FairFed: Enabling Group Fairness in Federated Learning

Yahya H. Ezzeldin*
University of Southern California
yessa@usc.edu

Shen Yan*
University of Southern California
shenyan@usc.edu

Chaoyang He
University of Southern California
chaoyang.he@usc.edu

Emilio Ferrara
University of Southern California
emiliofe@usc.edu

Salman Avestimehr
University of Southern California
avestimehr@ee.usc.edu

Abstract

As machine learning becomes increasingly incorporated in crucial decision-making scenarios such as healthcare, recruitment, and loan assessment, there have been increasing concerns about the privacy and fairness of such systems. Federated learning has been viewed as a promising solution for collaboratively learning machine learning models among multiple parties while maintaining the privacy of their local data. However, federated learning also poses new challenges in mitigating the potential bias against certain populations (e.g., demographic groups), which typically requires centralized access to the sensitive information (e.g., race, gender) of each data point. Motivated by the importance and challenges of group fairness in federated learning, in this work, we propose FairFed, a novel algorithm to enhance group fairness via a fairness-aware aggregation method, aiming to provide fair model performance across different sensitive groups (e.g., racial, gender groups) while maintaining high utility. The formulation can potentially provide more flexibility in the customized local debiasing strategies for each client. When running federated training on two widely investigated fairness datasets, Adult and COMPAS, our proposed method outperforms the state-of-the-art fair federated learning frameworks under a high heterogeneous sensitive attribute distribution.

1 Introduction

Federated learning (FL) has received significant of attention for its ability to train large-scale models in a decentralized manner without requiring direct access to user data [12, 24]. It also has been increasingly applied to facilitate decision-making in various crucial areas, such as healthcare, recruitment, loan grading, etc. Nevertheless, there are concerns regarding biased performance against certain populations in such ML-assisted decision-making systems [21, 3]. Despite the increasing attention on FL fairness, most existing studies [18, 16] focus on equalizing the performance across different participating devices/silos. Few works have discussed the *group fairness* [6] related to sensitive attributes (e.g., gender, race), which is a crucial requirement for responsible AI systems aiming to ensure the model treats the groups defined by sensitive attributes equally.

*co-first authors

1st NeurIPS Workshop on New Frontiers in Federated Learning (NFFL 2021), Virtual Meeting.

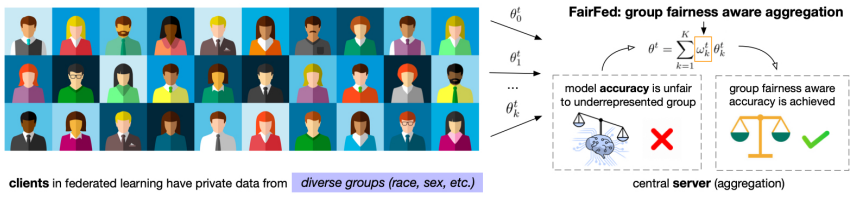

Figure 1: FairFed: Group fairness-aware federated learning framework.

Some recent studies [1] attempt to adopt the fair ML strategies in centralized learning to FL systems. However, most centralized fair ML methods require access to sensitive attributes; thus they can only be applied locally on each client in order to preserve the privacy of clients in FL. When the data distribution across clients exhibits high heterogeneity (e.g., different sensitive attributes distribution, such as when one client only has data regarding males while another client only has data regarding females), local debiasing cannot guarantee the fair performance on the overall population.

Driven by the importance and challenges of group fairness in FL, in this work, we propose a strategy to enhance group fairness via a fairness-aware aggregation method named *FairFed* (Figure 1). Our FairFed framework, introduced in Section 4, adaptively modifies aggregation weights at the server in each round. The weights are based on the mismatch between the global fairness metric (at the server) and the local fairness metric at each client, favoring clients whose local metrics match the global metric. The structure of FairFed gives it the following advantages over existing fair FL strategies:
• **Enhance group fairness under data heterogeneity:** One of the biggest challenges to FL group fairness is the heterogeneous distribution across different clients, which limits the impact of local debiasing efforts on the global data distribution. FairFed shows significant improvement in fairness performance under highly heterogeneous distribution settings, outperforming state-of-the-art methods for fairness in FL, indicating promising implications in real-life applications.
• **More freedom of customized modeling strategies on different clients:** As FairFed works as the aggregation method, it can potentially be more flexible to each client's modeling strategy (we expand on this notion in Section 4). For example, different clients can adopt different local debiasing methods based on the properties of their devices and data partitions.

2 Related work

In the classical centralized machine learning, substantial advancement has been made in group fairness in three categories: pre-processing [8, 7], in-processing [13, 25] and post-processing [17, 14] techniques. However, a majority of these techniques require access to the sensitive information (e.g., race, gender) of each data point, making it unsuitable for federated learning systems.

Fairness in federated learning. Federated learning can introduce new sources of bias through the collaborative learning process. Various definitions of fairness have been proposed to quantify such challenges in FL settings, such as *Collaborative fairness* and *agent-based fairness*. *Collaborative fairness* [18] is defined as rewarding a high-contribution participant with a better performing local model than is given to a low-contribution participant. *Client-based fairness* aims to equalize the model performance across different clients. Existing studies have focused on *uniforming the accuracy distribution over all clients* [16] and *maximizing the performance of the worst client* [20]. Due to the potential cross-device or cross-silo heterogeneity, these methods cannot prevent the discrimination against certain sensitive groups. For example, if the local models of all agents have the similar accuracy performance while being biased against the under-privileged group, then the system will satisfy the *client-based fairness* but will still display discrimination against certain sensitive groups.

Group fairness in federated learning Several recent works have made some progress on group fair FL. One research direction is to design an optimization objective with fairness constraints [26, 4], which requires each client to share the statistics of sensitive information to the server. Abay et al. also [1] investigated the effectiveness of adopting centralized debiasing mechanisms on each client. Local debiasing strategies are ineffective under the scenarios in which different clients have the heterogeneous sensitive attribute distribution, which limits its application in real-life scenarios. Compared to existing works, our proposed method does not restrict the local debiasing strategy of each participating client, thus increasing the flexibility of the system. Empirical evaluations also show that our method can yield more fairness improvements under higher data heterogeneity.

3 Preliminaries

In this section, we start by reviewing one of the most commonly used aggregation methods in FL – FedAvg [19]. We then introduce the definitions and metrics of group fairness and extend them to federated learning scenarios by defining the notions of global and local fairness.

3.1 Federated Averaging (FedAvg)

In FL, multiple clients collaborate to find a parameter θ that minimizes a weighted average of the loss across all clients. In particular:

$$\min_{\theta} f(\theta) = \sum_{k=1}^{K} \omega_k L_k(\theta), \tag{1}$$

where: K is the total number of clients; $L_k(\theta)$ denotes the local objective at client k; $\omega_k \geq 0$, and $\sum \omega_k = 1$. The local objective L_k's can be defined by empirical risks over the local dataset $\mathcal{D}_k$ of size n_k at client k, i.e., $f_k(\theta) = \frac{1}{n_k} \sum_{(\mathbf{x},y)\in\mathcal{D}_k} \ell(\theta, \mathbf{x}, y)$.

To minimize the objective in (1), the federated averaging algorithm FedAvg, proposed in [19], samples a subset of the Kclients per round to perform local training of the global model on their local datasets. The model updates are then averaged at the server weighted by the size of their respective datasets.

The FedAvg algorithm and subsequent improvements (e.g., FedOPT [22], FedNova [23]) allow for collaborative training of a high-performance global model while maintaining the privacy of the local dataset of each client. However, this collaborative training can result in a global model that is unfair towards an underlying demographic group of datapoints (similar to biases incurred in centralized training of machine learning models [6]). We highlight different notions of group fairness of a model in the following subsection.

3.2 Notions of group fairness

In sensitive machine learning applications, a data sample often contains private and sensitive demographic information that can lead to discriminatory. In particular, we assume that each data point is associated with a sensitive binary attribute A, such as gender or race. For a model with a binary output $\hat{Y}(\theta, \mathbf{x})$, the fairness is evaluated with respect to how it performs compared to the underlying groups defined by sensitive attribute A. We use $A = 1$ to represent the privileged group (e.g., male), while $A = 0$ is used to represent the under-privileged group (e.g., female). For the binary model output $\hat{Y}$ (and similarly the label Y), $\hat{Y} = 1$ is assumed to be the positive outcome. We now discuss how to evaluate the fairness of the output $\hat{Y}$ of the global model based on two group fairness definitions that are applied in centralized learning settings:

Definition 1 (Equal Opportunity [9]) *: Equal opportunity measures a binary predictor $\hat{Y}$ with respect to A and Y. The predictor is considered fair from the equal opportunity perspective if the true positive rate is independent of the sensitive attribute A (i.e., $\Pr(\hat{Y} = 1|A = 1, Y = 1) = \Pr(\hat{Y} = 1|A = 0, Y = 1)$). To measure this, we use the Equal Opportunity Difference (EOD), defined as*

$$EOD = \Pr(\hat{Y} = 1|A = 0, Y = 1) - \Pr(\hat{Y} = 1|A = 1, Y = 1). \tag{2}$$

with an ideal value equal to zero.

Definition 2 (Statistical Parity [6]) *: Statistical parity rewards the classifier for classifying each group as positive at the same rate. Thus, a binary predictor $\hat{Y}$ is fair from the statistical parity perspective if $\Pr(\hat{Y} = 1|A = 1) = \Pr(\hat{Y} = 1|A = 0)$. The Statistical Parity Difference (SPD) metric (ideally should have a value of zero) is defined as*

$$SPD = \Pr(\hat{Y} = 1|A = 0) - \Pr(\hat{Y} = 1|A = 1). \tag{3}$$

3.3 Global vs local group fairness in federated learning

The fairness definitions above can be readily applied to centralized model training to evaluate the performance of the trained model. However, in FL, clients typically have non-IID data distributions, which gives rise to low levels of fairness consideration in FL: *global fairness* and *local fairness*.

In particular, the **global fairness** performance of a given model takes into account the full dataset $\bar{\mathcal{D}} = \cup_k \mathcal{D}_k$ across the K clients participating in FL. In contrast, when only the local dataset $\mathcal{D}_k$ at client k is considered, we define the **local fairness** performance by applying the equations (2) and (3) on the data distribution at client k.

We further explain the two definitions below using the example of the *Equal Opportunity Difference* metric. For a trained classifier $\hat{Y}$, the global fairness EOD metric F_{global} is given by

$$F_{global} = \Pr(\hat{Y} = 1 | A = 0, Y = 1) - \Pr(\hat{Y} = 1 | A = 1, Y = 1), \tag{4}$$

where the probability above is based on the full dataset distribution (a mixture of the distributions across the clients). We can similarly define the local fairness metric F_k at client k as

$$F_k = \Pr(\hat{Y} = 1 | A = g, Y = 1, C = k) - \Pr(\hat{Y} = 1 | A = g, Y = 1, C = k), \tag{5}$$

where the parameter $C = k$ denotes that the k-th client and hence its local distribution (and dataset $\mathcal{D}_k\}$) is considered in the fairness evaluation.

4 FairFed: Fairness-aware Federated Learning

We first start by discussing the challenges to achieving group fairness in FL, then introduce our FairFed framework to address these challenges.

4.1 Challenges and the goal of fairness in federated learning

Given the notions of global/local group fairness defined in 3, we face the following challenges in FL:

• *Local fairness does not imply global fairness*: Due to the non-IID nature of the data distribution across clients, the full data distribution may not be represented by any single local distribution at any of the clients. Thus, for a classifier, $\hat{Y}$, the local fairness metrics $\{F_k\}_{k=1}^K$ and the global metric F_{global} may be quite different as an artifact of the difference between local and global distributions.

• *Local debias mitigation cannot improve the global group fairness:* Applying debias mitigation techniques – which are typically used in centralized training [6] – locally with FedAvg does not significantly improve the global group fairness of the model trained using FL (see experimental results in Table 1 from Section 5.2). In fact, in some cases, local debiasing can be counterproductive as the global minority group (e.g., African-American) might represent a local majority in the local dataset (e.g., credit union data in a black-majority city such as Detroit). Based on these observations, we can now define the ultimate goal of this work as follows.

Our Goal. The goal of this work is to develop an aggregation framework in FL that carefully benefits from the local bias mitigation techniques in order to output a high-performance global model that is also fair from a global group fairness perspective.

4.2 Our proposed approach

Recall that in the t-th iteration in FedAvg [19], local model updates $\{\theta_k^t\}_{k=1}^K$ are weighted averaged to get the new global model parameter θ^t as: $\theta^{t+1} = \sum_{k=1}^K \omega_k^t \, \theta_k^t$, where the weights $\omega_k^t = n_k / \sum_k n_k$ depend only on the number of data points at each client.

Note that naive aggregation favors clients with more data points. If the training performed in these clients results in locally biased models, then we end up with a biased global model since the weighted averaging exaggerates the contribution of the model update from these clients.

Based on this observation, in FairFed, we propose a method to achieve global group fairness F_{global} via adjusting the weights of different clients based on their local fairness metric F_k in addition to their local dataset sizes. In particular, given the global fairness metric F_{global}, the server gives a higher weight to clients that have a similar local fairness F_k to the global fairness metric.

Next, we illustrate how the aggregation weights for FairFed are computed at the server. A summary of the the steps performed while tracking the EOD metric in FairFed are summarized in Algorithm 1.

Algorithm 1: FairFed Algorithm

Server executes:

Initialize global model parameter θ_0;

Aggregate dataset statistics $\mathcal{S} = \{\Pr(A=1,Y=1), \Pr(A=0,Y=1)\}$ from clients through secure aggregation and send to clients;

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

  for *each client* $k = 1, \cdots K$ **in parallel do**

    $\theta_k^t,\ F_k^t,\ m_{\text{global},k}^t \leftarrow \text{ClientLocalUpdate}(k, \theta^t)$;

  end

  $F_{\text{global}}^t \leftarrow \sum_{k=1}^K m_{\text{global},k}^t$ *// Aggregate to get global metric as in (7);*

  $\bar{\omega_k}^t \leftarrow \exp(-\beta |F_k^t - F_{\text{global}}^t|) \cdot \frac{n_k}{\sum_{k=1}^K n_k}, \quad \forall k \in [K]$;

  $\omega_k^t \leftarrow \bar{\omega_k}^t / \sum_{i=1}^K \bar{\omega_i}^t, \quad \forall k \in [K]$;

  $\theta^{t+1} \leftarrow \sum_{k=1}^K \omega_k^t\, \theta_k^t$

end

ClientLocalUpdate(k, θ)**:**

$F_k \leftarrow \text{LocalFairnessMetric}(\theta, \mathcal{D}_k)$ *// Compute local fairness metric on θ^t using (5);*

$m_k \leftarrow \text{GlobalFairComponent}(\theta, \mathcal{D}_k, \mathcal{S})$ *// Get global fairness component as defined in (7);*

$\theta_k \leftarrow \text{LocalFairTraining}(\theta, \mathcal{D}_k)$ *// Local training at client k with local debiasing;*

Return θ_k, F_k, m_k to server

4.3 Computing the aggregation weights for FairFed at the server

Particularly, for the k-th client, we assign the weight ω_k based on the difference between the global fairness metric F_{global} and the local fairness metric F_k:

$$\bar{\omega}_k^t = \exp(-\beta |F_k^t - F_{global}^t|) \cdot \frac{n_k}{\sum_{k=1}^K n_k},\ \forall k \in \{1, \cdots, K\},$$
$$\omega_k^t = \bar{\omega}_k^t / \textstyle\sum_{i=1}^K \bar{\omega}_i^t,\ \forall k \in \{1, \cdots, K\}. \tag{6}$$

where β is a parameter that controls the fairness budget, which controls the trade-off between model utility and fairness. Higher values of β result in fairness metrics having a higher impact on the model optimization, while a lower β results in a reduced perturbation of the FedAvg weights due to fairness training; note that for $\beta = 0$, FairFed is equivalent to FedAvg. From the definition of ω_k in (6), the clients whose local fairness metric is similar to the global fairness metric will be assigned higher weights, while clients that have local metrics that significantly deviate from the global metric will have their FedAvg weights (i.e., $n_k / \sum_k n_k$) modified to lower values.

Thus, so far, the training process of FairFed at each iteration follows the following steps: 1.Each client reports their updated local model parameters and local fairness metric values F_k to the server based on the last global model; 2. The server computes the global fairness metric value F_{global}; 3. The server updates the aggregation weights ω_k based on the difference between the global metric and the participants local metrics as defined in (6); 4.The server aggregates local model updates and broadcasts the global model to the K clients.

4.4 How to privately compute the global metric at the server

One central assumption of FairFed presented above is the ability of the server to compute the global metric F_{global} in each iteration without the clients having to share their local dataset with the server. We demonstrate how the server can compute the F_{global} from non-private information sent by the clients by considering the EOD metric. Similar computations follow for the SPD metric. Note that the EOD metric in (4) can be rewritten as:

$$EOD = \Pr(\hat{Y}=1|A=0,Y=1) - \Pr(\hat{Y}=1|A=1,Y=1)$$
$$= \sum_{k=1}^K \underbrace{\frac{n_k}{\sum_{i=1}^K n_i} \sum_{a=0}^1 (-1)^a \Pr(\hat{Y}=1|A=a,Y=1,C=k) \frac{\Pr(A=a,Y=1|C=k)}{\Pr(Y=1,A=a)}}_{m_{global,k}}, \tag{7}$$

Thus the global EOD metric F_{global} can be computed by aggregating the values of $m_{global,k}$ from the K clients. Note that the conditional distributions in the definition of $m_{global,k}$ above are local performance metrics that can easily be computed by client k using its local dataset $\mathcal{D}_k$.

The only non-local terms in $m_{global,k}$ are the full dataset statistics $\mathcal{S} = \{\Pr(Y = 1, A = 0), \Pr(Y = 1, A = 1)\}$. These statistics $\mathcal{S}$ can be aggregated at the server using a single round of a secure aggregation scheme (e.g., [2]) at the start of training and then shared with the K participating clients to enable them to compute their global fairness component $m_{global,k}$.

Flexibility of FairFed to heterogenous debiasing. Note that the FairFed weights ω_k^t in (6) rely only on the values of the global/local fairness metrics and are not tuned towards a specific local debiasing method. Thus, we believe that FairFed is flexible to applying different debiasing methods at each client, and the server will incorporate the effects of these different methods by reweighting their respective clients based on their reported local fairness and the weight computation in (6).

5 Evaluation

5.1 Experimental setup

Implementation. We developed FairFed using `FedML` [10], which is a research-friendly FL library used to explore new algorithms. To accelerate the training, we used its parallel training paradigm, where each FL client is handled by an independent process using MPI (message passing interface). We conducted experiments in a server with `AMD EPYC 7502 32-Core` CPU Processor.

Datasets. In this work, we adopt as case studies two binary decision datasets that are widely investigated in fairness literature: the Adult [5] dataset and ProPublica COMPAS dataset [15]. For the Adult dataset, the race (defined as white or non-white) of each subject is considered the sensitive attribute in our experiments; for the COMPAS dataset, sex (defined as male or female) is considered the sensitive attribute. We provide a summary of the constitution of each dataset in Appendix A.1.

Configurable data heterogeneity for diverse sensitive attribute distributions. To fully understand our method and the baselines under different sensitive attribute distributions across clients, we need a configurable data synthesis method. In our context, we use a generic non-IID synthesis method based on the Dirichlet distribution proposed in [11] but apply it in a novel way for *configurable sensitive attribute distribution*: for each sensitive attribute value a, we sample $\mathbf{p}_a \sim \text{Dir}(\alpha)$ and allocate a portion $p_{a,k}$ of the data points with $A = a$ to client k. The parameter α controls the heterogeneity of the distributions at each client, where $\alpha \to \infty$ results in IID distributions. An example of this heterogeneous data distribution can be found in Appendix A.1.

Baselines. We adopt the following state-of-the-art solutions as our baselines:
• **FedAvg [19]:** the original federated learning algorithm for distributed training of private data. It does not consider the fairness of different demographic groups.
• **FedAvg + Local reweighting:** Each client adopts the reweighting strategy [13] to debias its local training data, then trains local models based on the processed data. FedAvg is used to aggregate the local model updates at the server.
• **FedAvg + Global reweighting [1]:** A differential-privacy approach to collect statistics such as the noisy number of samples with privileged attribute values and favorable labels from parties. The server will compute global reweighing weights based on the collected statistics and share them with parties. Parties assign the global reweighing weights to their data samples during FL training.

Hyperparameters. In our FedFair approach and the above baselines, we train a logistic regression for binary classification learning tasks on the aforementioned datasets. All experimental results are selected from the best accuracy obtained from grid search on important hyperparameters such as the fairness budget β and learning rate. For each hyperparameter configuration, we report the average performance of 20 random seeds. We summarize all hyperparameters in Appendix A.3.

5.2 Experimental results

Performance under the heterogeneous sensitive attribute distribution. Under partitions with different heterogeneous levels, we compared the performance of FedAvg, local reweighting, and FairFed. The results are summarized in Table 1. In high homogeneous data distributions (i.e., a large α value), FairFed does not provide significant gains in fairness performance. This is due to the fact

	Method	Adult ($\beta = 1$)					COMPAS ($\beta = 1$)				
		Heterogeneity Level α					Heterogeneity Level α				
		0.1	0.2	0.5	10	5000	0.1	0.2	0.5	10	5000
Acc.	FedAvg	**0.830**	**0.830**	**0.830**	**0.829**	**0.829**	**0.664**	**0.664**	**0.664**	**0.663**	0.663
	Local	0.829	0.829	0.829	**0.829**	**0.829**	**0.664**	0.663	0.664	0.662	0.663
	Global	0.829	0.829	0.829	0.828	0.828	0.663	0.663	0.662	0.662	0.662
	FairFed	0.818	0.822	0.828	0.828	**0.829**	0.648	0.644	0.652	0.662	**0.664**
EOD	FedAvg	-0.041	-0.040	-0.042	-0.041	-0.041	-0.068	-0.067	-0.069	-0.068	-0.066
	Local	-0.039	-0.038	-0.040	-0.039	-0.040	-0.068	-0.068	-0.067	-0.066	-0.066
	Global	-0.040	-0.039	-0.041	-0.039	**-0.039**	-0.067	-0.067	-0.067	-0.066	**-0.065**
	FairFed	**-0.005**	**-0.019**	**-0.033**	**-0.038**	-0.040	**-0.059**	**-0.055**	**-0.060**	**-0.065**	-0.066
SPD	FedAvg	-0.062	-0.062	-0.062	-0.062	-0.062	-0.128	-0.128	-0.129	-0.129	-0.128
	Local	-0.061	-0.061	-0.061	**-0.060**	**-0.060**	-0.128	-0.128	-0.130	-0.127	-0.128
	Global	-0.062	-0.062	-0.062	-0.061	-0.061	-0.129	-0.127	-0.130	-0.127	-0.127
	FairFed	**-0.041**	**-0.049**	**-0.058**	**-0.060**	**-0.060**	**-0.114**	**-0.110**	**-0.118**	**-0.126**	**0.126**

Table 1: Performance comparison of data partition with different heterogeneity levels α. A smaller α indicates a more heterogeneous distribution across clients. We report the average performance of 20 random seeds. For EOD and SPD metrics, values closer to zero indicate better fairness.

that under homogeneous sampling, the distributions of the local datasets are statistically similar (and reflect the original distribution with enough samples), resulting in similar weights being computed in all clients. For a larger level of heterogeneity in the sensitive attribute (i.e., a lower $\alpha = 0.1$), FairFed can improve EOD in Adult and COMPAS data by 87% and 13%, respectively, at the cost of an accuracy reduction to 0.818 and 0.648 (only a 1% and 2.5% decrease), respectively. In contrast, at the same heterogeneity level, both local and global reweighing strategies can only improve EOD by 7% and 1% for COMPAS and Adult datasets, respectively.

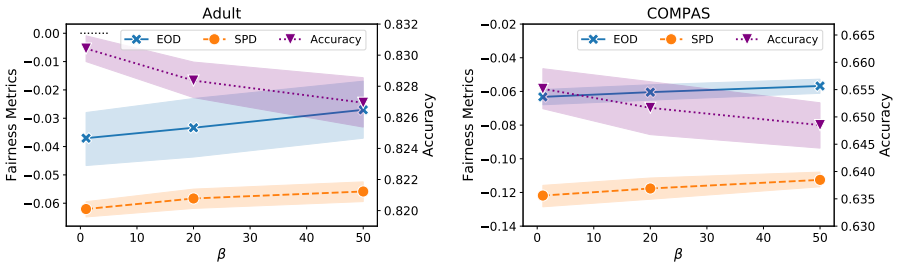

Figure 2: Effects of fairness budget β for $K = 5$ clients and heterogeneity parameter $\alpha = 0.5$.

Performance under different fairness budgets (β). In FairFed, we introduce a fairness budget parameter β, which controls the trade-off between accuracy and fairness (refer to Equation 6 for the explanation of β). Figure 2 visualizes the effects of β using heterogeneity level $\alpha = 0.5$ as an example. As the value of β increases, the fairness constraint has a bigger impact on the aggregation weights, yielding better fairness performance with a trade-off for the model accuracy.

In addition, we conducted experiments to verify the performance under different numbers of clients and also checked the system efficiency (see Appendix A.2 for details).

6 Conclusion and Future Works

In this work, motivated by the importance and challenges of group fairness in federated learning, we propose the FairFed algorithm to enhance group fairness via a fairness-aware aggregation method, aiming to provide fair model performance across different sensitive groups (e.g., racial, gender groups) while maintaining high utility. Though our proposed method outperforms the state-of-the-art fair FL frameworks under high data heterogeneity, limitations still exist. As such, we plan to further improve FairFed from these perspectives: 1) We report the empirical results on binary classification tasks in this work. We will extend the work to various application scenarios (e.g., regression tasks, natural language processing); 2) We will extend our study to scenarios of heterogeneous application of different local debiasing methods and understand how the framework can be tuned to incorporate updates from these different debiasing schemes; 3) We plan to improve the formulation of aggregation weights in FairFed in order to improve the trade-off between the gained fairness vs. lost accuracy; 4) The current FairFed algorithm mainly focuses on the group fairness. We plan to integrate FairFed with other fairness notions in FL, such as *collaborative fairness* and *client-based fairness*.

References

[1] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

[2] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

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

[4] 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.

[5] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[6] 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.

[7] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[8] 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 *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[10] Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, et al. Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*, 2020.

[11] Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[12] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary B. Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim Y. El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Oluwasanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, R. Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Xiaodong Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14:1–210, 2021.

[13] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[14] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254. ACM, 2019.

[15] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[16] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

[17] 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.

[18] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. In *Federated Learning*, pages 189–204. Springer, 2020.

[19] 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*, pages 1273–1282. PMLR, 2017.

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

[21] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[22] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[23] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*, 2020.

[24] 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*, 2021.

[25] 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. ACM, 2018.

[26] 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.

A Appendix

A.1 A Summary of datasets used in our experiments

The **Adult** dataset [5] contains 32,561 records of yearly income (represented as a binary label: over or under $50,000) and twelve categorical or continuous features including education, age, and job types. The race (defined as white or non-white) of each subject is considered as sensitive attribute.

The ProPublica **COMPAS** dataset [15] relates to recidivism, to assess if a criminal defendant will commit an offense within a certain future time. The dataset is gathered by ProPublica, with information on 6,167 criminal defendants who were subject to screening by *COMPAS*, a commercial recidivism risk assessment tool, in Broward County, Florida from 2013–2014. Features in this dataset include number of prior criminal offenses, age of the defendant, etc. The sex (classified as male or female) of the defendant is the sensitive attribute of interest.

Client ID	$\alpha = 0.1$		$\alpha = 10$	
	$A = 0$	$A = 1$	$A = 0$	$A = 1$
0	269	615	1505	3585
1	128	29839	876	5695
2	418	74	978	7261
3	43	392	601	5848
4	4196	203	1094	8734

Table 2: An example of the heterogeneous data distribution (non-IID) on the sensitive attribute A (race) used in experiments on the Adult dataset. The shown distributions are for $K = 5$ clients and two distribution heterogeneity parameters $\alpha = 0.1$ and $\alpha = 10$.

A.2 More experimental results

Performance under the different number of clients. Another factor that can impact FL fairness is the number of participating clients. In Figure 3, we compare the performance of FairFed with 5 and 10 clients over the Adult dataset. In general, FairFed performs better with smaller client counts. As the number of clients increases, our method can achieve more improvements to fairness with a more homogeneous data distribution.

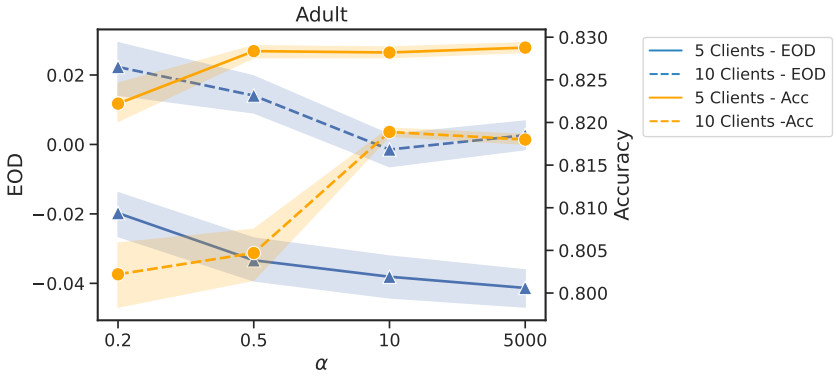

Figure 3: **Effects of number of clients**

System efficiency. At the current stage, we mainly focus on understanding the model and fairness performance for group fairness. With the help of our parallel training implementation, each training experiment can be finished in only a few minutes without GPU accelerators. As for the efficiency in an even larger scale datasets and number of clients, we hope to explore this in future works.

A.3 Hyperparameters in Experiments

In Table 3, we summarize the hyperparameter configurations used in our experiments. Each configuration of the hyperparameters is run for 20 random seeds.

Hyperparameter	Dataset	
	COMPAS	**Adult**
Optimizer	Adam, lr={0.01, 0.001}, wd=0.0001	Adam, lr={0.01, 0.001}, wd=0.0001
Local epochs	1	1
Communication rounds	20	20
Number of clients	5	{5,10}
β parameter in FairFed	{1,20,50}	{1,20,50}

Table 3: Hyperparameters used in Experiments on the COMPAS and Adult datasets.