Fairness Is Not Universal: Towards Contextualized and Autonomous Fairness in Federated Learning

Yi Zhou
University of Oxford
Oxford, UK
yi.zhou@keble.ox.ac.uk

Naman Goel
University of Oxford and Alan Turing Institute
Oxford, UK
naman.goel@alumni.epfl.ch

ABSTRACT

Federated learning (FL) enables distributed model training without centralizing data, but existing fairness-aware FL methods overwhelmingly treat fairness as a global property—a constraint or objective that can be defined and applied uniformly across all clients. In this paper, we argue that this paradigm is fundamentally insufficient. Fairness is contextual, population-dependent, and often governed by domain- or jurisdiction-specific values. To investigate the practical implications of this mismatch, we empirically evaluate existing methods against two client-side post-training interventions—model output post-processing and final-layer fine-tuning. Across four datasets of different modalities (tabular, signal, and image) and multiple heterogeneity settings, we show that even when all clients nominally commit to the same fairness metric, current fairness-aware FL algorithms fail to provide consistent context-sensitive fairness improvements. We find that the post-training interventions perform more reliably in our experiments than the fair FL methods we evaluate, but they remain only partial remedies and highlight the need for novel FL frameworks that explicitly support contextual and autonomous fairness.

CCS CONCEPTS

• **Computing methodologies** → **Distributed computing methodologies**; **Machine learning**; **Distributed artificial intelligence**; • **Security and privacy**; • **Human-centered computing**;

KEYWORDS

Algorithmic Fairness, Federated Learning

1 INTRODUCTION

Federated learning (FL) enables multiple parties to collaboratively train a machine learning model without sharing their raw data [32]. By keeping data decentralized, FL is designed to address settings in which privacy, data governance, or regulatory constraints make data pooling undesirable or infeasible. Example domains include healthcare, finance, and mobile sensing [23, 51]. Most FL frameworks, including the widely used FedAvg algorithm [32], are built around the principle of learning a single global model that aggregates information from all participating clients.

As FL applications are being explored in socially sensitive domains, interest has grown in incorporating algorithmic fairness into federated training. Fair machine learning refer to making machine learning models produce more fair outputs [5]. For example, in some application contexts, the output of a binary classifier, that has been trained with historical data using machine learning, may need to satisfy certain fairness properties like conditional independence from attributes like gender and race [15, 49].

A rapidly expanding body of work proposes fair FL methods to enforce fairness constraints or achieve parity across groups. However, most of these methods assume that fairness is a global property—a constraint or metric that can be applied uniformly across all clients [2, 10, 50, 52]. This assumption mirrors the centralized ML setting, where a single model serves a single context, but it conflicts with the sociotechnical reality of FL deployments.

Fairness is fundamentally contextual and situated. Numerous works in algorithmic fairness and ethics emphasize that fairness cannot be abstracted away from the local population, domain, legal environment, or institutional values in which a system operates [5, 14, 41]. While the literature in fair FL does acknowledges the heterogeneity across clients due to different data distributions and measures local fairness on local distributions, it ignores other types of heterogeneity. Clients within an FL system may serve different user groups, fall under different regulatory regimes, or prioritize different fairness notions depending on local harms, protected attribute definitions, risks, and operational constraints. Treating fairness as a single (or aggregated) global objective in FL ignores the contextual, local, and often jurisdiction-specific nature of fairness.

We argue that the literature on fairness in FL should not be restricted to methods that do not offer the clients the flexibility to operationalise fairness independently. Research in this area should therefore be extended to the design and study of methods that explicitly support contextual and autonomous fairness. As a step in that direction and to give a formal example of our position, we also introduce a simple post-training intervention based new baseline. In this approach, all clients first collaborate to perform global FL training without any fairness constraints. Once the global training is finished, each client performs local fairness or debiasing post-processing using its own dataset, tailoring the debiasing process towards its requirements and specific data distribution. This decoupled approach offers better flexibility and decentralization as it allows clients to apply different fairness constraints based on their specific needs. Since the debiasing is performed locally, it also adapts fairness to the local data distribution. Moreover, no information about sensitive attributes is required during the global training stage, thus offering better sensitive data protection, and there is no additional network communication cost for debiasing.

We demonstrate this approach using two post-training approaches (but other post-training approaches can also be similarly integrated). One is model output post-processing [15], which addresses fairness by solving a linear program and computing a derived predictor. It does not need to change the original model weights and instead post-processes the model output. The other one is the final layer fine-tuning technique, which fine-tunes the last layer of neural

network based models, with fairness constraints added to optimization objective. But it may be computationally more expensive and modifies model weights.

We conducted experiments to compare performance with three other methods: FedAvg [32], FairFed [13] and FairFed combined with Fair Linear Representation (FairFed/FR) [13]. We included four different datasets in our experiments. Two of them are commonly used tabular datasets in the fair ML literature, namely the Adult dataset [6] and the ProPublica COMPAS dataset [31]. The other two are: a signal dataset, called PTB-XL Electrocardiogram (ECG) dataset [44], and a medical images dataset, called NIH Chest X-Ray dataset [46]. To simulate different levels of data heterogeneity across clients, we use three heterogeneity settings.

Our experiments show that, even in a simple experimental setting in which clients nominally use the same fairness criteria and the only source of heterogeneity is local data distribution, the post-training intervention offers better fairness compared to other methods on most of the datasets. However, the post-training is neither sufficient (as it often fails to reduce unfairness sufficiently and consistently) nor optimal (as it often leads to a significant drop in accuracy). Therefore, further research is required for the development of advanced methods to support contextual and autonomous fairness in federated learning.

2 RELATED WORK

Federated Learning. In federated learning (FL) [32], different parties which are often called "clients" can collaborate with each other via a "server" without sharing their training data. During each global training round of FL, each client computes gradient updates on its local dataset, and then, instead of sharing their data directly, they only share the updates with the server. The server aggregates the updates from all the clients to compute an updated global model weights, and sends the new global modal to all clients for them to compute a new round of updates. This procedure repeats for a few global rounds until the global model reaches a certain performance or converges. One of the other major challenges in FL is that of heterogeneity across clients, with data heterogeneity being the most well-known type of heterogeneity [20, 33]. Personalized FL (PFL) [3, 24, 45] address data heterogeneity and individual needs by "personalizing" the global model towards a local objective. To the best of our knowledge, these works in PFL do not consider fairness.

Fairness in Machine Learning. Fairness in machine learning (and algorithmic decision-making systems broadly) is a very broad area of research [5, 18]. For example, group fairness requires that different demographic groups (often distinguished by sensitive attributes such as gender and race) should have the same chance to receive certain (conditional) outcomes [15]. For e.g., conditioned on "true" qualification or label, group fairness may require a college admission decision classifier to ensure equal admission rates in different demographic groups. There are different versions of group fairness in the ML literature such as demographic parity, equal odds, etc [5].

Fairness in FL. Most existing work for fairness in FL [39] trains a single global model with some fairness constraints. The literature does consider local fairness; but either adopts a narrow view of local fairness (e.g. measure a global fairness function on local data distributions) or considers local fairness as secondary (e.g. combine local fairness preferences into an aggregate global one).[1]

In FairFed [13], clients send ocal updates as well as local fairness measurement to the server in each global round. The server aggregates the local updates based on the fairness gap between each local fairness measurement and the global fairness (e.g. updates sent by "fairer client", i.e. client with the lowest fairness gap, have higher weight in global update). All clients receive the same model satisfying a global fairness notion after training. [9] and [28] are closer to FairFed [13] discussed above; they consider new ways of incorporating local fairness estimates during global training, but at the end of training all clients receive the same model satisfying a global fairness notion. [11] propose that clients use a post-processing function calculated by the server to achieve global fairness.

3 POST-TRAINING FAIRNESS INTERVENTION

To support clients in operationalising fairness independently, we introduce a simple baseline based on post-training intervention. Consider the following problem setting. There are K clients and one server participating in the FL training. Each client C_k ($k \in \{1, .., K\}$) has a local dataset D_k. All local datasets together form the global set $D = \cup_k D_k$. The distributions of D_k may differ for different k, due to heterogeneous data across clients.

The post-training intervention approach can be divided into mainly two stages, the federated training stage and the local debiasing stage. Only the training stage requires communication between clients and the server, and the debiasing stage is fully local.

3.1 Federated Training Stage

The training stage follows the training procedure of standard FL setting in FedAvg [32]. The general objective function in FedAvg can be written as follows: $f(\theta) = \sum_{k=1}^{K} w_k l_k(\theta)$, where θ denotes the model parameters, l_k denotes the local objective function on client k. The local objective function is fairness unaware, e.g. a vanilla loss function like cross-entropy loss. Minimizing function $f(\theta)$ finds model parameter θ that minimizes the weighted average of the local model losses across all clients.

3.2 Local Debiasing Stage

In the debiasing stage (which comes after the first stage is finished), the global model is sent to each client for them to evaluate the model locally on their dataset D_k, and apply different debiasing methods based on their local fairness constraint F_k. We discuss two examples of debiasing approaches that can be used in the post-training intervention: 1) output post-processing method for any binary classifier [15]; and 2) final layer fine-tuning method for deep neural networks [30].

For brevity, we will only use Equalized Odds (EOD) [15] as fairness metric in the examples (and the experiments), but note that due to the decoupling of global model training and local debiasing, different clients are not required to follow the same fairness definition or constraint. Clients can adjust to different fairness metrics,

[1] A related line of work in fair federated learning (e.g. [24, 25, 34]) is about utility fairness for clients participating in FL. In contrast, we focus on the social attributes based fairness for decision subjects.

different levels of fairness-accuracy trade-off, even different sensitive attributes based on their specific requirements. Clients can also use different post-process debiasing methods methods.

Equalized Odds (EOD) [15] defines fairness as groups having the same true positive rate (TPR) and false positive rate (FPR). For evaluating and comparing EOD between different methods, we will measure EOD as the maximum of the absolute difference of TPR and the absolute difference of FPR between different groups, as shown in Equation 1. This is consistent with popular fairness toolkits like IBM's AI Fairness 360 [7].

$$\begin{aligned} EOD =& Max(|TPR_{A=1} - TPR_{A=0}|, |FPR_{A=1} - FPR_{A=0}|) \\ =& Max(|Pr(\hat{Y}=1|A=1,Y=1) - Pr(\hat{Y}=1|A=0,Y=1)|, \\ & |Pr(\hat{Y}=1|A=1,Y=0) - Pr(\hat{Y}=1|A=0,Y=0)|) \end{aligned} \quad (1)$$

where A is the sensitive attribute, Y is the true label, and $\hat{Y}$ is the decision (or prediction).
For each client C_k, we measure local EOD metric as (Equation 2):

$$\begin{aligned} EOD_k =& Max(|TPR_{A=1,C_k} - TPR_{A=0,C_k}|, \\ & |FPR_{A=1,C_k} - FPR_{A=0,C_k}|) \\ =& Max(|Pr(\hat{Y}=1|A=1,Y=1,C_k) \\ & - Pr(\hat{Y}=1|A=0,Y=1,C_k)|, \\ & |Pr(\hat{Y}=1|A=1,Y=0,C_k) \\ & - Pr(\hat{Y}=1|A=0,Y=0,C_k)|) \end{aligned} \quad (2)$$

where $|, C_k$ denotes that measure is calculated for client C_k and their local data distribution (approximated using local data D_k).

3.2.1 FL Model Output Fairness Post-Processing. Algorithm 1 in Appendix A shows the pseudo-code of post-training intervention with FL model output fairness post-processing.

It starts by general FL training in lines 1-8 [32]. At the termination of the global `for` loop, each client $k \in 1, ..K$ has received a trained FL model with weights ω_T. In line 10, predictions $\hat{Y}_k$ are computed using the global model ω_T for the local dataset D_k for each client $C_k(k = 1, 2, ...K)$. In line 11, each client computes a derived predictor p_k based on the prediction $\hat{Y}_k$ and the local dataset D_k. The method of obtaining the derived predictor is adopted from Hardt et al. [15]. Each client computes a derived predictor separately based on their local context and requirements. For brevity, we do not distinguish between clients in the pseudo-code of Algorithm 1 (beyond the distinction in their datasets D_k). In principle, each client can also use different fairness definitions and post-processing methods in lines 10-11 and if a client doesn't wish to enforce fairness because of local application context, they can also skip lines 10-11.

Obtaining Derived Predictor [15] . Formally, a derived predictor can be defined using the following probabilities:

$$p_{ya} = Pr(\tilde{Y}=1|\hat{Y}=y, A=a)$$

The four probabilities $p = (p_{00}, p_{01}, p_{10}, p_{11})$ specify the derived prediction $\tilde{Y}_p$. Thus, the derived predictor probabilistically flips the prediction of the model depending on the value of the sensitive attribute and the prediction, in order to optimize EOD fairness.

A client computes their local derived predictor p using the following equation as the solution of a linear program:

$$\begin{aligned} \min_p \ & \mathbb{E}\, loss(\tilde{Y}_p, Y) \\ \text{s.t. } & \gamma_0(\tilde{Y}_p) = \gamma_1(\tilde{Y}_p) \\ & \forall_{y,a}\ 0 \le p_{ya} \le 1 \end{aligned} \quad (3)$$

where $\hat{Y}$ is the prediction by the model, $\tilde{Y}_p$ is the prediction by the derived predictor p, a is the value of sensitive attribute, and

$$\gamma_a(\tilde{Y}) \stackrel{\text{def}}{=} (Pr(\tilde{Y}=1|A=a, Y=0), Pr(\tilde{Y}=1|A=a, Y=1))$$

In other words, a client finds their derived predictor p by minimizing the expected loss between the derived prediction $\tilde{Y}_p$ and true label Y while satisfying the EOD fairness constraints $\gamma_0(\tilde{Y}_p) = \gamma_1(\tilde{Y}_p)$. Y and A come from their local dataset D_k. In our experiments, we also use the same definitions of *loss* as [15]

3.2.2 FL Model Final Layer Fairness Fine-Tuning. Algorithm 2 in Appendix A shows pseudo-code of post-training intervention with FL model final layer fairness fine-tuning for the debiasing stage.

Lines 1-8 are the same as Algorithm 1 since this is the global FedAvg model training stage. Each client receives a copy of the global model with weights ω_T at the end of this stage. In lines 9-15, we show the decoupled debiasing stage that is executed independently for each client. As was the case in the model output fairness post-processing method discussed previously, clients can decide their fairness definitions, metrics etc or skip fairness enforcement depending on their local context. The distinction between clients is not shown for brevity. On the other hand, unlike the model output fairness post-processing method, this method relies on modifying the weights in the final layer of a neural network model to improve fairness. Mao et al. [30] showed in centralized ML setting that fixing other model weights and only fine-tuning the last layer can effectively improve the fairness in neural networks. Besides, in FL, fine-tuning only the last layer can potentially be useful in preserving information learned through the data of other clients during global FL training.

In line 10, a client fixes the weights of the model layers except for the last layer L. Then r rounds of fine-tuning are performed based using the local dataset D_k at client k, with fine-tuning learning rate η and the loss function L that considers both local accuracy based loss (l) and local fairness F_k dependent loss (l'). A parameter α assigns relative weight to the two losses. In our experiments, we use a loss l' for EOD i.e. the sum of the differences in TPR and FPR for two demographic groups (see Section 3). Our implementation code will also be released for further clarity and reproducibility.

4 EXPERIMENT SETTINGS

4.1 Datasets

For empirical analysis, we use two tabular datasets that are widely used in fair machine learning literature, namely Adult [6] and ProPublica COMPAS [31]. In addition, we use two other datasets that are both larger in size and more complex in structure compared with the tabular datasets. One of them is an ECG signal dataset called PTB-XL [44], and the other is a chest X-ray image dataset called NIH Chest X-Ray [46]. Due to space constraints, the details

of datasets and data cleaning etc are provided in Appendix B. These datasets are only included for demonstration purposes—inclusion of a dataset in experiments doesn't indicate our position about suitability of any specific fairness metric for a dataset.

4.2 Data Split

To simulate the data distributions across different clients, we partition each dataset to form local datasets without overlap. Then for each client, we split the local dataset into a training set and a test set, and only the training set is used during model development.

Clients Data Split: We use a 4-client setup for all the datasets [42]. To generate heterogeneity among clients, we use Dirichlet Distribution to sample data points, which is commonly used in fair FL literature [13, 42, 47]. We sample $p_i \sim Dir(\alpha)$ for each client $i \in \{1, .., 4\}$, where α is a parameter controlling the level of data heterogeneity. Smaller α provides a more heterogeneous distribution, and as $\alpha \to \infty$, the data distribution gets close to homogeneous and can be considered as random split [13].

To simulate different levels of data heterogeneity from extremely imbalanced to (very close to) random split, we will use $\alpha = 0.5, 5, 100$ in the experiments (more discussion to follow in Section 6.1).

Train/Test Data Split: We split train and test sets on each client independently. On each client, we use 80% samples as training set and 20% as test set. The global model is trained using all local training sets in the first stage (i.e. federated training). During local debiasing stage, we use local training set at the respective client.

4.3 Other Evaluated Methods

We compare the post-training intervention introduced in Section 3 against the following methods:[2]

FedAvg: FedAvg refers to the originally proposed FL training framework without any fairness considerations [32] . Each client computes its local update and sends it to the server for aggregation and global model update.

FairFed: As discussed in Section 2, FairFed [13] is based on FedAvg framework and uses fairness-aware aggregation.

FairFed + Fair Representation (FairFed/FR): Fair Representation [48] is a pre-processing method which de-bias the input features by removing their correlations with sensitive attributes. We also include FairFed combined with FairRep[48], for the reason that the authors of the FairFed paper [13] presented the performance of FairFed combined with local pre-processing. Under their experimental setting, FairFed works well with Fair Representation with heterogeneous data distributions too.

4.4 Evaluation Metrics

The evaluation metrics used in experiments are summarized below and Appendix C contains formal expressions:

Accuracy: We use the local (client-level) test accuracy as one of the main performance evaluation metrics on the tabular datasets Adult and COMPAS.

Balanced Accuracy: For PTB-XL and NIH Chest X-Ray, we use balanced accuracy [8] instead of accuracy. Health datasets are often very imbalanced with labels for disease prediction or other classification tasks. In most cases, within a given dataset only a minority of patients is diagnosed with the specific disease, and the prediction accuracy can be high even if the model just predicts every sample as "No disease". Balanced accuracy addresses this issue by including both sensitivity (TPR) and specificity (TNR) providing a more reliable measurement for imbalanced datasets. It is widely used in health ML literature such as [17, 27].

EOD: For fairness, we use local (client-level) Equalized Odds (EOD) as introduced in Section 3. It is the maximum of the absolute TPR difference between different groups and the absolute FPR difference between groups (Equation 2).

Weighted Average: In addition, we also include weighted average of accuracy, balanced accuracy, and EOD. It can be interpreted as a measurement of the average performance across clients.

5 MODEL DEVELOPMENT

Adult and COMPAS tabular datasets are generally evaluated with simple architectures in the literature [49]. We use a simple one-layer model to train on the Adult and COMPAS datasets. The activation function ReLU [4] is used in the model. We use stochastic gradient descent (SGD) [38] as the optimizer and Binary Cross Entropy Loss (BCELoss) [29] as the loss function when training on Adult and COMPAS.

For the ECG dataset PTB-XL, we used a ResNet-based model [26] with residual blocks [16] to handle the uni-dimensional ECG signals. The model is composed of one convolutional layer and five residual blocks. Each residual block consists of two convolutional layers by batch normalization [19], activation function ReLU [4] and Dropout regularizer [43]. We use Adam optimizer [21] with weighted mean square error loss function for the ECG dataset.

In the case of NIH Chest X-Ray dataset, we initialize our model with the pre-trained model MobileNetV2 [40]. The MobileNetV2 model is trained on Imagenet. Prior works such as [37] have shown that utilizing pre-trained models (including MobileNet) for initialization is useful for classification tasks on X-Ray datasets.

Code Availability. To supplement the above information, we will release the implementation/code. All algorithms require hyperparameters to be tuned. For FairFed and FairFed/FR, we report results with hyperparameters settings based on the information available in the original paper and that showed best results for these method in our implementation. Appendix D provides the hyperparameter settings used for all methods, as well as other details likes libraries used and the computation resources.

6 RESULTS AND DISCUSSION

6.1 Performance on COMPAS Dataset

We first investigate the impact of data heterogeneity level. We partition the COMPAS dataset across four clients using Dirichlet distribution $Dir(\alpha)$ with $\alpha = 0.5, 1, 5, 100, 500$, and evaluate the performance of different methods. The third row in Figure 1 shows the distribution of data labels across clients under different heterogeneity settings. As expected, lower values of α (e.g., $\alpha = 0.5$ and $\alpha = 1$) result in more imbalanced data partitions both in terms of label distribution and dataset size. In contrast, higher values ($\alpha = 100$

[2]For comparison, we restrict experiment settings where clients nominally use the same fairness metric. This is without loss of generalisation for the post-training intervention that can also accommodate heterogeneous fairness metrics.

and $\alpha = 500$) generate more balanced splits, with $\alpha = 500$ closely simulating an identically distributed setting, mimicking random data splits across clients. The medium level of heterogeneity is represented by $\alpha = 5$.

The first row of Figure 1 shows the EOD values for each method as data heterogeneity increases. Experiments were repeated for 10 random seeds under each setting and the average and standard deviation are reported. We can observe that firstly, the EOD for the standard FL model without fairness constraints (FedAvg) decreases as α increases, which suggest that more balanced splits generally result in less bias (lower EOD values are better). As for FairFed and FairFed/FR, the best performance is observed at around medium to high levels of data heterogeneity ($\alpha = 0.5, 1$). However, we can see that the performance of FairFed and FairFed/FR is not very robust with a large standard deviation (std). A possible reason could be sensitivity w.r.t. the FairFed hyper-parameter β.

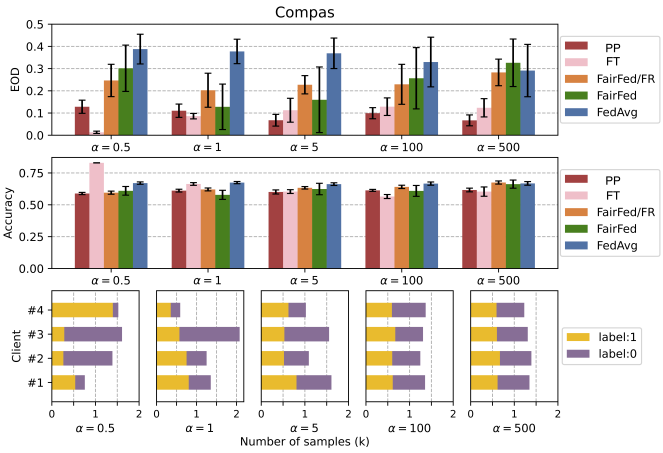

Figure 1: Experiment using COMPAS dataset with different data heterogeneity level α. Smaller α indicates more heterogeneous data partitions. Accuracy and EOD are weighted averaged across clients based on the dataset size; standard deviation (std) is shown with error bars. 1st row: Test accuracy (higher value is better). 2nd row: Equalized odds difference (lower value is better). 3rd row: Sample distribution on each client with different labels. FedAvg, FairFed, and FairFed/FR are methods described in Section 4.3. PP and FT refer to post-training intervention with model output post-processing and final layer fine-tuning respectively.

On the other hand, post-training intervention with both PP and FT methods shows good improvement in EOD under all heterogeneity settings with much smaller standard deviations. Simple output post-processing method can achieve the highest improvement when the data is not too imbalanced ($\alpha = 5, 100, 500$). When the heterogeneity is extremely high ($\alpha = 0.5, 1$), final-layer fine-tuning shows its effectiveness by reducing EOD close to zero.

The second row of Figure 1 provides the test accuracy with respect to data heterogeneity. Generally, models trained without any fairness constraints would obtain a higher accuracy compared with fairness fairness-constrained models. This situation can be observed in almost every method in Figure 1. However, when $\alpha = 0.5$, we can see that FT even achieves an increase in accuracy with a significant EOD improvement. The reason could be that, the fine-tuning procedure helps the model to optimize towards both the local fairness and accuracy goal as shown in line 12 of Algorithm 2, addressing data heterogeneity problem in fairness and prediction accuracy at the same time.

Henceforth, we will use three different data heterogeneity settings with $\alpha = 0.5, 5$ and 500 to simulate high, medium and very low (close to homogeneity) levels of data heterogeneity respectively.[3] Further, note that the final-layer fine tuning (FT) method is meant for deep neural networks. For the sake of completeness in Figure 1, we used a 2-layer neural network for COMPAS dataset, but in all further discussions, we will use a simple one-layer network for both tabular datasets Adult and COMPAS (as explained previously in Section 5). This also means that we will not discuss the FT method for tabular datasets henceforth.

Table 1 shows the client-wise performance of post-training intervention with model output post-processing (PP) method compared to other methods, under varying data heterogeneity. In general, PP shows a significant improvement in fairness at all clients as well as in the weighted average of local fairness under all heterogeneity settings. In contrast, the FairFed and FairFed/FR show lower fairness improvements. The irperformance is also less consistent, which suggests the sensitivity of FairFed-based approaches to experimental conditions. However, the improvement in fairness in PP comes at a cost of relatively more drop in accuracy.

Under extreme heterogeneity setting ($\alpha = 0.5$), post-training with PP, outperforms other methods in fairness improvement across clients and also show a significant improvement in weighted averaged EOD ($\sim$ 79% improvement over FedAvg). When the data partitions are less heterogeneous (with $\alpha = 5$ and $\alpha = 500$), PP becomes slightly more effective in improving fairness with an improvement of $\sim$ 85% for $\alpha = 500$ over FedAvg.

6.2 Performance on Adult Dataset

Compared with the COMPAS dataset, the Adult dataset is larger in size but more imbalanced in terms of label and sensitive attribute distribution with only 3.6% of samples having $label = 1$ and $sensitive\_attribute = 0$ (see Table 5 in Appendix 4.1).

Table 2 shows that post-training intervention with PP achieves the highest fairness improvement compared to other methods under all heterogeneity settings, with some accuracy decrease. FairFed provides a much smaller EOD improvement while maintaining a similar accuracy level as FedAvg. FairFed/FR generally appears to perform worse in terms of fairness-accuracy trade-off.

6.3 Performance on PTB-XL ECG Dataset

Table 3 presents the performance comparison of different methods on PTB-XL ECG dataset under different heterogeneity settings.

[3]We also considered $\alpha = 0.1$ as one of the heterogeneity settings. However, we found that due to the small dataset size and imbalanced label split, there would be missing labels within some groups with $\alpha = 0.1$. This creates problems in experiments (e.g. inability to even measure fairness on some clients). Therefore, we consider $\alpha = 0.5$ as the highest level of data heterogeneity. Similarly, we observed the data heterogeneity with $\alpha = 500$ to be very small and quite close to random splits. Thus we will consider $\alpha = 500$ as the lowest level of heterogeneity.

COMPAS Dataset

α	Metric	Client	Method			
			FedAvg	FairFed	FF+FR	PP
0.5	Acc.	Avg	0.670	**0.674**	0.652	0.610
		C1	0.632	**0.705**	0.628	0.550
		C2	0.633	**0.658**	0.638	0.532
		C3	**0.729**	0.705	0.684	0.706
		C4	**0.704**	0.682	0.656	0.697
	EOD	Avg	0.418	0.378	0.407	**0.088**
		C1	0.464	0.405	0.415	**0.078**
		C2	0.370	0.348	0.379	**0.064**
		C3	0.387	0.340	0.376	**0.101**
		C4	0.532	0.484	0.524	**0.137**
5	Acc.	Avg	**0.670**	0.663	0.660	0.602
		C1	**0.671**	0.661	0.661	0.583
		C2	0.648	0.640	**0.654**	0.591
		C3	0.708	**0.712**	0.703	0.679
		C4	**0.636**	0.613	0.599	0.527
	EOD	Avg	0.392	0.317	0.334	**0.086**
		C1	0.454	0.368	0.369	**0.062**
		C2	0.345	0.271	0.312	**0.118**
		C3	0.356	0.293	0.329	**0.075**
		C4	0.402	0.312	0.312	**0.106**
500	Acc.	Avg	**0.673**	0.671	0.672	0.603
		C1	0.671	0.666	**0.676**	0.602
		C2	**0.680**	0.679	0.678	0.607
		C3	**0.665**	0.661	0.664	0.611
		C4	0.675	**0.678**	0.668	0.589
	EOD	Avg	0.416	0.414	0.374	**0.064**
		C1	0.422	0.434	0.391	**0.053**
		C2	0.401	0.368	0.336	**0.061**
		C3	0.399	0.424	0.400	**0.080**
		C4	0.444	0.434	0.373	**0.061**

Table 1: Client-wise performance comparison on COMPAS dataset with different data heterogeneity level α. Smaller α indicates more imbalanced partitions. Acc.: Test accuracy (higher is better). EOD: Equalized odds difference (lower is better). Grey highlights indicate the best performance in each row. FedAvg, FairFed, and FF+FR are methods described in Section 4.3. PP refers to post-training intervention with model output post-processing.

Adult Dataset

α	Metric	Client	Method			
			FedAvg	FairFed	FF+FR	PP
0.5	Acc.	Avg	**0.847**	0.833	0.784	0.821
		C1	**0.804**	0.767	0.741	0.774
		C2	**0.739**	0.704	0.648	0.694
		C3	0.925	**0.958**	0.924	0.908
		C4	**0.725**	0.625	0.500	0.684
	EOD	Avg	0.193	0.158	0.199	**0.060**
		C1	0.154	0.108	0.122	**0.051**
		C2	0.324	0.295	**0.118**	0.214
		C3	0.224	0.205	0.299	**0.068**
		C4	0.171	0.110	0.076	**0.042**
5	Acc.	Avg	**0.842**	**0.842**	0.829	0.810
		C1	**0.784**	0.780	0.757	0.743
		C2	**0.797**	0.792	0.760	0.750
		C3	0.875	**0.879**	0.873	0.852
		C4	0.872	**0.874**	0.872	0.845
	EOD	Avg	0.164	0.149	0.153	**0.066**
		C1	0.197	0.186	0.173	**0.052**
		C2	0.122	0.108	0.107	**0.062**
		C3	0.164	0.140	0.177	**0.101**
		C4	0.171	0.164	0.149	**0.042**
500	Acc.	Avg	0.841	**0.842**	0.841	0.810
		C1	0.836	**0.838**	0.835	0.806
		C2	0.845	**0.846**	**0.846**	0.816
		C3	**0.842**	0.841	**0.842**	0.806
		C4	0.840	**0.841**	0.840	0.812
	EOD	Avg	0.173	0.159	0.160	**0.044**
		C1	0.178	0.152	0.152	**0.032**
		C2	0.175	0.167	0.163	**0.031**
		C3	0.144	0.147	0.148	**0.032**
		C4	0.195	0.171	0.177	**0.081**

Table 2: Client-wise performance comparison on Adult dataset with different data heterogeneity level α, where smaller α indicates more imbalanced partition. Acc.: Test accuracy (Higher is better). EOD: Equalized odds difference (Lower is better). Grey highlights indicate the best performance in each row. FedAvg, FairFed, and FF+FR are methods described in Section 4.3. PP refers to post-training intervention with model output post-processing.

Note that, for both PTB-XL ECG and NIH-Chest X-Ray datasets, FairFed/FR will not be included in the results. The Fair Linear Representation (FR) approach is based on correlations between data features and sensitive attributes and, to the best of our knowledge, does not apply to signal or image datasets [48]. Therefore, we will only use FairFed and FedAvg for the following experiments. On the other hand, as discussed in Section 6.1, now we will also include the results for final-layer fine-tuning method (FT) in post-training in addition to results with the PP method.

Model Output Post-Processing (PP). Post-training intervention with model output post-processing (PP) for fairness is still very effective on more complex datasets and models, and outperforms other methods in fairness improvement. We can see from Table 3 that the performance of PP is not influenced much by the change of data heterogeneity, with an improvement of EOD by 87%, 83% and 86% at $\alpha = 0.5$, $\alpha = 5$, $\alpha = 500$, respectively. On the other hand, the improvement in fairness comes with a decrease of balanced accuracy by 13% - 16% across different heterogeneity settings, with the most balanced data distribution having the largest accuracy decrease. One possible reason could be that on the dataset with a higher heterogeneous distribution, local post-processing could fit the local distribution better.

PTB-XL Dataset

α	Metric	Client	Method			
			FedAvg	FairFed	PP	FT
0.5	BA	Avg	0.790	0.780	0.684	**0.803**
		C1	0.760	**0.773**	0.642	0.771
		C2	0.810	**0.811**	0.712	0.809
		C3	0.734	0.767	0.581	**0.806**
		C4	**0.858**	0.841	0.800	0.853
	EOD	Avg	0.342	0.308	**0.044**	0.299
		C1	0.334	0.286	**0.074**	0.308
		C2	0.369	0.330	**0.074**	0.407
		C3	0.349	0.316	**0.021**	0.267
		C4	0.172	0.198	**0.080**	0.186
5	BA	Avg	0.770	0.769	0.648	**0.790**
		C1	0.791	0.787	0.672	**0.799**
		C2	0.774	**0.778**	0.675	0.776
		C3	0.771	0.768	0.636	**0.777**
		C4	0.755	0.750	0.615	**0.805**
	EOD	Avg	0.342	0.358	**0.055**	0.319
		C1	0.412	0.431	**0.066**	0.421
		C2	0.347	0.342	**0.065**	0.327
		C3	0.360	0.392	**0.051**	0.341
		C4	0.288	0.311	**0.042**	0.239
500	BA	Avg	0.770	0.772	0.645	**0.793**
		C1	0.754	0.753	0.629	**0.784**
		C2	0.760	0.770	0.630	**0.789**
		C3	0.777	0.774	0.648	**0.797**
		C4	0.788	0.791	0.671	**0.803**
	EOD	Avg	0.345	0.343	**0.046**	0.319
		C1	0.364	0.359	**0.029**	0.334
		C2	0.335	0.334	**0.049**	0.307
		C3	0.355	0.350	**0.047**	0.340
		C4	0.328	0.332	**0.055**	0.298

Table 3: Client-wise performance comparison on PTB-XL ECG dataset with different data heterogeneity levels α. BA: Balanced accuracy (Higher is better). EOD: Equalized odds difference (Lower is better). Grey highlights indicate the best performance of each row. FedAvg and FairFed are methods described in Section 4.3. PP and FT refer to post-training intervention with model output post-processing and final-layer fine-tuning respectively.

Final-layer fine-tuning (FT). Post-training intervention with final-layer fine tuning (FT) also outperforms FairFed and FedAvg in weighted averaged fairness improvement over all heterogeneity settings, even though it does not provide as satisfactory EOD reduction as PP. Final layer fine-tuning (FT) works better with more heterogeneous data, providing a fairness improvement of 12% with $\alpha = 0.5$, and a slightly lower improvement of 7% with $\alpha = 500$. However, it may be noted that FT also provides an increase of BA across all heterogeneity settings while reducing EOD. While FT overall does not appear satisfactory in our experiments for fairness, we leave it for future work to explore whether different hyper-parameter setting (e.g. for α_{ft}, lr, etc) can produce better results.

6.4 Performance on NIH Chest X-Ray Dataset

Table 4 shows the experiment results on the NIH Chest X-Ray dataset for different heterogeneity settings.

In general, post-training intervention with both PP and FT still provide a fairness improvement across all heterogeneity settings which decreases with the decrease of data heterogeneity, from around 36% to 26% and from 39% to 8% respectively. And the accuracy decrease is small as well. FairFed provides a higher fairness EOD improvement on average than post-training interventions while providing a slightly lower accuracy.

Compared with the results on other datasets we discussed so far, we can see that the best performance of different metrics are more unevenly distributed across different methods. For example, looking at weighted average fairness, FairFed appears to perform better. For client-wise fairness metrics, client 2 can achieve the lowest EOD with FT for $\alpha = 0.5$ and $\alpha = 500$ while client 4 can achieve that with PP instead. Thus, the winner trend is not as clear as it was in the case of previous datasets.

The difference between results on the NIH-Chest X-Ray dataset and the other datasets could be because the hyper-parameters are more sensitive on NIH Chest X-Ray dataset and thus requires more careful tuning and a larger search space to obtain the optimal model. This sensitivity may be due to many reasons. Firstly, we see from the tables that the EOD metrics on NIH Chest X-Ray obtained by FedAvg training without any fairness constraints are already small. Recall that EOD of FedAvg on COMPAS and PTB-XL is generally larger than 0.3, on Adult is more than 0.16, while on NIH Chest X-Ray is around 0.05 - 0.06. Secondly, recall from Table 8 (Appendix 4.1) that NIH Chest X-Ray dataset is very imbalanced with only 18% of the samples having positive labels ("Effusion"). After data is split across clients, the local sample size for certain groups could be very small which makes it more difficult to make fine-tuning work. These results highlight some of the limitations of the post-training intervention that we now summarize as we approach the end of our paper.

6.5 Summary of Observed Strengths and Limitations

Model Output Post-Processing: In the analyses of post-training intervention with PP (i.e. model output post-processing) method, we observed several of its strengths. Firstly, this method shows its effectiveness with significant fairness improvement across all datasets under various heterogeneity settings. It is especially effective when the original model or data contains relatively large EOD or when local datasets are highly heterogeneous. Secondly, this method is very efficient to apply because it is very fast to compute the derived predictor and requires no hyperparameter tuning. Additionally, it requires minimal computational resources at local clients, making it a low-cost solution (i.e. no GPUs required). On the negative side, we observed that PP comes with drop in accuracy in nearly all cases. While a drop in accuracy is commonly observed in the fairness literature and may often be due to data characteristics and assumptions [12, 22], but it also shows the importance

NIH-Chest X-Ray Dataset

α	Metric	Client	Method			
			FedAvg	FairFed	PP	FT
0.5	BA	Avg	0.846	0.840	0.844	**0.850**
		C1	**0.916**	0.900	0.915	0.911
		C2	0.577	**0.584**	0.573	0.574
		C3	**0.962**	0.961	0.960	0.960
		C4	0.888	0.870	0.885	**0.921**
	EOD	Avg	0.061	**0.031**	0.039	0.037
		C1	0.106	**0.040**	0.090	0.086
		C2	0.013	0.008	0.011	**0.004**
		C3	0.044	0.035	0.044	**0.007**
		C4	0.098	0.042	**0.010**	0.075
5	BA	Avg	**0.867**	0.862	0.864	0.865
		C1	**0.844**	0.839	0.839	**0.844**
		C2	0.892	**0.899**	0.893	0.889
		C3	**0.836**	0.831	0.832	0.830
		C4	0.890	0.871	0.887	**0.892**
	EOD	Avg	0.051	**0.031**	0.045	0.042
		C1	0.035	0.061	**0.015**	0.026
		C2	0.056	**0.019**	0.080	0.041
		C3	0.067	**0.023**	0.049	0.067
		C4	0.046	**0.025**	0.029	0.031
500	BA	Avg	**0.865**	**0.865**	0.862	0.864
		C1	0.860	**0.861**	0.857	0.857
		C2	**0.879**	0.873	0.877	0.878
		C3	0.860	0.861	0.858	**0.865**
		C4	0.860	**0.864**	0.856	0.859
	EOD	Avg	0.067	**0.024**	0.049	0.061
		C1	0.103	**0.038**	0.069	0.104
		C2	0.035	0.007	**0.006**	**0.006**
		C3	0.098	**0.011**	0.098	0.104
		C4	0.035	0.039	**0.024**	0.036

Table 4: Client-wise performance comparison on NIH Chest X-Ray dataset with different data heterogeneity level α. BA: Balanced accuracy (Higher is better). EOD: Equalized odds difference (Lower is better). Grey highlights indicate the best performance of each row. FedAvg and FairFed are methods described in Section 4.3. PP and FT refer to post-training intervention with model output post-processing and final-layer fine-tuning respectively.

of considering application and context in employing any fairness intervention. Application grounded discussions with ethical, legal, domain experts and various stakeholders must be taken into account to select the right fairness intervention.

Final Layer Fine-Tuning: In the analysis of post-training intervention with FT (i.e. fair final layer fine-tuning) method, we found it to be also effective in improving model fairness. It improves the EOD under all heterogeneity settings for most datasets. Similar to PP, this approach tends to achieve better fairness improvement with more heterogeneous data across clients. In contrast to PP, fine-tuning also provides a better model accuracy under more heterogeneous settings on most datasets. Notably, in cases of extreme data heterogeneity, fine-tuning often improves both fairness and accuracy simultaneously. While it does not always deliver as large a fairness boost as PP, it delivers fairness improvements with minimal or no impact on model performance.

Though it is not as computationally inexpensive as PP, FT is still efficient because only the last layer of the model is updated, while the other layers are kept fixed. This ensures that even with large models like the deep neural network (DNN) we used in the NIH Chest X-Ray experiments, the method is promising. Local clients can also tune their local model to suit their specific fairness requirements by tweaking parameters such as α_{ft} and the number of local training rounds. This provides a more flexible fairness option for clients compared to the PP approach.

However, this dependence on the fine-tuning parameter α_{ft} also means that achieving the optimal performance requires tuning the hyperparameters, which adds to the cost and complexity of applying it, especially on large datasets with increasing complexity. Finally, in cases where local datasets are small and thus, there are insufficient samples for certain labels/groups, fine-tuning tends to not work satisfactorily and hyperparameter tuning more difficult.

7 CONCLUSIONS AND FUTURE WORK

This paper argued for a shift in how fairness is conceptualized in federated learning. Rather than treating fairness as a single requirement to be satisfied by a global model, we highlight that fairness in practice is diverse, situated, and often contested; consequently, federated systems should not presume that one fairness definition, metric or constraint is appropriate for all clients.

This paper does not claim to offer a complete solution; instead, it demonstrates that existing global approaches and a simple new baseline (i.e. local post-training interventions) address only parts of a broader challenge. While post-training interventions offer flexibility and can be applied independently by each client, they also have limitations. Output post-processing often improves fairness with significant reduction in accuracy. Final-layer fine-tuning can achieve modest fairness gains without as much accuracy loss, but its performance is sensitive to hyperparameters and deteriorates when clients have limited local data. Both techniques rely heavily on the quality of the global model they receive. These observations indicate that fairness adapted only after global training, although useful, cannot fully address the broader misalignment between global model training and client-specific fairness needs.

Future work should investigate how training procedures can allow clients to express different fairness goals and explore how such goals can coexist without forcing a single fairness outcome. There is also room for work on robustness under small or imbalanced local datasets, better support for multiple fairness notions beyond group metrics, and extending contextualized fairness to multi-class prediction tasks. Finally, understanding how fairness interacts with domain requirements, legal obligations, and operational constraints across diverse clients remains an open and important challenge.

REFERENCES

[1] 2024. Python Imaging Library. https://en.wikipedia.org/w/index.php?title=Python_Imaging_Library&oldid=1225854061 Page Version ID: 1225854061.

[2] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. 2020. Mitigating Bias in Federated Learning. https://doi.org/10.48550/arXiv.2012.02447 arXiv:2012.02447 [cs, stat].

[3] Durmus Alp Emre Acar, Yue Zhao, Ruizhao Zhu, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. 2021. Debiasing Model Updates for Improving Personalized Federated Training. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 21–31. https://proceedings.mlr.press/v139/acar21a.html ISSN: 2640-3498.

[4] Abien Fred Agarap. 2019. Deep Learning using Rectified Linear Units (ReLU). https://doi.org/10.48550/arXiv.1803.08375 arXiv:1803.08375 [cs, stat].

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

[6] Ronny Kohavi Barry Becker. 1996. Adult. https://doi.org/10.24432/C5XW20

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://doi.org/10.48550/arXiv.1810.01943 arXiv:1810.01943 [cs].

[8] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*. 3121–3124. https://doi.org/10.1109/ICPR.2010.764 ISSN: 1051-4651.

[9] Xin Che, Jingdi Hu, Zirui Zhou, Yong Zhang, and Lingyang Chu. 2024. Training Fair Models in Federated Learning without Data Privacy Infringement. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 7687–7696.

[10] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. 2020. Fairness-aware Agnostic Federated Learning. https://doi.org/10.48550/arXiv.2010.05057 arXiv:2010.05057 [cs].

[11] Yuying Duan, Yijun Tian, Nitesh Chawla, and Michael Lemmon. 2024. Post-Fair Federated Learning: Achieving Group and Community Fairness in Federated Learning via Post-processing. *arXiv preprint arXiv:2405.17782* (2024).

[12] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*. PMLR, 2803–2813.

[13] Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A. Salman Avestimehr. 2023. FairFed: Enabling Group Fairness in Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 6 (June 2023), 7494–7502. https://doi.org/10.1609/aaai.v37i6.25911 Number: 6.

[14] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

[15] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. https://doi.org/10.48550/arXiv.1610.02413 arXiv:1610.02413 [cs].

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity Mappings in Deep Residual Networks. https://doi.org/10.48550/arXiv.1603.05027 arXiv:1603.05027 [cs].

[17] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C. Shen, George Shih, Ronald M. Summers, Yifan Peng, and Zhangyang Wang. 2022. Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study. In *Data Augmentation, Labelling, and Imperfections*, Hien V. Nguyen, Sharon X. Huang, and Yuan Xue (Eds.). Springer Nature Switzerland, Cham, 22–32. https://doi.org/10.1007/978-3-031-17027-0_3

[18] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[19] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. https://doi.org/10.48550/arXiv.1502.03167 arXiv:1502.03167 [cs].

[20] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning* 14, 1–2 (2021), 1–210.

[21] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/arXiv.1412.6980 arXiv:1412.6980 [cs].

[22] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[23] Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. 2020. A review of applications in federated learning. *Computers & Industrial Engineering* 149 (2020), 106854.

[24] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. 2021. Ditto: Fair and Robust Federated Learning Through Personalization. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 6357–6368. https://proceedings.mlr.press/v139/li21h.html ISSN: 2640-3498.

[25] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2020. Fair Resource Allocation in Federated Learning. https://doi.org/10.48550/arXiv.1905.10497 arXiv:1905.10497 [cs, stat].

[26] Emilly M. Lima, Antônio H. Ribeiro, Gabriela M. M. Paixão, Manoel Horta Ribeiro, Marcelo M. Pinto-Filho, Paulo R. Gomes, Derick M. Oliveira, Ester C. Sabino, Bruce B. Duncan, Luana Giatti, Sandhi M. Barreto, Wagner Meira Jr, Thomas B. Schön, and Antonio Luiz P. Ribeiro. 2021. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature Communications* 12, 1 (Aug. 2021), 5117. https://doi.org/10.1038/s41467-021-25351-7 Publisher: Nature Publishing Group.

[27] Caio B. S. Maior, João M. M. Santana, Isis D. Lins, and Márcio J. C. Moura. 2021. Convolutional neural network model based on radiological images to support COVID-19 diagnosis: Evaluating database biases. *PLOS ONE* 16, 3 (March 2021), e0247839. https://doi.org/10.1371/journal.pone.0247839 Publisher: Public Library of Science.

[28] Disha Makhija, Xing Han, Joydeep Ghosh, and Yejin Kim. 2024. Achieving fairness across local and global models in federated learning. *arXiv preprint arXiv:2406.17102* (2024).

[29] Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. https://doi.org/10.48550/arXiv.2304.07288 arXiv:2304.07288 [cs, stat].

[30] Yuzhen Mao, Zhun Deng, Huaxiu Yao, Ting Ye, Kenji Kawaguchi, and James Zou. 2023. Last-Layer Fairness Fine-tuning is Simple and Effective for Neural Networks. https://doi.org/10.48550/arXiv.2304.03935 arXiv:2304.03935 [cs].

[31] Mattu, Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[32] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html ISSN: 2640-3498.

[33] Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. 2022. Local Learning Matters: Rethinking Data Heterogeneity in Federated Learning. 8397–8406. https://openaccess.thecvf.com/content/CVPR2022/html/Mendieta_Local_Learning_Matters_Rethinking_Data_Heterogeneity_in_Federated_Learning_CVPR_2022_paper.html

[34] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. 2019. Agnostic Federated Learning. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 4615–4625. https://proceedings.mlr.press/v97/mohri19a.html ISSN: 2640-3498.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://doi.org/10.48550/arXiv.1912.01703 arXiv:1912.01703 [cs, stat].

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[37] Mana Saleh Al Reshan, Kanwarpartap Singh Gill, Vatsala Anand, Sheifali Gupta, Hani Alshahrani, Adel Sulaiman, and Asadullah Shaikh. 2023. Detection of Pneumonia from Chest X-ray Images Utilizing MobileNet Model. *Healthcare* 11, 11 (Jan. 2023), 1561. https://doi.org/10.3390/healthcare11111561 Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[38] Sebastian Ruder. 2017. An overview of gradient descent optimization algorithms. https://doi.org/10.48550/arXiv.1609.04747 arXiv:1609.04747 [cs].

[39] Teresa Salazar, Helder Araújo, Alberto Cano, and Pedro Henriques Abreu. 2024. A Survey on Group Fairness in Federated Learning: Challenges, Taxonomy of Solutions and Directions for Future Research. *arXiv preprint arXiv:2410.03855* (2024).

[40] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. https://doi.org/10.48550/arXiv.1801.04381 arXiv:1801.04381 [cs].

[41] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[42] Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. 2021. Personalized Federated Learning using Hypernetworks. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 9489–9502. https://proceedings.mlr.press/v139/shamsian21a.html ISSN: 2640-3498.

[43] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[44] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. 2022. PTB-XL, a large publicly available electrocardiography dataset. https://doi.org/10.13026/KFZX-AW45
[45] Tongnian Wang, Kai Zhang, Jiannan Cai, Yanmin Gong, Kim-Kwang Raymond Choo, and Yuanxiong Guo. 2024. Analyzing the Impact of Personalization on Fairness in Federated Learning for Healthcare. *Journal of Healthcare Informatics Research* 8, 2 (June 2024), 181–205. https://doi.org/10.1007/s41666-024-00164-7
[46] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3462–3471. https://doi.org/10.1109/CVPR.2017.369 arXiv:1705.02315 [cs].
[47] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Trong Nghia Hoang, and Yasaman Khazaeni. 2019. Bayesian Nonparametric Federated Learning of Neural Networks. https://doi.org/10.48550/arXiv.1905.12022 arXiv:1905.12022 [cs, stat].
[48] Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. A Geometric Solution to Fair Representations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020). https://dl.acm.org/doi/abs/10.1145/3375627.3375864
[49] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. 1171–1180. https://doi.org/10.1145/3038912.3052660 arXiv:1610.08452 [cs, stat].
[50] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. 2022. Improving Fairness via Federated Learning. https://doi.org/10.48550/arXiv.2110.15545 arXiv:2110.15545 [cs].
[51] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems* 216 (2021), 106775.
[52] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. 2020. 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)*. 1051–1060. https://doi.org/10.1109/BigData50022.2020.9378043

A PSEUDO CODE

See Algorithms 1 and 2 on the next page.

B DATASETS DETAILS

Adult Dataset. The target label in the Adult dataset [6] is "income" which is divided into two classes ("<=50K" and ">50K") for binary classification task. We use "sex" as the sensitive attribute with "male" as 1 and "female" as 0 in our experiments. We use one-hot encoding for all categorical features and apply `sklearn StandardScaler` [36] to standardize continuous features. The dataset originally consisted of 48842 samples. After dropping N/A values including values filled with abnormal values like '?', we have 44993 samples with 14662 (32.6%) samples being "female" and 30331 (67.4%) being "male". Other statistics of Adult dataset are in Table 5.

		income		
		1 (>50K)	0 (<=50K)	Total
sex	1 (Male)	9343 (20.8%)	20988 (46.6%)	30331 (67.4%)
	0 (Female)	1636 (3.6%)	13026 (29.0%)	14662 (32.6%)
	Total	10979 (24.4%)	34014 (75.6%)	44993 (100.0%)

Table 5: Statistics for the Adult dataset. *income* is used as target label, and *sex* is used as sensitive attribute.

COMPAS Dataset. Target variable in this dataset [31] is observed recidivism within 2 years. The dataset originally consisted of 7214 samples, and after dropping N/A values there are 6172 samples left. We use *race* as the sensitive attribute. For data pre-processing, we followed the procedures in [31] and [49]. Samples with the race "Caucasian" are categorized as the privileged group and samples with the race "African-American" as the unprivileged group. Samples in other race groups are dropped for the scope of experiments. 5278 samples are used for the experiments with 2103 samples (39.8%) being "Caucasian" and 3175 samples being (60.2%) "African-American". We only use a subset of the data features including "*age_cat*", "*sex*", "*priors_count*", "*c_charge_degree*", as well as sensitive attribute "*race*" and target label "*two_year_recid*". Similar to Adult data, we use one-hot encoding for categorical features and standardize numerical features using `sklearn StandardScaler`. Statistics of the COMPAS dataset in Table 6.

		two_year_recid		
	Total	1	0	Total
race	1 (Caucasian)	822 (15.5%)	1281 (24.3%)	2103 (39.8%)
	0 (African-American)	1661 (31.5%)	1514 (28.7%)	3175 (60.2%)
	Total	2483 (47.0%)	2795 (53.0%)	5278 (100.0%)

Table 6: Statistics for COMPAS dataset. *two_year_recid* is used as target label, and *race* is used as sensitive attribute.

PTB-XL ECG Signal Dataset. The PTB-XL dataset [44] is an electrocardiography (ECG) dataset consisting of 21799 clinical ECG signal records of 10-second length from 18869 patients. The ECGs are 12-lead with a sampling frequency of 100Hz. The 12 standard leads recorded in the dataset are lead I, II, III, aVF, aVR, aVL, V1, V2, V3, V4, V5, and V6, and each of them records the heart's electrical activity from a distinct viewpoint. Together, they provide a multidimensional and comprehensive view of the heart's electrical activity, which makes ECG very commonly used and powerful tool in cardiology. Each ECG record is annotated by one or two cardiologists and assigned a statement to diagnose if the patient has a normal ECG or certain heart disease. The likelihoods for each diagnostic statement are also provided. For pre-processing, we dropped N/A values as well as entries with abnormal ages like "300". Additionally, we only used samples with 100% confidence diagnosis for the experiments. We consider the target label as "normal" (ECG), and ECGs labeled with different heart diseases are considered as "abnormal". We use patient age as the sensitive attribute and patients with age larger than 60 are classified into one group and the rest as the other group. In the experiments, a total of 15766 samples are used with 5971 (37.9%) classified as "normal". Summary of PTB-XL in Table 7.

		normal		
	Total	1	0	Total
$age > 60$	1	1987 (12.6%)	6717 (42.6%)	8704 (55.2%)
	0	3984 (25.3%)	3078 (19.5%)	7062 (44.8%)
	Total	5971 (37.9%)	9795 (62.1%)	15766 (100.0%)

Table 7: Statistics for PTB-XL dataset. *normal* (ECG) is used as target label, and $age > 60$ is used as sensitive attribute.

Algorithm 1 FL Model Output Fairness Post-Processing

Initialize server with global model weights ω_0; K clients with local training dataset D_k

for each global round t=1, 2, ..., T **do** ▷ FedAvg starts
  for each client k=1, 2, ..., K in parallel **do**
    $\omega_t^k \leftarrow$ ClientLocalUpdate(ω_{t-1}, D_k) ▷ Compute local update at each client using data D_k
    CommunicateToServer(ω_t^k) ▷ Communicate local update to server
  end for
  $\omega_t \leftarrow$ Aggregate$(\{\omega_t^k\}_{k=1}^K)$ ▷ Aggregate local updates to compute global update at server
  CommunicateToClients(ω_t)
end for
for each client k=1, 2, ..., K in parallel **do** ▷ Post-processing starts
  $\hat{Y}_k \leftarrow$ Predict(ω_T, D_k) ▷ Compute prediction with FedAvg model for local data D_k
  $p_k \leftarrow$EqOdds$(\hat{Y}_k, D_k)$ ▷ Compute derived predictor for client
end for

Algorithm 2 FL Model Final Layer Fairness Fine-Tuning

Initialize global model with L layers and initial weights ω_0; K clients with local training dataset D_k; accuracy-based loss function l; fairness-based loss function l'; fine-tuning parameter α; fairness metrics F_k for each client k; fine-tuning learning rate η

for each global round t=1, 2, ..., T **do** ▷ FedAvg starts
  for each client k=1, 2, ..., K in parallel **do**
    $\omega_t^k \leftarrow$ ClientLocalUpdate(ω_{t-1}, D_k) ▷ Compute local update at each client using data D_k
    CommunicateToServer(ω_t^k) ▷ Communicate local update to server
  end for
  $\omega_t \leftarrow$ Aggregate$(\{\omega_t^k\}_{k=1}^K)$ ▷ Aggregate local updates to compute global update at server
  CommunicateToClients(ω_t)
end for
for each client k=1, 2, ..., K in parallel **do** ▷ Fine-tuning starts
  $\omega'^k \leftarrow$ FreezeLayers$(\{\omega_T\}_{l=1}^{L-1})$ ▷ Freeze weights for layer from 1 to (L-1)
  for each fine-tuning round r=1, 2, ..., R **do**
    $L = \alpha l(\omega'^k, D_k) + l'(\omega'^k, D_k, F_k)$ ▷ local weighted loss L including fairness F_k
    $\omega'^k \leftarrow \omega'^k - \eta \nabla_{\omega'^k}(L)$ ▷ Local update at client
  end for
end for

NIH Chest X-Ray Image Dataset. NIH Chest X-Ray dataset [46] is a medical imaging dataset comprised of 112120 X-ray images from 30805 patients. The X-ray images come labeled with up to 14 diseases and "No finding" by natural language processing (NLP) models based on the original radiological reports of each X-ray. We only select samples with "No Findings" and disease "Effusion" for the scope of our experiments. We use "Effusion" as the target label, and "Patient gender" as the sensitive attribute. After removing entries filled with N/A and abnormal values, we have a dataset with 73669 samples, and 13316 (18.1%) are labeled "Effusion". In addition, we resized each image into size (256 * 256 * 3) with 3 channels both for computational reasons and the requirement for using pre-trained models during the training process (to be discussed in Section 5). Images are also normalized using the required mean and standard deviation based on the pre-trained model used in the experiments [40]. Statistics of NIH Chest X-Ray are in Table 8.

In the NIH Chest X-Ray dataset, the feature *gender* refers to the biological sex of patients. Here we just kept the original feature name in the dataset, which is *gender*.

		Effusion		
	Total	1	0	Total
gender	1	7434 (10.1%)	33916 (46.0%)	41350 (56.1%)
	0	5882 (8.0%)	26437 (35.9%)	32319 (43.9%)
	Total	13316 (18.1%)	60353 (81.9%)	73669 (100.0%)

Table 8: Statistics for NIH-Chest X-Ray dataset. *Effusion* is used as target label and *gender* as sensitive attribute.

C EVALUATION METRICS

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of total samples}} = \frac{\text{TP} + \text{TN}}{(\text{TP+FN+FP+TN})} \tag{4}$$

$$\text{Balanced Accuracy (BA)} = \frac{\text{Sensitivity (TPR)} + \text{Specificity (TNR)}}{2} \tag{5}$$

We calculate the weighted average of x across clients as:

$$\text{Weighted Average} = \sum_{k=1}^{K}\left(\frac{n_k x_k}{\sum_{k=1}^{K} n_k}\right) \tag{6}$$

where n_k denotes the number of samples on client k and x_k denotes the local measure of x on client k.

D IMPLEMENTATION DETAILS

D.1 Hyperparameters

The hyperparameters we used for methods on different datasets can be found in Table 9. Note that although we only listed three methods in the table (FedAvg, FairFed and FT), all the methods used in the experiments are covered. The FL training stage of post-training with PP (output post-processing) and FT (fine-tuning) are the same as FedAvg, and PP does not require hyperparameter tuning. FairFed/FR uses the same parameters as FairFed and doesn't need extra hyperparameter tuning.

Dataset	FedAvg		FairFed			FT		
	lr	bs	lr	bs	β	lr	bs	α_{ft}
Adult	0.01	32	0.01	32	0.1	5e-3	256	1.0
COMPAS	0.01	32	0.01	32	0.5	5e-3	256	2.0
PTB-XL	5e-3	32	5e-3	32	0.1	5e-3	512	1.0
NIH-Chest	1e-4	64	1e-4	64	0.1	1e-4	64	0.1

Table 9: Hyperparameters used in the experiments for different methods on all the datasets. lr: learning rate. bs: local batch size. β: fairness budget in FairFed. α_{ft}: fine-tuning parameter in FT method. Both methods in post-training intervention use the same parameters as FedAvg during the FL training stage. FT denotes the final-layer fine-tuning stage of post-training intervention.

D.2 Libraries and Computational Resources

We use Python as the programming language and PyTorch [35] for the machine learning model development and experiments. We also use IBM AIF 360 [7] for the output post-processing approach and for fairness evaluation. Besides, libraries including sklearn, numpy, pandas, Pillow [1], and H5py are also used in the experiments for data processing and model training. For training on signal and image datasets, we use NVIDIA Tesla V100 for GPU acceleration.