0.75
0.08	-	78.9%	0.94	0.75
-	0.06	78.6%	0.87	0.90
-	0.08	80.5%	0.87	0.88

highest number of clients that achieve the lowest EOD and DP-Dis. These results demonstrate that by utilizing the proposed server-client co-design framework, GLOCALFAIR can achieve greater local group fairness among the clients compared to other group fairness methods.

12.5 Fairness vs. Utility Trade-off

In the design of GLOCALFAIR, we use γ to control the balance between fairness and utility of the global model. A larger γ value would result in a fairer model with lower utility. We evaluate GLOCALFAIR with different γ values on the CelebA dataset and summarize the results in Table 6. We can observe that as γ increases, the utility of the model decreases while the fairness performance saturates at $\gamma = 0.6$. Therefore, we set $\gamma = 0.6$ in our major experiments, which ensures the model’s fairness while still maintaining a good level of utility.

12.6 Customizable Fairness Configuration on Client Models

The developed GLOCALFAIR offers customizable fairness configurations on the client side, aiming to cater to diverse fairness requirements. To evaluate this feature, we intentionally alternate between fixing the FPR and FNR while relaxing the other fairness constraint, mirroring real-world scenarios where a preference for lower FPR or FNR varies. As shown in Table 7, across all scenarios, we observe consistently high utility scores. Constraining FPR results in high precision values, around 0.95, indicating accurate positive predictions. However, this often leads to a lower recall due to the unconstrained FNR, resulting in a higher number of false negatives. Conversely, when FNR is constrained, recall is notably high, signifying comprehensive positive predictions. Nevertheless, in these instances, precision tends to be lower due to the unconstrained FPR, which results in a higher number of false positives. These findings underscore the importance and effectiveness of customizable fairness configurations in GLOCALFAIR, as they enable the adaptation of model behavior to varying fairness requirements and real-world scenarios.

12.7 Additional Interpretability Results

In Figure 8, we extend the interpretability analysis of each baseline method using the CelebA dataset. Specifically, we show three examples from both classes (smiling and not smiling). From the visualization results, it is clear that GLOCAL-

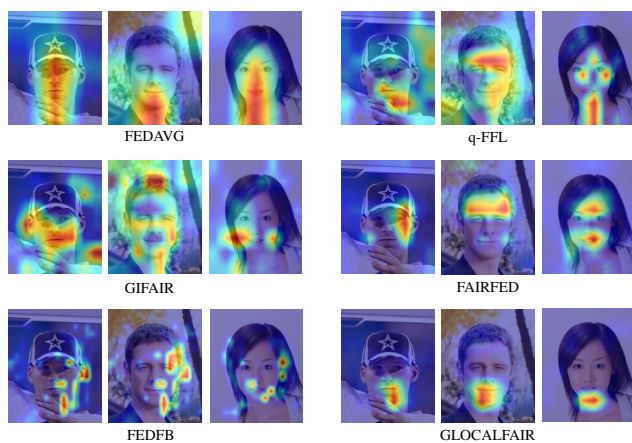


Figure 8: Comparison of feature heat maps produced by different baselines and GLOCALFAIR.

FAIR focuses on the areas that contain less sensitive attributes, which makes it fairer over the demographic groups compared to other baselines.