Distribution-Free Fair Federated Learning with Small Samples

Abstract

As federated learning gains increasing importance in real-world applications due to its capacity for decentralized data training, addressing fairness concerns across demographic groups becomes critically important. However, most existing machine learning algorithms for ensuring fairness are designed for centralized data environments and generally require large-sample and distributional assumptions, underscoring the urgent need for fairness techniques adapted for decentralized and heterogeneous systems with small-sample and distribution-free guarantees. To address this issue, this paper introduces FedFaiREE, a post-processing algorithm developed specifically for distribution-free fair learning in decentralized settings with small samples. Our approach accounts for unique challenges in decentralized environments, such as client heterogeneity, communication costs, and small sample sizes. We provide rigorous theoretical guarantees for both fairness and accuracy, and our experimental results further provide robust empirical validation for our proposed method.

Keywords: Federated learning; Algorithmic Fairness; Distribution-free
Mathematics Subject Classification (2020): 62H30, 62R07

1 Introduction

Federated learning (FL) enables collaborative model training across multiple clients without requiring data centralization (McMahan et al., 2017). This paradigm has become increasingly important in applications involving sensitive data, such as healthcare (Joshi et al., 2022; Antunes et al., 2022) and mobile systems (Li et al., 2020; Yang et al., 2021). As FL systems are deployed in high-stakes settings, ensuring *algorithmic fairness* across demographic groups has emerged as a critical concern.

Despite extensive progress in fairness-aware machine learning, most existing methods are designed for centralized settings and rely on large-sample or distributional assumptions. Directly applying these approaches in federated environments is nontrivial and often leads to

degraded performance or excessive communication overhead. Moreover, the decentralized nature of FL introduces additional challenges, including client heterogeneity, limited local sample sizes, and restricted data sharing, all of which complicate the enforcement of fairness constraints.

Recent works have attempted to address algorithmic fairness in the FL setting, including FairFed (Ezzeldin et al., 2023), FedFB (Zeng et al., 2021), FCFL (Cui et al., 2021), and AgnosticFair (Du et al., 2021). These methods typically enforce fairness by modifying local training objectives or adjusting aggregation weights. However, they suffer from two fundamental limitations. First, fairness is primarily enforced at the local level, while achieving *global* fairness is inherently challenging: fairness at the client level does not necessarily translate to fairness at the population level (Hamman and Dutta, 2023). Second, most existing approaches rely on asymptotic guarantees or fail to provide fairness guarantees in a *distribution-free* manner where no distributional assumptions are imposed, limiting their applicability in realistic scenarios with small and heterogeneous datasets.

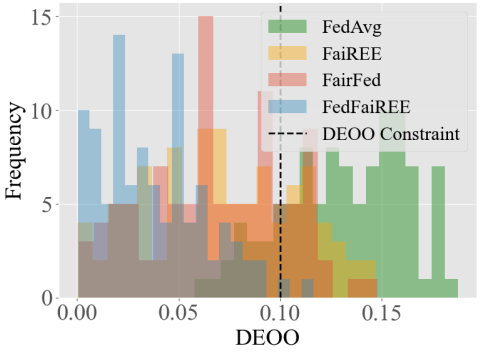

Figure 1: The distribution of $|DEOO|$ (defined in equation 2) for different methods on the Adult dataset (Dua et al., 2017). See Section 5 for details.

FaiREE (Li et al., 2022) is, to the best of our knowledge, the first method that provides group fairness guarantees that are both distribution-free and small-sample that ensures fairness with arbitrarily finite samples. However, it is restricted to centralized i.i.d. data and does not address the distinctive challenges of FL, such as decentralized data, communication constraints, and client heterogeneity. In heterogeneous FL settings, even if all training data are centralized, FaiREE can still suffer from bias due to cross-client distributional differences. These challenges are particularly pronounced in applications such as healthcare, where privacy regulations prevent data sharing and each institution only has access to a small, siloed dataset. This gap motivates the need for a method that achieves small-sample, distribution-free fairness guarantees while explicitly accounting for decentralization and heterogeneity.

To address these challenges, we propose FedFaiREE, a general post-processing framework for achieving *small-sample* and *distribution-free* fairness in federated learning. Our key insight is that controlling group fairness can be reduced to aligning order statistics of score distributions across groups, even under heterogeneous and decentralized data. This perspective enables us to transform fairness constraints into a rank-based selection problem, leading to both theoretical guarantees and practical efficiency. Our framework applies to a broad class of group fairness notions, including Equality of Opportunity (Hardt et al., 2016), Equalized Odds (Hardt et al., 2016), Demographic Parity (Agarwal et al., 2018), Predictive Equality (Hardt et al., 2016), and Overall Accuracy Equality (Zafar et al., 2017). The method is designed for realistic decentralized settings with heterogeneous client distributions, limited sample sizes, and communication constraints.

The key idea of FedFaiREE is to leverage distributed order statistics to construct a set of candidate classifiers that satisfy fairness constraints with high probability, and then select the

most accurate classifier within this feasible set. This formulation converts global fairness control into a rank-based problem across clients, enabling principled handling of heterogeneity and approximate local rank estimation under communication constraints. Importantly, this perspective decouples fairness enforcement from model training, leading to a flexible and theoretically grounded solution.

Our contributions are threefold:

1. We introduce FedFaiREE, a unified post-processing framework that takes any black-box classifiers as input and enforces group fairness constraints in federated learning under finite samples, client heterogeneity, and multiple protected groups. To the best of our knowledge, this is the first framework that provides distribution-free, small-sample fairness guarantees in the federated setting.

2. We establish rigorous guarantees showing that FedFaiREE satisfies fairness constraints with high probability in a distribution-free manner with arbitrarily finite sample, and achieves near-optimal accuracy under mild conditions on the input scorer.

3. Empirical validation. Through extensive experiments (Figure 1), we demonstrate that existing federated fairness methods often fail to control fairness under small-sample or heterogeneous settings, while FedFaiREE consistently satisfies the prescribed fairness constraints with competitive accuracy.

1.1 Additional Related Work

Fairness in federated learning has been studied from two complementary perspectives: fairness across clients and fairness across demographic groups. The former focuses on equitable performance or contribution across clients (Li et al., 2021; Lyu et al., 2020; Yu et al., 2020; Huang et al., 2020), while the latter aims to ensure equitable treatment across sensitive attributes such as race or gender, commonly referred to as *group fairness* (Dwork et al., 2012).

Existing Group Fairness Techniques. Existing approaches to group fairness can be approximately divided into three categories (Caton and Haas, 2020): pre-processing methods that directly perform debiasing on input data (Zemel et al., 2013; Johndrow and Lum, 2019); in-processing methods that incorporate fairness metrics into model training as part of the objective function (Goh et al., 2016; Cho et al., 2020); post-processing methods that adjust model outputs to enhance fairness (Li et al., 2022; Zeng et al., 2022; Fish et al., 2016). Our method falls into the post-processing category. Among these works, FaiREE (Li et al., 2022) is the most closely related, as it provides finite-sample, distribution-free fairness guarantees in the centralized setting. However, extending such guarantees to federated learning is nontrivial. In federated settings, optimizing local fairness objectives does not generally ensure global fairness(Hamman and Dutta, 2023). Instead, fairness must be enforced at the level of the global population, while the data remain distributed across clients with heterogeneous local distributions, limited local sample sizes, and communication constraints. To address this, we develop a distributed order statistics formulation that links global fairness control to client-level rank information, and further incorporate communication-efficient local rank estimation into the analysis. This yields finite-sample, distribution-free fairness guarantees in decentralized settings where FaiREE is not

directly applicable. In addition, our proposed method allows client correlation, while FaiREE requires independence among training samples. Our method is also applicable to a broader range of scenarios than FaiREE, including settings with label shift. See a more detailed discussion in Section D of the Appendix.

Group Fairness Approaches in Federated Learning. In recent years, there has been a growing amount of work focusing on group fairness in the context of Federated Learning (Ezzeldin et al., 2023; Cui et al., 2021; Zeng et al., 2021; Du et al., 2021; Rodríguez-Gálvez et al., 2021; Chu et al., 2021; Liang et al., 2020; Hu et al., 2022; Papadaki et al., 2022). Most of these studies aim to either introduce fairness principles into the local updates, adapt conventional fairness methods, or perform reweighting during aggregation, or a combination of these strategies. Specifically, Du et al. (2021) proposed AgnosticFair, a framework that utilizes kernel reweighing functions to adjust items in local objective functions, including both loss terms and fairness constraints. Zeng et al. (2021) introduced FedFB, a method that adapts Fair Batch, a centralized technique designed to improve fairness among groups by reweighting loss terms for different subgroups, for the FL setting. Ezzeldin et al. (2023) proposed FairFed, an approach that adjusts aggregate weights by considering the disparities between local fairness metrics and the global fairness metric in each training round.

2 Preliminaries

In this paper, we address the problem of predicting a binary label, denoted by Y, using a set of features. The features are divided into two categories: X and A. Here, $X \in \mathcal{X}$ represents non-sensitive features, while $A \in \mathcal{A} = \{0, 1, \cdots, A_0\}$ corresponds to sensitive features. A data point includes (x, y, a), which corresponds to (X, Y, A). For simplicity, we first introduce the concept of *Score-based classifier* (Chen et al., 2018; Zafar et al., 2019).

Definition 2.1. (Score-based classifier) A score-based classifier is an indication function $\hat{Y} = \phi(x, a) = \mathbb{1}\{f(x, a) > c\}$ for a measurable score function $f : \mathcal{X} \times \mathcal{A} \to [0, 1]$ and a constant threshold $c > 0$.

To assess the fairness of the classifier, several group fairness notions have been proposed in the literature. In the following, for illustration, we introduce two commonly used notions in the fairness literature, Equality of Opportunity (EOO) and Equalized Odds (EO), aiming to equalize true positive rate or false positive rate or both.

Definition 2.2. (Equality of Opportunity (Hardt et al., 2016)) A classifier satisfies Equality of Opportunity if it satisfies the same true positive rate among protected groups: $\mathbb{P}_{X|A=a,Y=1}(\widehat{Y} = 1) = \mathbb{P}_{X|A=0,Y=1}(\widehat{Y} = 1)$, where $a \in \{1, \cdots, A_0\}$.

Definition 2.3. (Equalized Odds (Hardt et al., 2016)) A classifier satisfies Equalized Odds if it satisfies the following equality: $\mathbb{P}_{X|A=1,Y=1}(\widehat{Y} = 1) = \mathbb{P}_{X|A=0,Y=1}(\widehat{Y} = 1)$ and $\mathbb{P}_{X|A=1,Y=0}(\widehat{Y} = 1) = \mathbb{P}_{X|A=0,Y=0}(\widehat{Y} = 1)$.

In practice, exact equality is often unattainable. Therefore, a tolerance parameter, denoted as α, is commonly introduced in Equality of Opportunity, as discussed in prior works (Zeng

et al., 2022; Li et al., 2022). To be more specific, given a classifier ϕ, the α difference tolerance in Equality of Opportunity within a binary group label can be defined as:

$$|\mathbb{P}_{X|A=1,Y=1}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=1}(\widehat{Y}=1)| \leq \alpha. \tag{1}$$

To be concise, in later sections, we use $DEOO$ to represent the left side of the inequality, i.e.,

$$DEOO = \mathbb{P}_{X|A=1,Y=1}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=1}(\widehat{Y}=1). \tag{2}$$

Similarly, the difference with respect to equalized odds can be defined as a two-dimensional vector

$$DEO = (\mathbb{P}_{X|A=1,Y=1}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=1}(\widehat{Y}=1), \mathbb{P}_{X|A=1,Y=0}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=0}(\widehat{Y}=1)). \tag{3}$$

We will extend our framework to multi-group settings in Section 6.2 later.

Additional Notation. To further simplify the formula in the article, we provide notations as follows: p_a signifies $P(A=a)$. $p_{Y,a}$ represents $P(Y=1 \mid A=a)$, and $q_{Y,a}$ is defined as $1-p_{Y,a}$. D and D_i represent the datasets for all clients and client i, respectively, where i belongs to the set $\{1,2,\ldots,S\}$. n denotes the size of dataset D. T represents the ordered scores of elements in dataset D. $D_i^{y,a}$ is used to denote the subset of dataset D_i where $Y=y$ and $A=a$. Similar notations apply to $T^{y,a}$ and $n^{y,a}$.

3 Fair Federated Learning Approach

In this section, we introduce FedFaiREE, a **Fed**erated Learning, **Fai**r, distribution-f**REE** algorithm. FedFaiREE is designed to ensure fairness under three key challenges: finite samples, distribution-free settings, and heterogeneous clients. Our key insight is that controlling group fairness can be reduced to aligning the *order statistics* (i.e., quantiles or ranks) of score distributions across sensitive groups. This perspective allows us to reformulate fairness constraints as a rank-based selection problem, which remains tractable even under heterogeneity and small samples.

To incorporate heterogeneity among clients, we adopt the following assumption.

Assumption 3.1. The training data points within the client i are drawn independently and identically (i.i.d) from distribution P_i, while the test data points are sampled from a global distribution that represents a mixture of $P_1,\cdots,P_S$ with weight $\{\pi_i\}_{i\in[S]} \in \Delta_S$. Specifically, we assume that

$$\left(X_k^i, Y_k^i\right) \sim P_i, \ \left(X^{\text{test}}, Y^{\text{test}}\right) \sim P^{mix} = \sum_{i=1}^{S} \pi_i P_i.$$

This implies that data points in client i are exchangeable. Specifically, we want to note that we do not make any assumptions among $P_1,\ldots,P_S$ here.

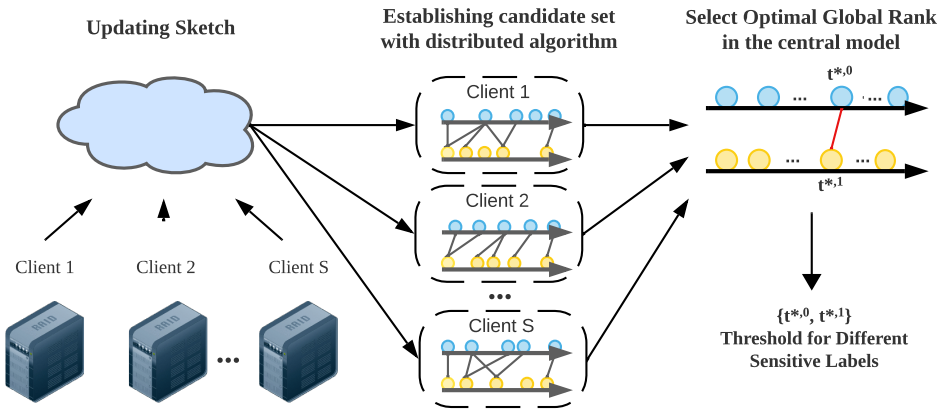

Figure 2: **Overview of FedFaiREE.** With S clients and a pre-trained model in consideration, each circle in the image symbolizes a datapoint score in the training set. The color of the circles represents different sensitive labels, while the gray edges depict local ranks of threshold pairs (each global classifier's threshold pair corresponds to S local ranks). Notably, the red edge signifies the chosen global classifier with thresholds $t^{*,0}, t^{*,1}$ for sensitive labels $A = 0$ and $A = 1$, respectively.

3.1 Problem formulation

Consider a scenario with S clients, each with a local dataset $D_i = \cup_{y\in\mathcal{Y},a\in\mathcal{A}} D_i^{y,a}$ and a pre-trained score-based classifier $\phi_0(x,a) = \mathbb{1}\{f(x,a) > c\}$. Here, $D_i^{y,a}$ denotes samples with label $Y = y$ and sensitive attribute $A = a$. Our goal is to construct a fair classifier of the form

$$\phi(x,a) = \mathbb{1}\{f(x,a) > \lambda_a\},$$

where λ_a is a group-specific threshold chosen to satisfy a fairness constraint such as $|DEOO| < \alpha$.

Piror work (Corbett-Davies et al., 2017; Menon and Williamson, 2018; Zeng et al., 2022) has shown that the classifier with optimal misclassification performance while adhering to specific fairness constraints requires group-wise thresholdings to the unconstrained Bayes-optimal classifier, we consider group-wise scores $t_{i,j}^{y,a} = f(x_{i,j}^{y,a})$ and denote the sorted scores on client i by $T_i^{y,a} = \{t_{i,1}^{y,a}, t_{i,2}^{y,a}, \cdots, t_{i,n_i^{y,a}}^{y,a}\}$. Let $T^{y,a}$ denote the global sorted scores aggregated across all clients. Instead of directly searching over thresholds, FedFaiREE takes a different perspective: we operate on the *ranks* of scores. On client i, this naturally leads us to the idea of transforming the problem of selecting optimal thresholds λ_a into determining the optimal "local ranks" $k_i^{1,a}$ of the score. However, as we concern about global fairness and misclassification error, we opt to seek the global rank $k^{1,a}$ (i.e., the rank in the sorted score set $T^{1,a}$), and $\phi(x,a) = \mathbb{1}\{f(x,a) > t_{(k^{1,a})}^{1,a}\}$. By mapping this to its corresponding "local ranks" $k_i^{1,a}$, we can leverage the properties of order statistics to ensure fairness under client heterogeneity. We will delve into the details of our approach and observations in the next subsection.

To this end, we present an overview of our algorithm in Figure 2, consisting of two main parts: 1). establishing a candidate set with a distributed algorithm that satisfies the fairness constraint with high probability, and 2). selecting the optimal rank pair to minimize the estimated misclassification error. In the next subsection, we first discuss a simple case: ensuring equality of opportunity under a binary-group and binary-label scenario, i.e., $\mathcal{Y} = \mathcal{A} = \{0,1\}$. As FedFaiREE is a general framework adaptable to various fairness notions and can accommodate even more diverse situations, we are going to study how to apply the method to equalized odds

in Section 3.4, and extend to other fairness notions in Section B in the appendix. The extension to the multi-group fairness scenario will be further studied in Section 6.2.

3.2 Candidate set construction with distributed quantile algorithm

We now describe how to construct a set of candidate rank pairs that satisfy the fairness notion equality of opportunity (EOO). Since EOO depends on differences in group-wise true positive rates, and these rates correspond to tail probabilities of score distributions, they can be controlled by the relative positions (ranks) of thresholds within each group. Therefore, fairness can be enforced by selecting rank pairs whose induced quantiles are sufficiently aligned across groups. To formalize this, we leverage the theory of order statistics. Specifically, we consider score sets that $k^{1,a}$ represents the rank in the sorted $T^{1,a}$. To account for heterogeneity among clients, we further introduce the notation $k_i^{1,a}$ to denote the corresponding rank of $t^{1,a}_{k^{1,a}}$ within the sorted set $T_i^{1,a}$, where $i \in [S]$ and $k_i^{1,a}$ satisfies $t^{1,a}_{i,(k_i^{1,a})} \leq t^{1,a}_{(k^{1,a})} < t^{1,a}_{i,(k_i^{1,a}+1)}$. For simplicity, we further define $\boldsymbol{k}^{1,a} = (k_1^{1,a}, \cdots, k_S^{1,a})$, and $Q(\alpha, \beta)$ represents independent variable following a Beta(α, β) distribution. We present the following observation regarding fairness control.

Proposition 3.2. *Under Assumption 3.1, for $a \in \{0,1\}$, consider $k^{1,a} \in \{1, \ldots, n^{1,a}\}$, the corresponding $k_i^{1,a}$ for $i \in [S]$ and the score-based classifier $\phi(x,a) = \mathbb{1}\{f(x,a) > t^{1,a}_{(k^{1,a})}\}$. Define*

$$h_{y,a}(\boldsymbol{u}, \boldsymbol{v}) = \mathbb{P}\Big(\sum_{i=1}^{S} \pi_i^{y,a} Q\left(u_i, n_i^{y,a} + 1 - u_i\right) - \sum_{i=1}^{S} \pi_i^{y,1-a} Q\left(v_i, n_i^{y,1-a} + 1 - v_i\right) \geq \alpha \Big). \tag{4}$$

Then we have:

$$\mathbb{P}(|DEOO(\phi)| > \alpha) \leq h_{1,0}(\boldsymbol{k}^{1,0} + \mathbf{1}, \boldsymbol{k}^{1,1}) + h_{1,1}(\boldsymbol{k}^{1,1} + \mathbf{1}, \boldsymbol{k}^{1,0}), \tag{5}$$

where $\pi_i^{1,a} = \mathbb{P}(x \text{ from client } i \mid x \text{ with } Y = 1, A = a)$.

This proposition enables us to select classifiers that satisfy fairness constraints with arbitrary finite samples and no distributional assumption. Moreover, $Q(\alpha, \beta)$ can be efficiently estimated by Monte Carlo simulations in applications. Specifically, we approximated $Q(\alpha, \beta)$ by conducting random sampling 1000 times in our experiment, yielding a highly satisfactory approximation.

Due to the need of computing local ranks to make use of Proposition 3.2, it is crucial to consider the tradeoff between accuracy and communication cost in real applications. We can adopt distributed quantile algorithms to reduce communication costs while controlling errors in calculating local ranks. Therefore, we present an alternative formulation of Proposition 3.2 to allow errors in the local rank calculation. To begin with, we introduce the concept of approximate quantiles and ranks (Luo et al., 2016; Lu et al., 2023).

Definition 3.3. (ε-approximate β-quantile and rank of a given set) For an error $\varepsilon \in (0,1)$, the ε-approximate β-quantile of a given set is any element with rank between $(\beta-\varepsilon)N$ and $(\beta+\varepsilon)N$, where N is the total number of elements in set. Further, the ε-approximate rank of an element in a given set is any rank between $(\beta-\varepsilon)N$ and $(\beta+\varepsilon)N$ where βN represents the real rank.

In other words, an ε-approximate rank or quantile may differ from its exact counterpart by at most an ε-fraction of the sample size, which provides a natural way to characterize the error

introduced by communication-efficient distributed rank estimation. Under Definition 3.3, if the rank estimation method produces ε-approximate ranks, it is possible to correspondingly modify Proposition 3.2.

Proposition 3.4. *Under Assumption 3.1, for $a \in \{0,1\}$, consider $k^{1,a} \in \{1,\dots,n^{1,a}\}$, the corresponding $\hat{k}_i^{1,a}$ for $i \in [S]$ which are ε-approximate ranks and the score-based classifier $\phi(x,a) = \mathbb{1}\{f(x,a) > t^{1,a}_{(k^{1,a})}\}$. Define*

$$h_{y,a}(\boldsymbol{u},\boldsymbol{v}) = \mathbb{P}\Big(\sum_{i=1}^{S} \pi_i^{y,a} Q\left(u_i, n_i^{y,a}+1-u_i\right) - \sum_{i=1}^{S} \pi_i^{y,1-a} Q\left(v_i, n_i^{y,1-a}+1-v_i\right) \geq \alpha\Big). \tag{6}$$

Then we have:

$$\mathbb{P}(|DEOO(\phi)| > \alpha) \leq h_{1,0}(\boldsymbol{M}^{1,0}, \boldsymbol{m}^{1,1}) + h_{1,1}(\boldsymbol{M}^{1,1}, \boldsymbol{m}^{1,0}), \tag{7}$$

where $\pi_i^{1,a}$ is defined in Proposition 3.2, $\boldsymbol{M}^{1,a} = (M_1^{1,a}, \cdots, M_S^{1,a})$, $\boldsymbol{m}^{1,a} = (m_1^{1,a}, \cdots, m_S^{1,a})$, $M_i^{1,a} = max(\lceil \hat{k}_i^{1,a} + \varepsilon n_i^{1,a} \rceil, n_i^{1,a}+1)$, $m_i^{1,a} = min(\lceil \hat{k}_i^{1,a} - \varepsilon n_i^{1,a} \rceil, 0)$. Especially, $Q(0,\beta) = 0$ and $Q(\alpha,0) = 1$ for $\alpha, \beta \neq 0$.

Proposition 3.4 shows that the fairness guarantee remains valid even when the exact local ranks are replaced by ε-approximate ones. The effect of rank approximation is captured through the inflated upper and lower rank bounds $\boldsymbol{M}^{1,a}$ and $\boldsymbol{m}^{1,a}$, which enlarge the uncertainty region in a conservative manner; as ε decreases, this bound becomes tighter and recovers Proposition 3.2 in the exact-rank case.

In practical distributed settings, calculating the exact local rank in Proposition 3.4 is generally hard due to communication constraints. By adopting approximate ε and related parameters in a distributed quantile algorithm, we strike a balance between accuracy and communication cost, enabling the effective implementation of our algorithm in distributed environments.

In our experiments, we implemented the Q-digest (Shrivastava et al., 2004), a tree-based sketching distributed quantile algorithm commonly used for efficiently approximating quantiles and ranks computation with rigorous theory controlling the error. Due to the inherent characteristics of the Q-digest algorithm, it only yields approximate quantiles and ranks that tend to be greater than their true values. However, considering the adaptability of other distributed quantile algorithms and aiming to reduce the absolute value of ε, we take into account both upward and downward estimation deviations as described in Definition 3.3.

By Proposition 3.4, we construct the candidate set K as

$$K = \{(k^{1,0}, k^{1,1}) | L(\boldsymbol{k}^{1,0}, \boldsymbol{k}^{1,1}) < 1-\beta\}, \tag{8}$$

where $\boldsymbol{k}^{1,a} = (\hat{k}_1^{1,a}, \cdots, \hat{k}_S^{1,a})$ are estimated corresponding "local ranks" of $k^{1,a}$, and $L(\boldsymbol{k}^{1,0}, \boldsymbol{k}^{1,1}) = h_{1,0}(\boldsymbol{M}^{1,0}, \boldsymbol{m}^{1,1}) + h_{1,1}(\boldsymbol{M}^{1,1}, \boldsymbol{m}^{1,0})$ corresponding to the right-hand side of the inequality equation 7.

Algorithm 1 FedFaiREE for EOO

Input: Train dataset $D_i = D_i^{0,0} \cup D_i^{0,1} \cup D_i^{1,0} \cup D_i^{1,1}$; pre-trained classifier ϕ_0 with function f; fairness constraint parameter α ; Confidence level parameter β; Weights of different clients π
Output: classifier $\hat{\phi}(x,a) = \mathbb{1}\{f(x,a) > t^{1,a}_{(k^{1,a})}\}$

1: **Client Side:**
2: **for** i=1,2,..,S **do**
3: Score on train data points in D_i and get $T_i^{y,a} = \{t_{i,1}^{y,a}, t_{i,2}^{y,a}, \cdots, t_{i,n_i^{y,a}}^{y,a}\}$
4: Sort $T_i^{y,a}$ and calculate q-digest of $T_i^{y,a}$ on client i
5: Update digest to server
6: **end for**
7: **Server Side:**
8: Construct K by $K = \{(k^{1,0}, k^{1,1}) | L(\boldsymbol{k}^{1,0}, \boldsymbol{k}^{1,1}) < 1-\beta\}$ {Establishing a set that satisfies fairness constraints and confidence requirements using order statistics. The search for $(k^{1,0}, k^{1,1})$ can be simplified using technique in Appendix C.1.}
9: Select optimal (k_0, k_1) by minimizing equation 9 using estimated values $\hat{p}_a^i$, $\hat{p}_{Y,a}^i$ and $\hat{q}_{Y,a}^i$

3.3 Selection for the optimal threshold

In this subsection, we elaborate on our method for selecting the optimal threshold. For a given pair $(k^{1,0}, k^{1,1})$ from the candidate set, we exploit the properties of order statistics to compute the estimated misclassification error and then select the pair minimizing the estimated error.

To facilitate this, we need to compute the approximate ranks of $t^{1,0}_{(k^{1,0})}$ and $t^{1,1}_{(k^{1,1})}$ in the sorted sets $T_i^{0,0}$ and $T_i^{0,1}$, where $i \in [S]$, respectively. Specifically, we determine $k_i^{0,a}$ such that $t^{0,a}_{i,(k_i^{0,a})} \leq t^{1,a}_{(k^{1,a})} < t^{0,a}_{i,(k_i^{0,a}+1)}$ for $a \in \{0,1\}$. To simplify, in the following sections, we assume the corresponding $\hat{k}_i^{1,a}$ for $i \in [S]$ are ε-approximate ranks and the estimated quantiles presented by distributed quantile algorithm are ε-approximate quantiles. Then, we commence by presenting our observation on the estimation of misclassification error through the following proposition.

Proposition 3.5. *Under Assumption 3.1, the misclassification error can be estimated through* $\hat{\mathbb{P}}\left(\hat{\phi}(x,a) \neq Y\right)$, *which equals to*

$$\sum_{i=1}^{S} \pi_i \Big[\frac{\hat{k}_i^{1,0} + 0.5}{n_i^{1,0}+1} p_0^i p_{Y,0}^i + \frac{\hat{k}_i^{1,1} + 0.5}{n_i^{1,1}+1} p_1^i p_{Y,1}^i + \frac{n_i^{0,0} + 0.5 - \hat{k}_i^{0,0}}{n_i^{0,0}+1} p_0^i q_{Y,0}^i + \frac{n_i^{0,1} + 0.5 - \hat{k}_i^{0,1}}{n_i^{0,1}+1} p_1^i q_{Y,1}^i \Big]. \tag{9}$$

Further, the discrepancy between empirical error and true error is upper bounded:

$$\left| \mathbb{P}\left(\hat{\phi}(x,a) \neq Y\right) - \hat{\mathbb{P}}\left(\hat{\phi}(x,a) \neq Y\right) \right| \leq \theta, \tag{10}$$

where $\theta = \sum_{i=1}^{S} \pi_i [e_i^{0,0} p_0^i q_{Y,0}^i + e_i^{0,1} p_0^i p_{Y,0}^i + e_i^{1,0} p_1^i q_{Y,1}^i + e_i^{1,1} p_1^i p_{Y,1}^i], e_i^{y,a} = \frac{2\lfloor \varepsilon n_i^{y,a} \rfloor + 1}{2(n_i^{y,a}+1)}$.

Proposition 3.5 provides a method for estimating the overall misclassification error using data from the training set with equation 9. However, we may not have exact knowledge of the probabilities p_a^i and $p_{Y,a}^i$. In such cases, we can use the estimated values $\hat{p}_a^i = \frac{n_i^{0,a}+n_i^{1,a}}{n_i^{0,0}+n_i^{0,1}+n_i^{1,0}+n_i^{1,1}}$, $\hat{p}_{Y,a}^i = \frac{n_i^{1,a}}{n_i^{0,a}+n_i^{1,a}}$, $\hat{q}_{Y,a}^i = 1 - \hat{p}_{Y,a}^i$ to calculate the empirical error. We will further present a

theorem to show that we can achieve a desirable accuracy using the estimated values in Section 4.

At the end of this section, we provide a concise summary of our algorithm in Algorithm 1. It is worth noting that while in our experiment, we assume that π_i is proportional to n_i, we may not know the exact values of π_i in real applications. To enhance the robustness of our approach in such real-world scenarios, one can consider introducing a hypothesis space denoted as $H(\pi)$ to model the range of π and incorporate $\max_{\pi \in H(\pi)}$ into Equations equation 8 and equation 9.

3.4 Extension to Equalized Odds

We now extend FedFaiREE from EOO to Equalized Odds (EO), which requires simultaneously controlling both the true positive rate and the false positive rate across groups.

Under EO, fairness constraints involve two types of conditional distributions (for $Y = 1$ and $Y = 0$). Similar to the EOO case, these quantities can be characterized by the relative positions (i.e., ranks or quantiles) of score thresholds within each group. Therefore, EO can be enforced by jointly aligning order statistics for both $Y = 1$ and $Y = 0$ across groups. For notational simplicity, let $n_i^{y,a}$ denote the number of samples in client i with label $Y = y$ and attribute $A = a$. The corresponding rank variables $k^{y,a}$ and $\hat{k}_i^{y,a}$ are defined analogously to the EOO case.

We now present the following extension of Proposition 3.4.

Proposition 3.6. *Under Assumption 3.1, for $a \in \{0,1\}$, consider $k^{1,a} \in \{1,\dots,n^{1,a}\}$, the corresponding $\hat{k}_i^{1,a}$ for $i \in [S]$ which are ε-approximate ranks and the score-based classifier $\phi(x,a) = \mathbb{1}\{f(x,a) > t^{1,a}_{(k^{1,a})}\}$. Define*

$$h_{y,a}(\boldsymbol{u},\boldsymbol{v}) = \mathbb{P}\Big(\sum_{i=1}^{S} \pi_i^{y,a} Q\left(u_i, n_i^{y,a} + 1 - u_i\right) - \sum_{i=1}^{S} \pi_i^{y,1-a} Q\left(v_i, n_i^{y,1-a} + 1 - v_i\right) \geq \alpha\Big).$$

Then we have:

$$\mathbb{P}(|DEO(\phi)| \preceq (\alpha,\alpha)) \geq 1 - h^*_{1,1} - h^*_{1,0} - h^*_{0,1} - h^*_{0,0} \tag{11}$$

where the definitions of $M_i^{y,a}$, $m_i^{y,a}$, $\pi_i^{y,a}$, $Q(A,B)$ are similar to Proposition 3.4, $h^*_{1,1} = h_{y,a}(\boldsymbol{M}^{y,a}, \boldsymbol{m}^{y,a})$.

Compared to the EOO case, the EO constraint introduces two additional terms corresponding to the $Y = 0$ population. As a result, fairness control requires simultaneous alignment of order statistics across both positive and negative classes, leading to a union bound over four terms.

Building upon Proposition 3.6, we construct the candidate set under EO constraints and apply the same rank-selection framework as before.

Algorithm 2 FedFaiREE for EO

Input: Train dataset $D_i = D_i^{0,0} \cup D_i^{0,1} \cup D_i^{1,0} \cup D_i^{1,1}$; pre-trained classifier ϕ_0 with function f; fairness constraint parameter α ; Confidence level parameter β; Weights of different clients π
Output: classifier $\hat{\phi}(x,a) = \mathbb{1}\{f(x,a) > t^{1,a}_{(k^{1,a})}\}$

1: **Client Side:**
2: **for** i=1,2,..,S **do**
3: Score on train data points in D_i and get $T_i^{y,a} = \{t_{i,1}^{y,a}, t_{i,2}^{y,a}, \cdots, t_{i,n_i^{y,a}}^{y,a}\}$
4: Sort $T_i^{y,a}$
5: Calculate q-digest of $T_i^{y,a}$ on client i
6: Update digest to server
7: **end for**
8: **Server Side:**
9: Construct K by $K = \{(k^{1,0}, k^{1,1}) | L(\boldsymbol{k}^{1,0}, \boldsymbol{k}^{1,1}) < 1-\beta\}$, where $L = h^*_{1,1} - h^*_{1,0} - h^*_{0,1} - h^*_{0,0}$ corresponding to the right hand side of equation 11
10: Select optimal (k_0, k_1) by minimizing equation 9 using estimated values $\hat{p}_a^i = \frac{n_i^{0,a}+n_i^{1,a}}{n_i^{0,0}+n_i^{0,1}+n_i^{1,0}+n_i^{1,1}}$ and $\hat{p}^i_{Y,a} = \frac{n_i^{1,a}}{n_i^{0,a}+n_i^{1,a}}$

4 Theoretical Guarantees

In this section, we establish theoretical guarantees for FedFaiREE, showing that it achieves both *finite-sample fairness control* and *near-optimal accuracy*.

At a high level, our results show that: 1). FedFaiREE satisfies the target fairness constraint with high probability in a distribution-free manner; and 2). the resulting classifier achieves a misclassification error close to that of the fair Bayes-optimal classifier, up to controllable approximation and estimation errors.

To derive these results, we first introduce a mild regularity condition on the score function.

Assumption 4.1. The distribution of $f(x,a)$ exhibits the following property. When conducting N independent samplings to form a sample set, let q_0 be the β-quantile of the sample set. There exist function $\delta : \mathbb{N} \to \mathbb{R}$, constant $\gamma > 0$, such that $\lim_{N\to\infty} \delta(N) = 0$ and with a probability of at least $1 - \delta(N)$, for any q considered as an ε-approximate β-quantile of the sample set, it satisfies that , q lies within the $\gamma\varepsilon$-neighborhood of q_0.

Assumption 4.1 ensures that approximate quantiles obtained via distributed rank estimation are stable, in the sense that small rank errors translate into controlled perturbations of the corresponding thresholds. This condition is analogous to a Lipschitz-type regularity of the score distribution near the quantile of interest, and is standard in quantile-based analysis.

In the following theorem, we establish a theoretical basis for the accuracy of FedFaiREE. To facilitate comparison, we introduce the notion of the *fair Bayes-optimal classifier*, defined as the classifier achieving the lowest misclassification error under the fairness constraint. The formal definition is given in Lemma A.2. We denote the standard (unconstrained) Bayes-optimal classifier by

$$\phi^*(x,a) = \mathbb{1}\{f^*(x,a) > 1/2\},$$

where $f^* \in \arg\min_f \mathbb{P}(Y \neq \mathbb{1}\{f(x,a) > 1/2\})$. We now present the theoretical guarantees of Algorithm 1 proposed in the last section.

Theorem 4.2. *Under Assumptions 3.1 and 4.1, let $\alpha' < \alpha$, and let $\hat{\phi}$ be the output of FedFaiREE. Then:*

(1) Fairness guarantee.

$$|DEOO(\hat{\phi})| < \alpha$$

holds with probability $(1-\delta)^N$, where N is the size of the candidate set.

(2) Accuracy guarantee. *Suppose the density of f^* under $A = a, Y = 1$ is continuous. If the input classifier satisfies $|f(x,a) - f^*(x,a)| \leq \epsilon_0$, then for any $\epsilon > 0$ such that $F^*_{(+)}(\epsilon + \gamma\varepsilon) \leq \frac{\alpha-\alpha'}{2} - F^*_{(+)}(2\epsilon_0)$, we have*

$$\mathbb{P}(\hat{\phi}(x,a) \neq Y) - \mathbb{P}(\phi^*_{\alpha'}(x,a) \neq Y) \leq 2F^*_{(+)}(2\epsilon_0) + 2F^*_{(+)}(\epsilon + \gamma\varepsilon) + 8\epsilon^2 + 20\epsilon + 2\theta, \tag{12}$$

with probability at least $1 - 4\sum_{a=0}^{1}\sum_{i=1}^{S} e^{-2n_i^{0,a}\epsilon^2} - \prod_{i=1}^{S}(1 - F^{1,0}_{i(-)}(2\epsilon))^{n_i^{1,0}} - \prod_{i=1}^{S}(1 - F^{1,1}_{i(-)}(2\epsilon))^{n_i^{1,1}} - \delta$, where $\delta = \delta^{1,0}(n^{1,0}) + \delta^{1,1}(n^{1,1})$, θ is defined in Proposition 3.5 and the definition of $F_{(+)}$ and $F_{(-)}$ are shown in Lemma A.4.

The excess risk bound consists of three components: 1). approximation error $F^*_{(+)}(2\epsilon_0)$, which reflects how close the input score function f is to the Bayes-optimal f^*; 2). quantile estimation error $F^*_{(+)}(\epsilon + \gamma\varepsilon)$, which arises from approximate rank estimation and is controlled by ε; and 3). statistical error terms (in ϵ and θ), which vanish as sample sizes increase. Together, these terms show that FedFaiREE achieves near-optimal accuracy with DEOO constraints, provided that the initial score function is sufficiently accurate and the quantile approximation error is controlled, underscoring the effectiveness of our approach in minimizing errors when ensuring fairness in a distribution-free and small-sample manner.

Similarly, we provide theoretical guarantees for DEO fairness.

Theorem 4.3. *Under Assumptions 3.1 and 4.1, given $\alpha' < \alpha$. Suppose $\hat{\phi}$ is the final output of FedFaiREE with target DEO constraint. We then have:*

(1) Fairness guarantee.

$$|DEO(\hat{\phi})| < \alpha$$

holds with probability $(1-\delta)^N$, where N is the size of the candidate set.

(2) Accuracy guarantee. *Suppose the density distribution functions of f^* under $A = a, Y = 1$ are continuous. When the input classifier f satisfies $|f(x,a) - f^*(x,a)| \leq \epsilon_0$, for any $\epsilon > 0$ such that $F^*_{(+)}(\epsilon + \gamma\varepsilon) \leq \frac{\alpha-\alpha'}{2} - F^*_{(+)}(2\epsilon_0)$, we have*

$$\mathbb{P}(\hat{\phi}(x,a) \neq Y) - \mathbb{P}(\phi^*_{\alpha'}(x,a) \neq Y) \leq 2F^*_{(+)}(2\epsilon_0) + 2F^*_{(+)}(\epsilon + \gamma\varepsilon) + 2\theta + O(\epsilon) \tag{13}$$

with probability $1 - 4\sum_{a=0}^{1}\sum_{i=1}^{S} e^{-2n_i^{0,a}\epsilon^2} - \prod_{i=1}^{S}\left(1 - F^{1,0}_{i(-)}(2\epsilon)\right)^{n_i^{1,0}} - \prod_{i=1}^{S}\left(1 - F^{1,1}_{i(-)}(2\epsilon)\right)^{n_i^{1,1}} - \delta$, where the definitions of δ, θ, $F_{(+)}$, $F_{(-)}$ are same as Theorem 4.2.

Compared to the DEOO case, the DEO result involves both positive and negative classes, but the overall structure of the bound remains similar. This shows that our framework extends naturally to stronger fairness notions without degrading statistical guarantees.

5 Numerical Experiments

In this section, we empirically evaluate FedFaiREE on multiple benchmark datasets under both semi-synthetic and real heterogeneous settings, and compare it with existing baselines in terms of the fairness–accuracy trade-off.

5.1 Numerical results for EOO

We evaluate FedFaiREE on two semi-synthetic datasets, Adult (Dua et al., 2017) and Compas (Dieterich et al., 2016), where decentralized data are generated following the pipeline of Ezzeldin et al. (2023), as well as two real-world datasets, ACSIncome (Ding et al., 2021) and CelebA (Liu et al., 2015), which naturally exhibit heterogeneous federated structures.

We apply FaiREE and FedFaiREE as post-processing methods on top of standard federated learning baselines, including FedAvg (McMahan et al., 2017), FedFB (Zeng et al., 2021), and FairFed (Ezzeldin et al., 2023). All models are trained using two-layer neural networks, except for CelebA where we adopt ResNet18. Further implementation details are provided in Appendix C.

5.1.1 Semi-synthetic Data

Adult dataset (Dua et al., 2017), which is employed for the prediction task that determines whether an individual's income exceeds \$50,000, comprises 45,222 samples, featuring various attributes including age, education, and more. Compas dataset (Dieterich et al., 2016), whose task is to predict whether a person will conduct crime in the future, comprises 7214 samples. Gender is chosen as the sensitive feature for both datasets.

Data Processing. To realize the decentralized conditions and account for heterogeneity across clients, we adopt the approach introduced by Ezzeldin et al. (2023). Specifically, we initiated the process by randomly sampling proportions for various sensitive attributes within each client, using the Dirichlet distribution. Subsequently, we partitioned the dataset into client-specific subsets based on these proportions. Within each of these subsets, we performed an 80-20 split, allocating 80% of the data as the local client training set and reserving the remaining 20% for the test set. For the numerical experiments, we repeated this procedure 100 times on both Adult and Compas datasets.

Result and Analysis. Table 1 summarizes the performance on the Adult and Compas datasets. Across all backbone models (FedAvg, FedFB, FairFed), FedFaiREE consistently achieves substantially lower $|DEOO|$ after post-processing, compared to both the original models and post-processing by FaiREE. More importantly, the empirical $|DEOO|_{95}$ aligns closely with the target confidence level $\beta = 0.95$, demonstrating that our high-probability fairness guarantee is effective in practice. In addition, while FaiREE improves fairness in centralized settings, its performance degrades under heterogeneous client distributions. In contrast, FedFaiREE explicitly accounts for client-level variability via local order statistics, leading to stable fairness control across all settings. Further, despite enforcing strict fairness constraints, FedFaiREE maintains accuracy comparable to the baselines. This validates our theoretical result that fairness can be achieved with minimal loss in predictive performance. Overall, these results confirm that FedFaiREE effectively balances fairness and accuracy under finite-sample and heterogeneous

federated settings. More numerical experiments with varying values of α and β, as well as additional ablation studies are deferred to Appendix C.

Table 1: **Results on Adult and Compas dataset.** We conducted 100 experimental repetitions for each model on both datasets and compared the accuracy and fairness indicators of different models. The Method and α columns indicate whether FedFaiREE or FaiREE was used or not and the fairness constraint. Confidence level β is set to be 95% throughout the experiments. $\overline{ACC}$ and $\overline{|DEOO|}$ represent the averages of accuracy and DEOO (defined in equation 2). $|DEOO|_{95}$ represents the 95% quantile of DEOO since we set the confidence level of FedFaiREE to 95% in our experiments.

		Adult				**Compas**			
Model	Method	α	$\overline{ACC}$	$\overline{\lvert DEOO\rvert}$	$\lvert DEOO\rvert_{95}$	α	$\overline{ACC}$	$\overline{\lvert DEOO\rvert}$	$\lvert DEOO\rvert_{95}$
FedAvg	/	/	0.844	0.131	0.178	/	0.662	0.126	0.223
	FaiREE	0.10	0.844	0.071	0.128	0.15	0.659	**0.051**	**0.137**
	FedFaiREE	0.10	0.843	**0.038**	**0.083**	0.15	0.659	**0.051**	**0.137**
FedFB	/	/	0.850	0.057	0.117	/	0.642	0.107	0.174
	FaiREE	0.10	0.850	0.055	0.109	0.15	0.641	**0.062**	**0.125**
	FedFaiREE	0.10	0.850	**0.036**	**0.083**	0.15	0.641	**0.062**	**0.125**
FairFed	/	/	0.842	0.069	0.118	/	0.648	0.097	0.166
	FaiREE	0.10	0.842	0.066	0.112	0.15	0.645	**0.047**	**0.114**
	FedFaiREE	0.10	0.841	**0.037**	**0.081**	0.15	0.645	**0.047**	**0.114**

5.1.2 Real Data Analysis

To validate the effectiveness of FedFaiREE in scenarios with naturally heterogeneous distributions, we further consider the ACSIncome dataset (Ding et al., 2021) and CelebA dataset (Liu et al., 2015). In the ACSIncome dataset, the task is to predict whether an individual's income is above \$50,000, with the sensitive label being Race (white/non-white), and the data partitioned across 50 states. CelebA dataset is a large-scale face attributes dataset comprising more than 200k images of 10k target celebrities, each annotated with 40 attributes. We create clients by grouping every 20 celebrities together, resulting in 508 clients. Following prior work (Park et al., 2022), we use Attractive as target variable and Male as sensitive attribute.

Table 2 reports the results on ACSIncome and CelebA, which exhibit natural heterogeneity. We observe that FedFaiREE significantly reduces $|DEOO|$ across both datasets while maintaining competitive accuracy. Notably, the improvement is particularly pronounced on CelebA, where data heterogeneity is substantial due to the large number of clients. This demonstrates that the rank-based formulation of FedFaiREE is especially effective in realistic federated environments, where distributional differences across clients are inherent and cannot be ignored. Detailed hyperparameter selection and experimental set-up are provided in Sections C.2 and C.3 of the Appendix respectively.

Table 2: **Results on ACSIncome and CelebA datasets.**

Model	FedFaiREE	ACSIncome			CelebA		
		α	$\overline{ACC}$	$\overline{\lvert DEOO\rvert}$	α	$\overline{ACC}$	$\overline{\lvert DEOO\rvert}$
FedAvg	✗	/	0.808	0.126	/	0.709	0.280
	✓	0.10	0.806	**0.041**	0.15	0.684	**0.099**
FairFed	✗	/	0.773	0.092	/	0.721	0.324
	✓	0.10	0.771	**0.044**	0.15	0.697	**0.086**

5.2 Numerical results for EO

We evaluate FedFaiREE under the Equalized Odds (DEO) constraint, as introduced in Section 3.4. Since Equalized Odds requires controlling both the true positive rate and the false positive rate across groups, we report DEOO and DPE, where

$$\text{DPE} = \mathbb{P}_{X|A=1,Y=0}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=0}(\widehat{Y}=1).$$

The results are summarized in Tables 3 and 4. We observe that FedFaiREE consistently improves both DEOO and DPE across all models. Importantly, these fairness gains are achieved while maintaining comparable accuracy, demonstrating that our method effectively handles the stricter DEO constraint without significant loss in predictive performance.

Moreover, the improvements are consistent across different backbone methods (FedAvg, FedFB, and FairFed), indicating that the proposed rank-based framework is robust and broadly applicable in federated settings.

Table 3: **Results of FedFaiREE for DEO on Adult dataset.** We conducted 100 experimental repetitions for each model on both datasets and compared the accuracy and fairness indicators of different models. The "FedFaiREE" and "α" columns indicate whether FedFaiREE was used or not."$\overline{ACC}$", "$\overline{|DEOO|}$" and "$\overline{|DPE|}$" represent the averages of accuracy, DEOO (defined in equation 2) and DPE (defined in equation 35), respectively. "$|DEOO|_{95}$" and "$|DPE|_{95}$" represent the 95% quantile of DEOO and DPE since we set the confidence level of FedFaiREE to 95% in our experiments.

Model	FedFaiREE	Adult					
		α	$\overline{ACC}$	$\overline{\lvert DEOO\rvert}$	$\lvert DEOO\rvert_{95}$	$\overline{\lvert DPE\rvert}$	$\lvert DPE\rvert_{95}$
FedAvg	✗	/	0.844 (0.003)	0.131 (0.030)	0.178	0.088 (0.005)	0.097
	✓	0.10	0.843 (0.003)	**0.037** (0.025)	**0.082**	**0.064** (0.007)	**0.075**
FedFB	✗	/	0.850 (0.003)	0.057 (0.034)	0.117	0.066 (0.007)	0.077
	✓	0.10	0.850 (0.003)	**0.036** (0.025)	**0.083**	**0.061** (0.006)	**0.070**
FairFed	✗	/	0.842 (0.003)	0.069 (0.034)	0.118	0.072 (0.006)	0.083
	✓	0.10	0.841 (0.003)	**0.037** (0.026)	**0.081**	**0.063** (0.006)	**0.071**

Table 4: **Results of FedFaiREE for DEO on Compas dataset.** We conducted 100 experimental repetitions for each model on both datasets and compared the accuracy and fairness indicators of different models. All notations are the same as in Table 3

		Compas					
Model	FedFaiREE	α	$\overline{ACC}$	$\overline{\lvert DEOO\rvert}$	$\lvert DEOO\rvert_{95}$	$\overline{\lvert DPE\rvert}$	$\lvert DPE\rvert_{95}$
FedAvg	✗	/	0.662 (0.011)	0.126 (0.056)	0.223	0.083 (0.032)	0.136
	✓	0.15	0.652 (0.036)	**0.049** (0.045)	**0.137**	**0.028** (0.024)	**0.072**
FedFB	✗	/	0.642 (0.011)	0.107 (0.043)	0.174	0.066 (0.028)	0.112
	✓	0.15	0.642 (0.010)	**0.062** (0.040)	**0.125**	**0.036** (0.024)	**0.081**
FairFed	✗	/	0.648 (0.011)	0.097 (0.047)	0.166	0.087 (0.036)	0.148
	✓	0.15	0.642 (0.029)	**0.047** (0.036)	**0.114**	**0.037** (0.028)	**0.085**

6 Extension

In this section, we extend FedFaiREE to several practically relevant settings, including label shift at test time and multi-group fairness. These extensions demonstrate the flexibility of our rank-based framework and its applicability beyond the basic binary-group setting.

6.1 Label Shift in Test Set

We first consider the label shift setting, where the test distribution differs from the training distribution in class proportions. This scenario is common in real-world deployments (Plassier et al., 2023; Tian et al., 2023), where the marginal distribution of labels may evolve over time while the conditional distribution $P(X, A \mid Y)$ remains stable.

To accommodate this setting, we introduce the counterpart of Assumption 3.1 in the following assumption.

Assumption 6.1. The training data points on client i are i.i.d drawn from the distribution P_i, and we further assume the global distribution P is a mixture of $P_1, \cdots, P_S$ with weight $\{\pi_i\}_{i\in[S]} \in \Delta_S$, while the test data points are sampled from another distribution P_{S+1}, heterogeneity between P and which induced due to label shift, that is, we assume that

$$P^{mix} = \sum_{i=1}^{S} \pi_i P_i = P(X, A|Y) * P^{mix}(Y), \ \left(X^{\text{test}}, Y^{\text{test}}\right) \sim P_{S+1} = P(X, A|Y) * P_{S+1}(Y). \quad (14)$$

This assumption captures the classical label-shift setting, in which only the class proportions change between training and testing, while the conditional distribution of covariates and sensitive attributes given the label remains stable. We adapt FedFaiREE to this setting by modifying the empirical objective used for rank selection. Specifically, we replace equation 9 with a reweighted error estimator that accounts for the discrepancy between training and test label distributions:

$$\begin{aligned}\hat{\mathbb{P}}\left(\hat{\phi}(x,a) \neq Y\right) = \sum_{i=1}^{S} \pi_i [&\frac{\hat{k}_i^{1,0} + 0.5}{n_i^{1,0} + 1} p_0^i p_{Y,0}^i w^{1,0} + \frac{\hat{k}_i^{1,1} + 0.5}{n_i^{1,1} + 1} p_1^i p_{Y,1}^i w^{1,1} \\ &+ \frac{n_i^{0,0} + 0.5 - \hat{k}_i^{0,0}}{n_i^{0,0} + 1} p_0^i q_{Y,0}^i w^{0,0} + \frac{n_i^{0,1} + 0.5 - \hat{k}_i^{0,1}}{n_i^{0,1} + 1} p_1^i q_{Y,1}^i w^{0,1}],\end{aligned} \quad (15)$$

Algorithm 3 FedFaiREE for label shift case

Input: Train dataset $D_i = D_i^{0,0} \cup D_i^{0,1} \cup D_i^{1,0} \cup D_i^{1,1}$; pre-trained classifier ϕ_0 with function f; fainess constraint parameter α ; Confidence level parameter β; Weights of different clients π
Output: classifier $\hat{\phi}(x,a) = \mathbb{1}\{f(x,a) > t^{1,a}_{(k^{1,a})}\}$

Client Side:
for i=1,2,..,S **do**
  Score on train data points in D_i and get $T_i^{y,a} = \{t_{i,1}^{y,a}, t_{i,2}^{y,a}, \cdots, t_{i,n_i^{y,a}}^{y,a}\}$
  Sort $T_i^{y,a}$
  Calculate q-digest of $T_i^{y,a}$ on client i
  Update digest to server
end for
Server Side:
Construct K by $K = \{(k^{1,0}, k^{1,1}) | L(\boldsymbol{k}^{1,0}, \boldsymbol{k}^{1,1}) < 1-\beta\}$
Select optimal (k_0, k_1) by minimizing equation 15 using estimated values $\hat{p}_a^i = \frac{n_i^{0,a}+n_i^{1,a}}{n_i^{0,0}+n_i^{0,1}+n_i^{1,0}+n_i^{1,1}}$ and $\hat{p}_{Y,a}^i = \frac{n_i^{1,a}}{n_i^{0,a}+n_i^{1,a}}$

where $w^{y,a} = \frac{p_a^{S+1} p_{Y,a}^{S+1}}{p_a p_{Y,a}}$. The weights $w^{y,a}$ reweight each subgroup to reflect its prevalence under the test distribution, effectively correcting the bias introduced by label shift. As a result, similar to Proposition 3.2, we have the following proposition.

Proposition 6.2. *Under Assumption 6.1, the misclassification error can be estimated by equation 15. Further, discrepancy between empirical error and true error is limited by following inequality:*

$$\left| \mathbb{P}\left(\hat{\phi}(x,a) \neq Y\right) - \hat{\mathbb{P}}\left(\hat{\phi}(x,a) \neq Y\right) \right| \leq \theta' \tag{16}$$

where $e_i^{y,a} = \frac{2\lfloor \varepsilon n_i^{y,a} \rfloor + 1}{2(n_i^{y,a}+1)}$ *and*

$$\theta' = \sum_{i=1}^{S} \pi_i [e_i^{0,0} p_0^i q_{Y,0}^i w^{0,0} + e_i^{0,1} w^{0,1} p_0^i p_{Y,0}^i + e_i^{1,0} w^{1,0} p_1^i q_{Y,1}^i + e_i^{1,1} w^{1,1} p_1^i p_{Y,1}^i].$$

Proposition 6.2 shows that the reweighted estimator consistently approximates the test-time misclassification error, despite the distribution shift. This result enables the server to select candidate classifiers based on an empirical objective that accurately tracks test-time performance. The overall procedure is summarized in Algorithm 3. We further establish fairness and accuracy guarantees analogous to Theorem 4.2.

Theorem 6.3. *Under Assumptions 4.1 and 6.1, given $\alpha' < \alpha$. Suppose $\hat{\phi}$ is the final output of FedFaiREE. We then have:*

(1) Fairness guarantee.

$$|DEOO(\hat{\phi})| < \alpha$$

holds with probability $(1-\delta)^N$, where N is the size of the candidate set.

(2) Accuracy guarantee. *Under the assumptions specified in Appendix A.5, with high probability we have*

$$\mathbb{P}(\hat{\phi}(x,a) \neq Y) - \mathbb{P}\left(\phi^*_{\alpha'}(x,a) \neq Y\right) \leq 2F^*_{(+)}(2\epsilon_0) + 2F^*_{(+)}(\epsilon + \gamma\varepsilon) + 2\theta' + O(\epsilon), \tag{17}$$

Theorem 6.3 shows that FedFaiREE retains both fairness control and near-optimal accuracy under label shift. This is particularly important in deployment scenarios where label distributions may change over time.

6.2 Extension to the Multi-Group Setting

We now extend FedFaiREE to the multi-group setting, where the sensitive attribute takes values in $\mathcal{A} = \{0, 1, \ldots, A_0\}$ with $A_0 > 1$. This setting is common in practice, where fairness must be enforced across multiple demographic groups. In such settings, the fairness constraint must ensure that the classifier behaves similarly across all protected groups, rather than only between two groups.

We begin by introducing the notion of Equality of Opportunity in the multi-group setting.

Definition 6.4. (Equality of Opportunity; Multiple Groups) A classifier satisfies Equality of Opportunity if the true positive rate is equal across all protected groups:

$$\mathbb{P}_{X|A=0,Y=1}(\widehat{Y} = 1) = \mathbb{P}_{X|A=a,Y=1}(\widehat{Y} = 1), \quad \forall a \in \mathcal{A}.$$

This definition generalizes the binary-group notion of Equality of Opportunity by requiring parity of true positive rates across all groups.

To quantify deviations from this condition, we define

$$DEOOM = \max_a \left| \mathbb{P}_{X|A=a,Y=1}(\widehat{Y} = 1) - \mathbb{P}_{X|A=0,Y=1}(\widehat{Y} = 1) \right|.$$

Thus, controlling $DEOOM$ amounts to ensuring that no protected group differs too much from the reference group in terms of true positive rate. We now extend the fairness control result underlying FedFaiREE to this multi-group setting.

Proposition 6.5. *Under Assumption 3.1, for $a \in \{0, 1, \cdots, A_0\}$, consider $k^{1,a} \in \{1, \ldots, n^{1,a}\}$, the corresponding $\hat{k}_i^{1,a}$ for $i \in [S]$ which are ε-approximate ranks and the score-based classifier $\phi(x, a) = \mathbb{1}\{f(x, a) > t^{1,a}_{(k^{1,a})}\}$. Define*

$$\begin{aligned} h^*_{y,a} =& \mathbb{P}\left(\sum_{i=1}^{S} \pi_i^{y,a} Q\left(M_i^{1,a}, n_i^{y,a} + 1 - M_i^{1,a} \right) - \sum_{i=1}^{S} \pi_i^{y,0} Q\left(m_i^{1,0}, n_i^{y,0} + 1 - m_i^{1,0} \right) \geq \alpha \right) \\ &+ \mathbb{P}\left(\sum_{i=1}^{S} \pi_i^{y,0} Q\left(M_i^{1,0}, n_i^{y,0} + 1 - M_i^{1,0} \right) - \sum_{i=1}^{S} \pi_i^{y,a} Q\left(m_i^{1,a}, n_i^{y,a} + 1 - m_i^{1,a} \right) \geq \alpha \right). \end{aligned}$$

Then we have:

$$\mathbb{P}(|DEOOM(\phi)| > \alpha) \leq \sum_{a=1}^{A_0} h^*_{1,a} \tag{18}$$

where $\pi_i^{1,a}$, $\pi_i^{1,0}$ are similarly defined as in Proposition 3.4. $M_i^{1,a} = max(\lceil \hat{k}_i^{1,a} + \varepsilon n_i^{1,a} \rceil, n_i^{1,a} + 1)$, $m_i^{1,a} = min(\lceil \hat{k}_i^{1,a} - \varepsilon n_i^{1,a} \rceil, 0)$, $M_i^{1,0}$ and $m_i^{1,0}$ are similarly defined. $Q(\alpha, \beta)$ are independent random variables and $Q(\alpha, \beta) \sim Beta(\alpha, \beta)$. Especially, we define $Q(0, \beta) = 0$ and $Q(\alpha, 0) = 1$ for $\alpha, \beta \neq 0$.

Algorithm 4 FedFaiREE for Multi-Groups

Input: Train dataset $D_i = D_i^{0,0} \cup D_i^{0,1} \cup D_i^{1,0} \cup D_i^{1,1}$; pre-trained classifier ϕ_0 with function f; fairness constraint parameter α ; Confidence level parameter β; Weights of different clients π
Output: classifier $\hat{\phi}(x,a) = \mathbb{1}\{f(x,a) > t^{1,a}_{(k^{1,a})}\}$

Client Side:
for i=1,2,..,S **do**
  Score on train data points in D_i and get $T_i^{y,a} = \{t_{i,1}^{y,a}, t_{i,2}^{y,a}, \cdots, t_{i,n_i^{y,a}}^{y,a}\}$
  Sort $T_i^{y,a}$
  Calculate q-digest of $T_i^{y,a}$ on client i
  Update digest to server
end for
Server Side:
Construct K by $K = \{(k^{1,0}, k^{1,1}, \cdots, k^{1,A_0}) | L < 1-\beta\}$, where L is defined by the right-hand side of Inequality 18
Select optimal $(k^{1,0}, k^{1,1}, \cdots, k^{1,A_0})$ by minimizing equation 19 using estimated values $\hat{p}_a^i$ and $\hat{p}_{Y,a}^i$

Proposition 6.5 is the multi-group analogue of Proposition 3.4. It shows that the probability of violating the fairness constraint can still be controlled by aggregating client-level uncertainty through Beta-distributed order-statistics bounds. Compared to the binary-group case, fairness control now requires bounding deviations for each group relative to the reference group. This leads to a summation over all groups in equation 18, reflecting a union bound over group-wise fairness violations. Next, we extend the error estimation result to the multi-group setting.

Proposition 6.6. *Under Assumption 3.1, the misclassification error can be estimated by*

$$\hat{\mathbb{P}}\left(\hat{\phi}(x,a) \neq Y\right) = \sum_{i=1}^{S}\left[\pi_i \sum_{a=0}^{A_0}\left(\frac{\hat{k}_i^{1,a}+0.5}{n_i^{1,a}+1} p_a^i p_{Y,a}^i + \frac{n_i^{0,a}+0.5-\hat{k}_i^{0,a}}{n_i^{0,a}+1} p_a^i q_{Y,a}^i\right)\right] \tag{19}$$

Further, the discrepancy between empirical error and true error is upper bounded by the following:

$$\left|\mathbb{P}\left(\hat{\phi}(x,a) \neq Y\right) - \hat{\mathbb{P}}\left(\hat{\phi}(x,a) \neq Y\right)\right| \leq \theta, \tag{20}$$

where $\theta = \sum_{i=1}^{S}\left[\pi_i \sum_{a=0}^{A_0}\left(e_i^{0,a} p_a^i q_{Y,a}^i + e_i^{1,a} p_1^i q_{Y,a}^i\right)\right]$, $e_i^{y,a} = \frac{2\lfloor \varepsilon n_i^{y,a} \rfloor + 1}{2(n_i^{y,a}+1)}$.

Proposition 6.6 shows that, the error estimator retains the same structure as in the binary case, with an additional summation over groups. This shows that the rank-based formulation scales naturally with the number of groups.

Based on Propositions 6.5 and 6.6, we obtain the following extension of FedFaiREE, summarized in Algorithm 4. We next state the corresponding theoretical guarantee.

Theorem 6.7. *Under Assumption 3.1 and 4.1, given $\alpha' < \alpha$. Suppose $\hat{\phi}$ is the final output of FedFaiREE, we have:*

(1) Fairness guarantee.

$$|DEOOM(\hat{\phi})| < \alpha$$

holds with probability $(1-\delta)^N$, where N is the size of the candidate set.

(2) Accuracy guarantee. *Suppose the density distribution functions of f^* under $A = a, Y = 1$ are continuous. When the input classifier f satisfies $|f(x,a) - f^*(x,a)| \leq \epsilon_0$, for any $\epsilon > 0$ such that $F^*_{(+)}(\epsilon + \gamma\varepsilon) \leq \frac{\alpha - \alpha'}{2} - F^*_{(+)}(2\epsilon_0)$, we have*

$$\mathbb{P}(\hat{\phi}(x,a) \neq Y) - \mathbb{P}\left(\phi^*_{\alpha'}(x,a) \neq Y\right) \leq 2F^*_{(+)}(2\epsilon_0) + 2F^*_{(+)}(\epsilon + \gamma\varepsilon) + 2\theta + O(\epsilon) \tag{21}$$

with probability at least $1 - 4\sum_{a=0}^{A_0}\sum_{i=1}^{S} e^{-2n_i^{0,a}\epsilon^2} - \sum_{a=0}^{A_0}\prod_{i=1}^{S}\left(1 - F^{1,a}_{i(-)}(2\epsilon)\right)^{n_i^{1,a}} - \delta$. Here $\delta = \sum_{a=0}^{A_0}\delta^{1,a}(n^{1,a})$, θ is defined in Proposition 6.6 and the definition of $F_{(+)}$ and $F_{(-)}$ are shown in Lemma A.4

The theorem shows that the main guarantees of FedFaiREE extend seamlessly to the multi-group setting. In particular, the algorithm continues to enforce fairness with high probability, while achieving near-optimal accuracy with explicit finite-sample bounds.

Overall, these extensions highlight the generality of the proposed framework. FedFaiREE can accommodate distribution shift and complex fairness notions without altering its core rank-based structure. We defer the extension to multi-label settings to Appendix A.6.

7 Conclusion

In this paper, we introduce FedFaiREE, a small-sample and distribution-free approach to guarantee fairness constraints under the federated learning setting. FedFaiREE addresses concerns that commonly exist in federated learning, such as client heterogeneity, small samples, and limited communication costs. The FedFaiREE framework can be applied to a wide range of group fairness notions and various scenarios, including label shifts and multi-group settings.

For future work, an exploration of more efficient distributed quantile algorithms for rank and quantile calculations within the FedFaiREE framework could significantly enhance its scalability and performance. Moreover, exploring a broader range of application scenarios and assessing its performance in conjunction with in-processing fair federated learning frameworks could yield valuable insights.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–23, 2022.

Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*, 2021.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. *Advances in neural information processing systems*, 33:15088–15099, 2020.

Lingyang Chu, Lanjun Wang, Yanjie Dong, Jian Pei, Zirui Zhou, and Yong Zhang. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. Addressing algorithmic disparity and performance inconsistency in federated learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26091–26102. Curran Associates, Inc., 2021.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4):1–36, 2016.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. *Fairness-aware Agnostic Federated Learning*, pages 181–189. 2021.

Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017.

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.

Yahya Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A. Avestimehr. Fairfed: Enabling group fairness in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:7494–7502, 06 2023.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 144–152. SIAM, 2016.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. *Advances in neural information processing systems*, 29, 2016.

Faisal Hamman and Sanghamitra Dutta. Demystifying local and global fairness trade-offs in federated learning using information theory. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

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

Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Fair federated learning via bounded group loss. *arXiv preprint arXiv:2203.10190*, 2022.

Tiansheng Huang, Weiwei Lin, Wentai Wu, Ligang He, Keqin Li, and Albert Y Zomaya. An efficiency-boosting client selection scheme for federated learning with fairness guarantee. *IEEE Transactions on Parallel and Distributed Systems*, 32(7):1552–1564, 2020.

James E Johndrow and Kristian Lum. An algorithm for removing sensitive information. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

Madhura Joshi, Ankit Pal, and Malaikannan Sankarasubbu. Federated learning for healthcare domain-pipeline, applications and challenges. *ACM Transactions on Computing for Healthcare*, 3(4):1–36, 2022.

Puheng Li, James Zou, and Linjun Zhang. Fairee: fair classification with finite-sample and distribution-free guarantee. In *The Eleventh International Conference on Learning Representations*, 2022.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Nicholas B Allen, Randy P Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Tianci Liu, Haoyu Wang, Yaqing Wang, Xiaoqian Wang, Lu Su, and Jing Gao. Simfair: A unified framework for fairness-aware multi-label classification. *arXiv preprint arXiv:2302.09683*, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, pages 22942–22964. PMLR, 2023.

Ge Luo, Lu Wang, Ke Yi, and Graham Cormode. Quantiles over data streams: experimental comparisons, new analyses, and further improvements. *The VLDB Journal*, 25:449–472, 2016.

Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. *Federated Learning: Privacy and Incentive*, pages 189–204, 2020.

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.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.

Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax demographic group fairness in federated learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 142–159, 2022.

Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398, 2022.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In *Proceedings of the 40th International Conference on Machine Learning*, pages 27907–27947. PMLR, 2023.

Borja Rodríguez-Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. *arXiv preprint arXiv:2109.08604*, 2021.

Nisheeth Shrivastava, Chiranjeeb Buragohain, Divyakant Agrawal, and Subhash Suri. Medians and beyond: new aggregation techniques for sensor networks. In *Proceedings of the 2nd international conference on Embedded networked sensor systems*, pages 239–249, 2004.

Qinglong Tian, Xin Zhang, and Jiwei Zhao. ELSA: Efficient label shift adaptation through the lens of semiparametric models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 34120–34142. PMLR, 2023.

Chengxu Yang, Qipeng Wang, Mengwei Xu, Zhenpeng Chen, Kaigui Bian, Yunxin Liu, and Xuanzhe Liu. Characterizing impacts of heterogeneity in federated learning upon large-scale smartphone data. In *Proceedings of the Web Conference 2021*, pages 935–946, 2021.

Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. A fairness-aware incentive scheme for federated learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 393–399, 2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Xianli Zeng, Edgar Dobriban, and Guang Cheng. Bayes-optimal classifiers under group fairness. *arXiv preprint arXiv:2202.09724*, 2022.

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