OursFed: Provable Group Fairness-Aware Federated Learning Against Distrust and Fragility

Yun Xin[1], Jianfeng Lu[2,3*], Gang Li[4*], Shuqing Cao[2], Guanghui Wen[5], Kehao Wang[6]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology, China
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, China
[3]Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education, China
[4]College of Computer Science, Inner Mongolia University, China
[5]School of Automation, Southeast University, China
[6]School of Information Engineering, Wuhan University of Technology, China
{yunxin, lujianfeng}@wust.edu.cn, gli@imu.edu.cn, shuqincao@wust.edu.cn, ghwen@seu.edu.cn, kehao.wang@whut.edu.cn

Abstract

With the increasing application of high-stakes decision-making application in Federated Learning (FL), ensuring fairness across different populations to prevent biases against certain groups has become crucial. However, achieving group fairness (GF) in FL presents a formidable challenge due to its decentralization, which complicates the global GF estimation by the server. Moreover, distrust and fragility hinder the server from gathering GF values from unreliable clients. This challenge motivates our proposal of OursFed, a provable GF-aware FL framework that integrates a privacy pair-based contract and robust GF estimation method to address issues of distrust and fragility. Methodologically, we categorize client unreliability into two categories: active unreliability stemming from distrust and passive unreliability arising from fragility. To mitigate active unreliability, we design a privacy pair-based contract to guarantee truthful GF reporting, and enhance multivariate analysis by identifying relationships among multiple private data. To counteract passive unreliability, we develop a robust GF estimation using non-parametric techniques to smooth data and estimate probability densities and regression functions, improving per-client GF accuracy under multi-dimensional data perturbation. Theoretically, we demonstrate the efficacy of OursFed by analyzing its convergence, GF stability, and accuracy deviation. Experimentally, evaluations on two real datasets show that OursFed improves GF by 28.61% with at most 2.7% trade-off versus state-of-the-art baselines, and synthetic experiments further confirm its effectiveness in handling fragility and distrust.

Introduction

Federated learning (FL) has emerged as a novel distributed paradigm that enables effective collaborative learning while protecting remote clients' data privacy (McMahan et al. 2017). With the widespread application of FL in high-stakes critical decision-making, fairness concerns arising from the low priority given to equitable treatment of clients have received significant attention (Mehrabi et al. 2021). Fairness challenges arise throughout the FL process, including client selection (Zhang et al. 2023), model optimization (Lu et al. 2023), and reward allocation (Murhekar et al. 2024). Due to heterogeneity of data distribution, quantity, and quality, FL models may exhibit biases towards certain groups, embedding and even amplifying societal biases among different demographics, such as job application screening (Raghavan et al. 2020), and criminal sentencing recommendations (Kleinberg et al. 2018). Thus, among various fairness considerations, ensuring fair decision-making across diverse populations remains as a critical research priority in FL.

To ensure consistently performance among different demographics, Group Fairness (GF) has been proposed to quantify model non-discrimination (Dwork et al. 2012). Prominent GF metrics, such as demographic parity (DP) (Feldman et al. 2015), equal opportunity (Hardt et al. 2016), and predictive parity (Chouldechova 2017), have been applied in machine learning to increase decision-making fairness, enhance social trust, and reduce bias across populations (Su et al. 2024; Cao et al. 2025). However, these metrics may not be suitable in fragile scenarios due to data perturbation, which can lead to inaccurate estimates of GF by clients. Although the extended GF metric (Jiang et al. 2022) addresses the aforementioned issue, it is inappropriate for decentralized settings where selfish clients may benefit from actively misreporting GF, leading to unfair reward allocation and poor model performance. Therefore, the development of a GF-aware FL framework that bolsters GF in fragile and distrust scenarios emerges as a pressing concern.

To bridge this gap, the inherent unreliability of clients poses significant challenges for GF reporting and estimation. *(i) Distrust in GF reporting.* The server is unable to access GF from a comprehensive global viewpoint. Instead, it relies on each client to self-assess and report its GF in the distributed FL setting. However, rational and selfish clients may strategically misreport their fairness values to gain higher returns (Deng et al. 2022; Ding, Gao, and Huang 2023; Huang et al. 2022). *(ii) Fragility of GF estimation.* Data variability caused by dynamic changes among clients may lead to turbulent GF estimations, and clients may passively report GF

*Corresponding Authors are Jianfeng Lu and Gang Li.

inaccurately (Jiang et al. 2022; Li et al. 2021). As a result, distrust and fragility make it challenging for FL to execute fairness aggregation and make proper reward assignments.

To overcome the abovementioned challenges, we propose a provable GF-aware framework named OursFed, i.e., a rObust and truthful group fairness-aware Federated learning, which integrates a privacy pair-based contract and robust GF estimation method to address issue of distrust and fragility. To address active unreliability, we design a privacy pair-based contract that represents clients' privacy attributes (e.g., GF, cost) as information pairs instead of linear combinations. This contract mitigates distrust by discouraging clients from misreporting private information and enables multivariate analysis without the information loss or parameter tuning issues of linear combination methods (Wang et al. 2023; Hu et al. 2022). To address passive unreliability, we extend DP (Feldman et al. 2015) using non-parametric methods, particularly kernel density estimation method (Tang and He 2015) and Nadaraya-Watson Kernel Estimator (Bierens 1994), to handle multi-dimensional data perturbation in fragility scenarios. Smooth density and regression curves improve robustness and accuracy in GF optimization. We summarize our main contributions as follows:

- Methodologically, we propose OursFed, the first provable GF-aware FL framework that jointly mitigates distrust and fragility via game-theoretic and statistical foundations. Our Stackelberg equilibrium analysis reveals GF misreporting as a dominant strategy, motivating a privacy pair-based contract for truthful reporting. Our robust estimator addresses multi-dimensional perturbations by bounding fragility within a specified error.
- Theoretically, we prove that OursFed converges to an approximate optimal solution under practical FL constraints, exhibiting remarkable stability in preserving both fairness and accuracy even when GF deteriorates. Particularly, we derive the lower bound for the probability of GF deteriorates, and the upper bound for the probability of accuracy deviation from a standard value.
- Experimentally, rich results show that OursFed achieves a 28.61% improvement in GF with at most 2.7% accuracy loss compared to state-of-the-art baselines on real-world datasets. Synthetic experiments further verify its effectiveness in ensuring truthful reporting under distrust and alleviating GF estimation bias in fragile scenarios.

Related Work

Group Fairness in FL. Fairness issues pervade all stages of FL, including client selection (Zhang et al. 2023), model optimization (Guo et al. 2023), and reward allocation (Xiao et al. 2023). Among them, GF plays a key role in mitigating decision bias across demographics (Shi, Yu, and Leung 2023). Recent works have explored GF from different perspectives. For instance, Cao et al. [2025] and Long et al. [2024] adopted standard fairness metrics, which fail under data fragility. To address instability, Jiang et al. [2022] and Cho et al. [2020] extended these metrics, yet their methods remain ineffective under multi-dimensional perturbations. Considering the decentralized nature of FL, Ezzeldin et al. [2023] proposed GF-aware aggregation, but it lacks robustness in distrustful settings and a theoretical analysis of the GF–accuracy trade-off. Overall, existing approaches handle either fragility or fairness, often assuming single-dimensional perturbations or client trustworthiness. In contrast, OursFed simultaneously addresses both issues—enforcing truthful GF reporting for actively unreliable clients via a privacy pair–based contract and enabling robust GF estimation for passively unreliable clients through an enhanced demographic parity method.

Incentive Mechanisms for FL. In FL with incomplete information, incentive mechanisms encourage selfish clients to truthfully report private information and contribute data cooperatively to enhance model quality. Prior works, such as Zhao et al. [2023] and Baudry et al. [2024], employed auction-based schemes, while Wang et al. [2022] introduced multi-dimensional contracts to foster cooperation. However, contracts based on linear combinations of private information fail to capture nonlinear feature dependencies. In contrast, our proposed contract represents multi-dimensional private information via independent feature pairs, offering two advantages: (i) intuitive multi-dimensional visualization and (ii) complete preservation of feature integrity for rigorous multivariate analysis.

System Model and Problem Formulation

In this section, we introduce the system setting of a typical FL, provide a basic estimation method for GF, and propose the OursFed framework.

System Setting

A typical FL involves a server and a set of clients $\mathcal{N} = \{1, \cdots, N\}$ collaborating to train a global model $\boldsymbol{\omega}$. If client $n \in \mathcal{N}$ selects a subset $\mathcal{Z}_n$ with Z_n samples from the local data $\mathcal{D}_n$ for local training, i.e. $\mathcal{Z}_n \subseteq \mathcal{D}_n$, the average loss on client n is

$$l_n(\boldsymbol{\omega}, \mathcal{Z}_n) = \frac{1}{Z_n} \sum_{m \in \mathcal{Z}_n} l_n^m(\boldsymbol{\omega}), \tag{1}$$

where $l_n^m(\boldsymbol{\omega})$ as the prediction loss function for the m-th data sample of client n with parameters $\boldsymbol{\omega}$ (Fan et al. 2022). FL aims to minimize the global loss function on the entire dataset (Saha et al. 2022), i.e.,

$$\min_{\omega} l(\boldsymbol{\omega}) = \min_{\omega} \sum_{n \in \mathcal{N}} \theta_n l_n(\boldsymbol{\omega}, \mathcal{Z}_n), \tag{2}$$

where $\theta_n = \frac{Z_n}{Z}$, and $Z = \sum_{n \in \mathcal{N}} Z_n$.

A single FL round that updates the global model from ω^t to ω^{t+1} as follows. Firstly, the server broadcasts the latest model ω^t to all clients. Secondly, each client n performs local learning on $\mathcal{Z}_n$ using ω^t to obtain ω_n^{t+1} via $\boldsymbol{\omega}_n^{t+1} = \boldsymbol{\omega}^t - \eta \nabla l_n(\boldsymbol{\omega}^t, \mathcal{Z}_n)$, where η is the learning rate and $\nabla l_n(\boldsymbol{\omega}^t, \mathcal{Z}_n)$ denotes the gradient of local average loss. Thirdly, the server aggregates local models using the standard federated averaging method: $\boldsymbol{\omega}^{t+1} = \sum_{n \in \mathcal{N}} \theta_n \boldsymbol{\omega}_n^t$.

Group Fairness

We consider a classifier for target value prediction in FL, which can seamlessly extend to multi-class prediction tasks (Cho, et al. 2020). Each sample often contains sensitive demographic information, which can lead to discrimination, manifested as unequal predictions between protected groups (Ezzeldin et al. 2023; Blandin et al. 2023). To investigate these issue, we incorporate sensitive attributes into each sample and define client n's dataset as $\mathcal{D}_n = \{(\mathbf{x}_n^m, y_n^m, s_n^m)\}_{m=1}^{D_n}$, where D_n is the number of samples. For each sample, m is sample index, $\mathbf{x}_n^m \subseteq \mathbb{R}$ represents the input features, $y_n^m \in \{0,1\}$ is the ground truth label, and $s_n^m \in [0,1]$ is the continuous sensitive value. The predicted label $\hat{y}_n^m$ is derived by the output scores $\tilde{y}_n^m = h(\mathbf{x}_n^m)$ at a threshold: $\hat{y}_n^m = \mathbb{I}\{\tilde{y}_n^m \geq \tau | \tau \in [0,1]\}$, where $h : \mathbf{x} \to \tilde{y}$ is a hypothesis function, and $\mathbb{I}\{\cdot\}$ is a indicator function.

Among various GF metrics (Hardt et al. 2016; Chouldechova 2017), we focus on DP (Feldman et al. 2015), one of the most important metrics for quantifying the model's decision bias on the rate of positive outcomes across groups. The predictor satisfies DP if the predicted label $\hat{y}_n$ is independent of the discrete sensitive attribute $s_n \in \{0,1\}$, i.e., $P(\hat{y}_n|s_n) = P(\hat{y}_n)$. To emphasize the statistical properties of the overall data distribution, we omit the superscript m, focusing on the model's expected behavior across all samples from a client rather than any specific one. We define the DP metric $b_n \in [0,1]$ to quantify client n's group bias, where lower values indicate better fairness.

Definition 1 *(DP) The DP metric b_n is defined as:*

$$b_n = \sum_{s \in s_n} [P(\tilde{y}_n \geq \tau | s) - P(\tilde{y}_n \geq \tau)], n \in \mathcal{N}. \quad (3)$$

Given the heterogeneity in data distribution, quantity, and quality across demographics in FL, ensuring GF in the global model is essential. Conventional aggregation, which prioritizes clients with larger datasets, overlooks GF and leads to unfair outcomes across sensitive groups. Local group bias in certain clients may propagate and amplify globally. To alleviate it, we aggregate models using the GF metrics $f_n = 1 - b_n$ and data contribution z_n as follows:

$$\omega^t = \omega^{t-1} + \sum_{n \in \mathcal{N}} \frac{z_n f_n}{\sum_{m \in \mathcal{N}} z_m f_m} \omega_n^t. \quad (4)$$

Group Fairness-Aware Federated Learning

To address group bias in FL, we propose OursFed, illustrated in Figure 1. Given the decentralized nature of FL, the server collect GF metrics from clients. Each round proceeds as follows: the server broadcasts global parameters and reward rules (step ①); clients estimate GF locally (step ②), perform local updates (step ③), and submit their models and private information pairs (step ④); finally, the server rewards clients based on model quality and aggregates the models (step ⑤).

Although this GF-aware workflow is easy to conduct, it suffers from serious performance degradation when the clients are unreliable. More specifically, *unreliable clients may either actively misreport GF to achieve a more favorable outcome at step ④, or passively distort them because of data perturbation at step ②*. To address issues of distrust and fragility, OursFed is designed to maximize the server's utility $u_s = w_s - r_s$. This framework not only considers the welfare w_s derived from the model's quality, but also uses reward r_s to incentivize clients to participant in FL and provide high-quality data. We can formulate the GF-aware FL against distrust and fragility problem as follows:

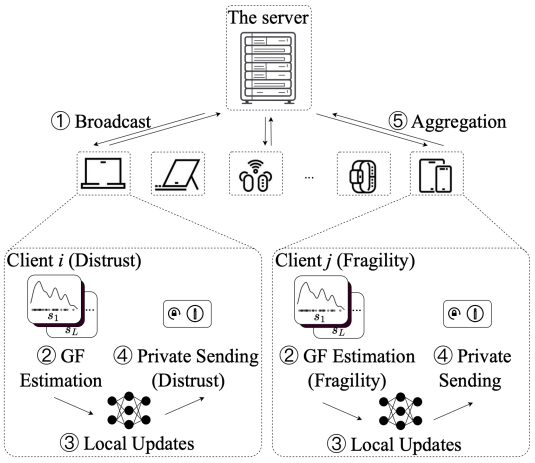

Figure 1: An overview of OursFed.

$$\mathbb{P}_1 : \max u_s, \quad \text{s.t.} \begin{cases} u_n(f_n) \geq u_n(f_n'), \\ P(|f_n - \hat{f}_n| \geq \epsilon) \leq \delta. \end{cases} \quad (5)$$

To overcome distrust issues in FL, we need to ensure that clients can obtain higher utility u_n when reporting truthful GF f_n rather than f_n', i.e., $u_n(f_n) \geq u_n(f_n')$. To resolve fragility issue, we need to ensure that the probability of the difference between the evaluated GF $\hat{f}_n$ and real GF f_n exceeding ϵ is at most δ, i.e., $P(|\hat{f}_n - f_n| \geq \epsilon) \leq \delta$.

Design of OursFed

This section presents a privacy pair-based contract for truthful GF reporting in distrustful settings and a robust DP method for accurate GF estimation under data fragility.

Privacy Pair-based Contract Design

We formulate a Stackelberg game with a unique equilibrium where clients misreport GF, and introduce a privacy pair-based contract that enforces truthful reporting and optimizes server utility under distrust.

False Report on Group Fairnes Clients are characterized by two-dimensional private information (f_n, c_n), corresponding to GF degree and marginal data-usage cost. Without loss of generality, we sort f_n in descending order and classify N clients into I types, denoted by $f_1 \geq \cdots \geq f_I$; similarly, we sort c_n in ascending order and group clients into J types, denoted by $c_1 \leq \cdots \leq c_J$. The information pair (f_i, c_j) induces an $I \times J$ type space over the client set $\mathcal{N}$. Let $\mathcal{U}_{ij} = \{1, \cdots, U_{ij}\}$ denote the set of clients

with GF f_i and cost c_j, where $\sum_{i=1}^{I}\sum_{j=1}^{J} U_{ij} = N$. Consider a simple scenario where the server offers total reward R_{ij} to clients in $\mathcal{U}_{ij}$, which is proportionally distributed among them based on their data contributions. Each client can strategically report type (i, j), corresponding to private information (f_i, c_j) , to enter $\mathcal{U}_{ij}$ and claim a portion of R_{ij}. Therefore, the self-interested clients will deliberately misreport their private information to gain a greater portion of the allocated reward and maximize their own utilities. Then, the utility function of client n can be formulated as:

$$\arg\max_{(i,j)} u_n = \frac{Z_n}{\sum_{m\in\mathcal{U}_{ij}} Z_m} R_{ij} - c_n Z_n. \tag{6}$$

Model accuracy loss decreases with more training data from clients, following a convex pattern that reflects diminishing marginal effects in economics. Intuitively, larger data contribution and fairer data yield better model performance. Therefore, the value generated by the client set $\mathcal{U}_{ij}$ is $\ln(1 + \alpha\sum_{n\in\mathcal{U}_{ij}} f_n Z_n)$. The server will determine the optimal reward R_{ij} by maximizing its utilities as follows:

$$\arg\max_{R_{ij}} u_s = \ln(1 + \alpha f_i \sum_{n\in\mathcal{U}_{ij}} Z_n) - R_{ij}. \tag{7}$$

With the above analyses, we can formulate a False Report (FR) Game for FL client participation with incomplete information, where clients are unaware of the strategies chosen by other clients when making decisions. In this game, the server, as a leader, broadcasts rewards, while each client, as a follower, strategically submits her preferred type.

Definition 2 *(FR) A FR Game can be formulated as a 3-tuple $(\mathcal{N}, \mathcal{S}, \mathcal{U})$, i.e., a client set $\mathcal{N}$, a reporting strategy set $\mathcal{S}$, and a utility set $\mathcal{U}$.*

- $\mathcal{N} = \{1, \cdots, N\}$*, the set of N clients.*
- $\mathcal{S} = \{(i, j))|i \in [1, I] \cap \mathbb{Z}, j \in [1, J] \cap \mathbb{Z}\}$*, where each client n submits her preferred type (i, j).*
- $\mathcal{U} = \{u_1, \dots u_N\}$*, where each client $n \in \mathcal{N}$ aims to maximize her utility u_n.*

Next, we define the Nash Equilibrium (NE) of the FR game and derive its corresponding NE.

Definition 3 *(NE) A NE $\mathcal{S} = \{s_n^*|\forall n \in N\}$ of a FR game can be achieved if $u(s_n^*, \mathbf{s}_{-n}^*) \geq u(s_n, \mathbf{s}_{-n}^*), \forall n \in \mathcal{N}$.*

At NE, no client can benefit more by deviating from her optimal strategy. We here derive the best response of client n that maximizes her utility given the strategies of others.

Lemma 1 *A NE of the FR game can be achieved when each client n reports her type as $(1, j')$ to maximize her utility, i.e., $f_n = f_1$ and $c_n = c_{j'}$, where $j' = \arg\max_j [1 - \frac{(U_{ij}+1)c_n}{U_{ij}c_j}]^2[1 - \frac{U_{ij}c_j}{\alpha(U_{ij}-1)f_i}]$.*

From Lemma 1, clients prefer to misreport their types as a dominant strategy when clients are not in $\mathcal{U}_{1j'}$, which brings the highest utility to the client regardless of others' strategies. Therefore, it's necessary to design an incentive mechanism that ensures clients report their types truthfully.

Contract Definition Clients exploit fake GF reporting to strategically misreport their types for utility maximization. To encourage clients cooperation and truthful reporting, we introduce contract theory to address this issue. We propose a privacy pair-based contract to ensure clients report their types truthfully, visualizing multiple private information items in a multi-dimensional space. The server generates $I \times J$ contract packages $\{\phi_{ij}\}$ for clients with f_i and c_j in $\mathcal{U}_{ij}$. Each contract package is defined as follows:

Definition 4 *A privacy pair-based contract package ϕ_{ij} is represented as a 3-tuple $((f_i, c_j), z_{ij}, r_{ij})$, i.e., a two-dimensional private information of a client with GF f_i and cost c_j, a required data size z_{ij}, and a reward r_{ij}.*

Each client strategically selects her preferred contract package at the beginning of training. If a client chooses the contract item ϕ_{ij}, it indicates that she belongs to group $\mathcal{U}_{ij}$, with GF report f_i, cost c_j, and data contribution z_{ij} for local training. Empirically, higher GF indicates better data quality, so the client receives $f_n r_{ij}$ reward per global round. If client n accepts the contract ϕ_{ij}, her utility is

$$u_n^{ij} = f_n r_{ij} - c_n z_{ij}. \tag{8}$$

According to Eq. (8), each client can maximize her own utility by selecting her optimal contract ϕ_{ij}, and we can formulate this as the following optimization problem:

$$\begin{aligned} \mathbb{P}_2 : \max_{ij} u_n^{ij} &= f_n r_{ij} - c_n z_{ij}, \\ \text{s.t. } z_{ij} &\leq D_n, i \in [1, I] \cap \mathbb{Z}, j \in [1, J] \cap \mathbb{Z}. \end{aligned} \tag{9}$$

Optimal Design of Contract To ensure that clients select the appropriate contract package based on their truthful types, the privacy pair-based contract design must satisfy Individual Rationality (IR) (Tang, Peng, and Wong 2024) and Incentive Compatibility (IC) (Y. Saputra 2022).

Definition 5 *(IR) The IR criterion indicates that each client $n \in \mathcal{N}$ obtain a non-negative utility by receiving ϕ_{ij}, i.e.,*

$$u_n^{ij} \geq 0, \forall n \in \mathcal{N}, i \in [1, I] \cap \mathbb{Z}, j \in [1, J] \cap \mathbb{Z}. \tag{10}$$

Definition 6 *(IC) The IC criterion indicates that each client $n \in \mathcal{N}$ can maximize her utility by receiving contract package ϕ_{ij} for her truthful type, rather than other $\phi_{i'j'}$, i.e.,*

$$u_n^{ij} \geq u_n^{i'j'}, \forall n \in \mathcal{N}, i \in [1, I] \cap \mathbb{Z}, j \in [1, J] \cap \mathbb{Z}. \tag{11}$$

From the aforementioned IR and IC constraints, we can reformulate $\mathbb{P}_2$ as:

$$\begin{aligned} \mathbb{P}_3 : \max_{ij} u_n^{ij} &= f_n r_{ij} - c_n z_{ij}, \\ \text{s.t. } &\begin{cases} u_n^{ij} \geq 0, \\ u_n^{ij} \geq u_n^{i'j'}, \\ z_{ij} \leq D_n, i \in [1, I] \cap \mathbb{Z}, j \in [1, J] \cap \mathbb{Z}. \end{cases} \end{aligned} \tag{12}$$

Since IC constraints with multiple clients are complex to solve directly, we simplify these constraints and derive the following specific properties.

Lemma 2 *Let* $\{\phi_{ij}\} = \{((f_i, c_j), z_{ij}, r_{ij})\}$ *denote any contract that satisfies IC constraint, where* $i \in [1, I] \cap \mathbb{Z}$ *and* $j \in [1, J] \cap \mathbb{Z}$*, the following properties are unheld:*

- *If* $f_i \geq f_{i'}$ *and* $c_j \leq c_{j'}$*, then* $z_{ij} \geq \max(z_{ij'}, z_{i'j})$ *and* $r_{ij} \geq \max(r_{ij'}, r_{i'j})$.
- *The relationship between any two neighboring contract packages is* $\underline{r_{ij}} \leq r_{ij} \leq \overline{r_{ij}}$*, where* $\underline{r_{ij}} = \max\{r_{(i+1)j} + \frac{c_j}{f_i}(z_{ij} - z_{(i+1)j}), r_{i(j+1)} + \frac{c_j}{f_i}(z_{ij} - z_{i(j+1)})\}$*, and* $\overline{r_{ij}} = \min\{r_{(i+1)j} + \frac{c_j}{f_{i+1}}(z_{ij} - z_{(i+1)j}), r_{i(j+1)} + \frac{c_{j+1}}{f_i}(z_{ij} - z_{i(j+1)})\}$.

Using the abvenmentioned simplified constraints, $\mathbb{P}_3$ can be reformulated as:

$$\mathbb{P}_4 : \max_{ij} u_n^{ij} = u_n^{ij} = f_n r_{ij} - c_n z_{ij}, \quad \text{s.t.} \begin{cases} f_n r_{ij} - c_n z_{ij} \geq 0, \\ \underline{r_{ij}} \leq r_{ij} \leq \overline{r_{ij}}, \\ z_{ij} \leq D_n, i \in [1, I] \cap \mathbb{Z}, j \in [1, J] \cap \mathbb{Z}. \end{cases} \tag{13}$$

In real-world settings with privacy constraints, the server knows only the type distribution, where p_{ij} is the probability of a client belonging to $\mathcal{U}_{ij}$ (Xiao et al. 2023). According to Eq.(7), the server's utility obtained from group $\mathcal{U}_{ij}$ is

$$u_s = \sum_{n=1}^{N} p_{ij}^n [\ln(1 + \alpha n f_i z_{ij}) - n f_i r_{ij}], \tag{14}$$

where $p_{ij}^n = \binom{N}{n}(p_{ij})^n (1 - p_{ij})^{N-n}$. Based on Eq. (14), the optimal contract design of $\mathbb{P}_4$ is as follows.

Theorem 1 *The optimal privacy pair-based contract takes the form* $\{\phi_{ij}\} = \{((f_i, c_j), z_{ij}^*, r_{ij}^*)\}$*, where the optimal data contribution* z_{ij}^* *satisfies* $\sum_{n=1}^{N} \frac{p_{ij}^n n}{1 + \alpha n f_i z_{ij}^*} = c_j \frac{\sum_{n=1}^{N} p_{ij}^n n}{\alpha f_i}$*, and the optimal reward* r_{ij}^* *is given by*

$$r_{ij}^* = \begin{cases} \frac{c_j z_{ij}^*}{f_i}, & \text{if } i = I, j = J, \\ \frac{c_j z_{ij}^*}{f_i} + c_j \sum_{n=i+1}^{I} \triangle \frac{1}{f_n} z_{nj}^*, & \text{if } i < I, j = J, \\ \frac{c_j z_{ij}^*}{f_i} + \frac{1}{f_i} \sum_{n=j+1}^{J} \triangle c_n z_{in}^*, & \text{if } i = I, j < J, \\ \frac{c_j z_{ij}^*}{f_i} + \max\{c_j \sum_{n=i+1}^{I} \triangle \frac{1}{f_n} z_{nj}^*, \\ \frac{1}{f_i} \sum_{n=j+1}^{J} \triangle c_n z_{in}^*\}, & \text{if } i < I, j < J, \end{cases} \tag{15}$$

here $\triangle \frac{1}{f_i} = \frac{1}{f_i} - \frac{1}{f_{i-1}}$*, and* $\triangle c_j = c_j - c_{j-1}$.

Therefore, the optimal privacy pair-based contract ensures that actively unreliable clients report multiple privacy information truthfully, without assuming linear relationships among them. The server can maximize its utility by assigning more rewards and requiring larger data contributions from clients with lower costs and higher GF.

Robust Group Fairness Estimation Design

Although the DP metric is a convenient method to estimate GF, data perturbations can cause inaccuracies, leading to inaccurate GF metrics being reported by passive clients in fragility scenarios. (Jiang et al. 2022). To overcome these shortcomings, we extend the standard DP metrics to a robust demographic parity (RDP), by employing the kernel density estimator (Tang and He 2015) and Nadaraya-Watson kernel estimator (Nadaraya 1964; Bierens 1994) to provide smooth density and regression curves, ensuring accuracy and effectively addressing multi-dimensional data perturbation.

We omit the subscript n here to focus on the expected GF of the model across all samples from all clients. We reorder the data samples $(\tilde{y}, s)$ based on s and then uniformly divide them into L subsets. The subset interval is $[s_{l-1}, s_l]$, where l is the subset index. Let $s_l = \frac{s_{l-1} + s_l}{2}$ approximate represent the values in each subset according to histogram estimation (Blandin and Kash 2023). Each subset contains D_{s_l} clients, and the total number of clients is given by $\sum_{l=0}^{L} D_{s_l} = D$. $K(\cdot)$ is a symmetric one-dimensional smoothing kernel function, and $h > 0$ is the kernel bandwidth. The definition of RDP is as follows.

Definition 7 *(RDP) The RDP metric* b_r *under a continuous sensitive attribute* $s \in [0, 1]$ *is defined as:*

$$\begin{cases} b_r = \int_0^1 |b(s) - \bar{b}| p(s) ds, \\ s.t. \begin{cases} b(s) = \frac{\sum_{l=1}^{L} \Delta(s_l) K(\frac{s_l - s}{h})}{\sum_{l=1}^{L} K(\frac{s_l - s}{h})}, \\ \bar{b} = \frac{\sum_{l=1}^{L} \Delta(s_l)}{L}, \\ p(s) = \frac{1}{Lh} \sum_{l=1}^{L} K(\frac{s_l - s}{h}), \end{cases} \end{cases} \tag{16}$$

where $\Delta(s_l) = \int_{\tau}^{\infty} p(\tilde{y}|s) d\tilde{y} - \sum_{l=1}^{L} \frac{D_{s_l}}{D} \int_{\tau}^{\infty} p(\tilde{y}|s_l) d\tilde{y}$ *measures the bias in a sensitive group* s_l*, and the probability distribution function of the output scores in a sensitive group is* $p(\tilde{y}|s) = \frac{1}{D_s h} \sum_{m=1}^{D_s} K(\frac{\tilde{y} - \tilde{y}^m}{h})$.

To illustrate the robustness of the proposed metrics, we derive the following theorem that the lower bound of the GF estimation deviation is ϵ when perturbation occurs. Let $\Delta'(s_l)$ denotes the GF estimation under perturbation, $\Delta(s_l)$ represent the GF estimation without perturbation for the sensitive attribute s_l, and assume $\tilde{y}$ follows a uniform distribution under s_l, i.e., $\tilde{y} \sim U(\alpha_{s_l}, \beta_{s_l})$.

Theorem 2 *The GF estimation deviation exceeds* ϵ *with probability at most*

$$P(|\Delta'(s_l) - \Delta(s_l)| \geq \epsilon) \leq 2 \exp(-\frac{2\epsilon^2}{D_{s_l} c^2}), \tag{17}$$

where $c = \frac{4M^2(\beta_{s_l} - \tau)^2 (D - L D_{s_l})^2}{h^2 D_{s_l}^2 D^2}$*, and* M *denotes the upper bound of the kernel function* $K(\cdot)$.

Theorem 2 implies that a higher D_{s_l} leads to tighter probability bounds, while a higher β_{s_l} relaxes them. Due to OursFed's ability to incentivize clients with higher GF to contribute more data, the proposed metric significantly decreases the instability effect on output scores. Sensitive attributes can be similarly analyzed for perturbation effects, and thus detailed analysis is omitted here.

Theoretical Analysis of OursFed

In this section, we demonstrate the efficacy of OursFed by rigorously analyzing its convergence, and analyze the GF through two metrics: the stability and the accuracy deviation.

Convergence

By introducing the following standard assumptions with detailed description refer to Appendix, we prove that OursFed can converge to an approximate optimal solution.

Assumption 1 *(Fan et al. 2022; Lu et al. 2023) A differential loss function $l : \mathbb{R}^n \to \mathbb{R}$ satisfying: γ-smooth, β-strongly convex, and L-Lipschitz condition, while a stochastic gradient function $g : \mathbb{R}^n \to \mathbb{R}^n$ satisfying: a bounded expectation of the norm and the second-order norm.*

Theorem 3 *Under Assumption 1, OursFed converges to an approximate optimal solution ω^* as follows:*

$$l(\omega^{t+1}) - l(\omega^*) \leq (1 - \frac{\beta\eta\underline{g}}{2L})^{t-1}(l(\omega^1) - l(\omega^*) - \frac{L^2\eta\nu^2}{\beta\underline{g}}) + \frac{L^2\eta\nu^2}{\beta\underline{g}}. \quad (18)$$

The Stability of GF

For the ease of mathematical exposition, we restate the hypothesis function of FL as follows:

$$h^q(\mathbf{x}) = \sum_{n\in\mathcal{N}} \theta_n^q h_n(\mathbf{x}), q \in \{0, 1\}, \quad (19)$$

where $\theta_n^q = \frac{z_n f_n^q}{\sum_{i\in\mathcal{N}} z_i f_i^q}$, $q = 0$ denoting the standard FL, and $q = 1$ as the proposed GF-aware framework. Given clients' local group biases $\{b_n | n \in \mathcal{N}\}$, and their assigned weights $\{\theta_n^q | n \in \mathcal{N}, q \in \{0, 1\}\}$, the global group bias is regarded as a convex combination $b^q = \sum_{n\in\mathcal{N}} \theta_n^q b_n$, where deterioration of b_n can change b^q at most θ_n^q.

The stability of GF (SGF) of the global model depends on the model aggregator in Eq. (19), which quantifies the impact of sharp fairness deterioration in clients on the global model's GF via McDiarmid's inequality (Long et al. 2024; McDiarmid 1989). To compare SGF differences under the aggregator, we assume all clients initially have perfect GF ($f_n^q = 1$) and denote the expected global GF as $\mathbb{E}(b^q)$. Then, we analyze the impact of the proposed framework on the GF when a subset of clients $\mathcal{N}_m \in \mathcal{N}$ deteriorates by $\xi_m \in (0, 1]$, i.e., $f_m = 1 - \xi_m, m \in \mathcal{N}_m$.

Theorem 4 *OursFed offers probabilistic guarantees for SGF:*

$$P(b^q - \mathbb{E}(b^q) < \epsilon) > 1 - \delta, \quad (20)$$

with the deviation probability bound δ given by:

$$\delta = \begin{cases} 0, & \text{if } q = 1, \xi_m \to 1, \\ \delta^q, & \text{otherwise}, \end{cases} \quad (21)$$

where $\delta^q = \exp(-2\varepsilon^2(\frac{\sum_{n\in\mathcal{N}/\mathcal{N}_m} z_n}{\sum_{m\in\mathcal{N}_m} z_m(1-\xi_m)^q} + 1))$.

According to Theorem 4, OursFed achieves superior SGF by mitigating GF deterioration, as indicated by $\delta^1 \geq \delta^0$. As the level of deterioration increases, represented by a larger $\sum_{m\in\mathcal{N}_m} z_m\xi_m$, OursFed maintains higher GF. In the extreme case where all ξ_m approach one, the global GF remains unaffected and δ equals zero.

The Accuracy Deviation

The metrics of accuracy deviation is used to analyze the probability of different predictions between the proposed and standard FL for the same input, i.e., $P(h^1(\mathbf{x}) \neq h^0(\mathbf{x})) \leq \varepsilon$. Empirically, a higher degree of deviation ε from the current classic function indicates worse accuracy. We reference α to describe the classifier confidence. A hypothesis value within the interval $[0, \tau - \alpha] \cup [\tau + \alpha, 1]$ indicates high confidence. In contrast, a hypothesis value within the interval $[\tau - \alpha, \tau + \alpha]$ indicates an ambiguous prediction result. We can derive the accuracy deviation as follows.

Theorem 5 *The exponentially growing bound of the accuracy deviation under $h^q(\mathbf{x}) \sim (\mathbb{E}(h(\mathbf{x}), \mathbb{D}(h(\mathbf{x})))$ is*

$$P(h^1(\mathbf{x}) \neq h^0(\mathbf{x})) \leq 2\exp(-\frac{8\alpha^2(\sum_{n\in\mathcal{N}} z_n)^2}{(\sum_{m\in\mathcal{N}_m} z_m)^2}) + 2\gamma, \quad (22)$$

where $P(|h^q(\mathbf{x}) - \tau| < \alpha) < \gamma$.

Theorem 5 implies that the probability of different outputs from h^1 and h^0 for the same input is exponentially related to the number of training samples from clients with GF deterioration. OursFed significantly reduces the probability by incentivizing clients with higher GF to contribute more data.

Experiments

In this section, we employ two real datasets to investigate the performance of OursFed and two synthetic datasets to show its effectiveness in addressing distrust and fragility.

Experimental Setup

Real datasets. We evaluate the performance of OursFed on two real datasets: **COMPAS** (Larson et al. 2016) and **Adult** (Dua and Graff 2017). For COMPAS, we use a 3-layer neural network (NN) with 16 nodes per hidden layer and a learning rate of 0.001. For Adult, we employ a 2-layer NN with 10 nodes in the hidden layer and a learning rate 0.01. Detailed experimental settings are provided in the Appendix.

Synthetic datasets. For distrust scenarios, we generate the data (f, c) from a bivariate uniform distribution $N([0, 1] \times [0, 1])$. For fragility scenarios, we sample the data $(\tilde{y}, s)$ from a bivariate Gaussian distribution $N(\mu, \Sigma)$ with mean $\mu = [0.2, 0.4]$ and covariance matrix $\sigma = [[0.1, 0.05], [0.05, 0.2]]$, and generate few perturbation data from uniform distribution $N([0.9, 1] \times [0, 0.1])$.

Baselines To evaluate model performance, we compare OursFed with three typical baselines: **FedAvg** (McMahan et al. 2017), **FedAvg-W** (Yang et al. 2019), and **FairBatch** (Li et al. 2022), as well as an additional custom baseline, **OGF**, which is developed for comparative analysis. Detailed description of the baselines are provided in the Appendix.

Experimental Results

Results on model performance. Figures 2(a) and 2(b) present the accuracy and GF estimates on COMPAS and Adult, comparing OursFed against four baselines over 100 random trials. FedAvg exhibits poor accuracy and GF, as it ignores data contribution and GF. FedAvg-W considers data

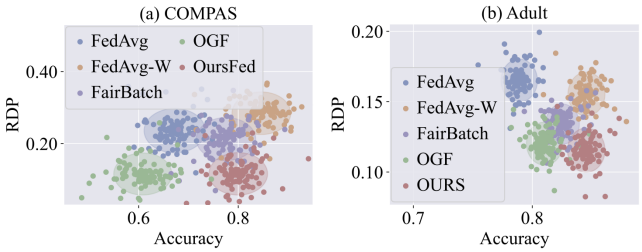

Figure 2: Accuracy-GF evaluate.

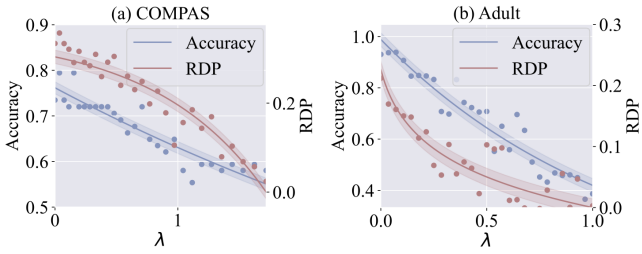

Figure 3: Accuracy-GF trade-off.

Methods	RDP		Accuracy	
	COMPAS	Adult	COMPAS	Adult
FedAvg	0.24±0.04	0.17±0.01	0.68±0.04	0.79±0.01
FedAvg-W	0.29±0.02	0.16±0.01	0.84±0.04	0.85±0.01
FairBatch	0.21±0.03	0.13±0.02	0.77±0.04	0.83±0.01
OGF	0.11±0.04	0.12±0.01	0.60±0.01	0.81±0.02
OursFed	**0.12±0.03**	**0.12±0.01**	**0.80±0.04**	**0.85±0.01**

Table 1: Results on COMPAS and Adult.

contribution but ignores GF, achieving the highest accuracy of 0.85 on COMPAS and 0.845 on Adult as shown in Table 1, but with poor GF. A similar issue arises when using OGF, it displays poor accuracy but the best GF of 0.114 on COMPAS and 0.12 on Adult, as it only considers GF without data contribution. FairBath reduces biases by balancing data distribution and batching to improve GF compared to FedAvg-W. OursFed achieves near-optimal GF value of 0.117 on COMPAS and optimal GF value of 0.115 on Adult, and sub-optimal accuracy of 0.796 on COMPAS and 0.845 on Adult. It achieves the best GF with minimal accuracy loss, demonstrating the optimal trade-off between accuracy and fairness. This is because it jointly considers accuracy and GF, incentivizing high-GF clients contribute more data via contract. OursFed improves GF by at least 28.61% with an average accuracy loss of no more than 2.7%.

Figures 3(a) and 3(b) show the tradeoff between accuracy and GF on COMPAS and Adult using turning knob λ, with each point representing an average value over 5 trials with different random seeds. Specifically, greater emphasis on GF leads to better fairness but lower accuracy. This is due to the existence of heterogeneity in data distribution, quantity and quality among different clients.

Results on overcoming distrust. Figure 4 shows the allocation of data contribution and rewards among different client types. Clients are categorized based on GF and cost, each with 10 types, yielding 10×10 types in total, and each cell indicating the data contribution required from each client type and the reward they can receive. Clients with higher GF and lower costs contribute more data as shown in Figure 4(a), and can receive greater reward assignments as shown in Figure 4(b). On the one hand, the privacy pair-based contract can provide higher data quality and save the server's budget to maximize the server's utility. On the other hand, it can prevent accuracy deviation.

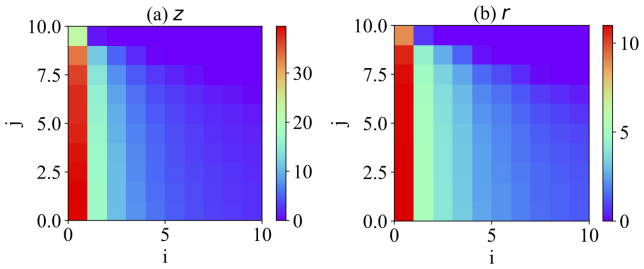

Figure 4: Contract design on data contribution and reward.

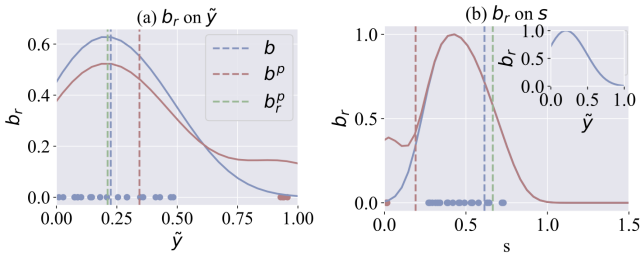

Figure 5: Perturbation on labels and sensitive attributes.

Results on overcoming fragility. Figure 5 shows the GF estimation under perturbation on $\tilde{y}$ or s. As shown in Figure 5(a), the group bias estimate b_r^p under RDP with perturbation on $\tilde{y}$ is closer to the expected group bias b without perturbation than the group bias estimate b^p under DP with perturbation. This is due to the smooth density function, which alleviate the effect of perturbation on $\tilde{y}$. As shown in Figure 5(b), b_r^p with perturbation on s is closer to b without perturbation than b_p, since RDP provide smooth regression on s to ensure robustness and accuracy of GF estimation.

Conclusion

In this paper, we have proposed OursFed to tackle the challenges of distrust and fragility in FL. To address client unreliability from strategic GF misreporting and inaccurate passive estimation, we proposed a privacy pair-based contract that encourages truthful reporting via multivariate analysis, and a robust GF estimation method that improves evaluation accuracy under multi-dimensional data perturbations. Particularly, the convergence of OursFed was proved, and its GF was analyzed through SGF and accuracy deviation. Finally, extensive experiments were conducted on both real and synthetic datasets validate the effectiveness of OursFed.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62325304, 62372343, 62402352, 62562050, 62172313, U22B2046, 62072411), in part by the Natural Science Foundation of Jiangsu Province of China (No. BK20253020), and in part by the Key Research and Development Program of Zhejiang Province (No. 2025C01055), and in part by the Open Fund of Key Laboratory of Social Computing and Cognitive Intelligence (Dalian University of Technology), Ministry of Education (No. SCCI2024TB02).

References

Baudry, D.; Richard, H.; Cherifa, M.; Calauzènes, C.; and Perchet, V. 2024. Optimizing the coalition gain in Online Auctions with Greedy Structured Bandits. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 15–24. Vancouver Convention Centre, Vancouver, BC, Canada: Curran Associates, Inc.

Bierens, H. 1994. The nadaraya-watson kernel regression function estimator. *Topics in Advanced Econometrics Cambridge University Press*, 212–247.

Blandin, J.; and Kash, I. 2023. Generalizing Group Fairness in Machines Learning via utilities. *Journal of Artificial Intelligence Research*, 78: 747–780.

Cao, S.; Cheng, R.; and Wang, Z. 2025. AGR: Age Group fairness Reward for Bias Mitigation in LLMs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2025)*. Hyderabad, India: IEEE.

Cho, J.; Hwang, G.; and Suh, C. 2020. A Fair Classifier Using Kernel Density Estimation. In *Advances in Neural Information Processing Systems (NeurIPS-33)*, 15088–15099. Vancouver, Canada: Curran Associates, Inc.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5: 1556–1571.

Deng, Y.; Lyu, F.; Ren, J.; Chen, Y.; Yang, P.; Zhou, Y.; and Zhang, Y. 2022. Improving Federated Learning With Quality-Aware User Incentive and Auto-Weighted Model Aggregation. *IEEE Transactions on Parallel and Distributed Systems*, 33: 4515–4529.

Ding, N.; Gao, L.; and Huang, J. 2023. Joint Participation Incentive and Network Pricing Design for Federated Learning. In *IEEE International Conference on Computer Communications (INFOCOM-40)*, 1–10. Vancouver, Canada: IEEE.

Dua, D.; and Graff, C. 2017. Machine Learning Repository.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Ezzeldin, Y.; Yan, S.; He, C.; Ferrara, E.; and Avestimehr, A. 2023. FairFed: Enabling Group Fairness in Federated Learning. In *Association for the Advancement of Artificial Intelligence Conference (AAAI-37)*, volume 37, 2061–2070. Washington, D.C., USA.

Fan, Z.; Fang, H.; Zhou, Z.; Pei, J.; Friedlander, M.; Liu, C.; and Zhang, Y. 2022. Fairness for Data Valuation in Horizontal Federated Learning. In *IEEE International Conference on Data Engineering (ICDE-38)*, 2361–2372. Dallas, TX, USA: IEEE.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Guo, J.; Liu, Z.; Tian, S.; Huang, F.; Li, J.; Li, X.; Igorevich, K.; and Ma, J. 2023. TFL-DT: A Trust Evaluation Scheme for Federated Learning in Digital Twin for Mobile Networks. *IEEE Journal on Selected Areas in Communications*, 41: 3548–3560.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS-30)*, 3315–3323. Barcelona, Spain: Curran Associates, Inc.

Hu, M.; Yang, W.; Luo, Z.; Liu, X.; Zhou, Y.; Chen, X.; and Wu, D. 2022. AutoFL: A Bayesian Game Approach for Autonomous Client Participation in Federated Edge Learning. *IEEE Transactions on Mobile Computing*, 23: 194–208.

Huang, G.; Chen, X.; Ouyang, T.; Ma, Q.; Chen, L.; and Zhang, J. 2022. Collaboration in Participant-Centric Federated Learning: A Game-Theoretical Perspective. *IEEE Transactions on Mobile Computing*, 22: 6311–6326.

Jiang, Z.; Han, X.; Fan, C.; Yang, F.; Mostafavi, A.; and Hu, X. 2022. Generalized Demographic Parity for Group Fairness. In *International Conference on Learning Representations (ICLR-10)*, 1095–1114. Vienna, Austria: ICLR.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2018. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133: 237–293.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica*.

Li, X.; Qu, Z.; Zhao, S.; Tang, B.; Lu, Z.; and Liu, Y. 2021. LoMar: A Local Defense Against Poisoning Attack on Federated Learning. *IEEE Transactions on Dependable and Secure Computing*, 20: 437–450.

Long, C.; Hsu, H.; Alghamdi, W.; and Calmon, F. 2024. Individual Arbitrariness and Group Fairness. In *Advances in Neural Information Processing Systems (NeurIPS-38)*, 1–12. San Francisco, CA, USA: Curran Associates, Inc.

Lu, R.; Zhang, W.; Wang, Y.; Li, Q.; Zhong, X.; Yang, H.; and Wang, D. 2023. Auction Based Clustered Federated Learning in Mobile Edge Computing System. *IEEE Transactions on Parallel and Distributed Systems*, 34: 1145–1158.

McDiarmid, C. 1989. On the Method of Bounded Differences. *Surveys in combinatorics*, 204: 145–188.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. 2017. Communication-efficient learning of deep networks from decentralized data. *Artificial Intelligence and Statistics*, 54: 1273–1282.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6): 1–35.

Murhekar, A.; Yuan, Z.; Chaudhury, B.; Li, B.; and Mehta, R. 2024. Incentives in Federated Learning: Equilibria, Dynamics, and Mechanisms for Welfare Maximization. In *Advances in Neural Information Processing Systems (NeurIPS-38)*, 1–14. San Francisco, CA, USA: Curran Associates, Inc.

Nadaraya, E. 1964. On estimating regression. *Theory of Probability and Its Applications*, 9(1): 141–142.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 469–481.

Saha, R.; Misra, S.; Chakraborty, A.; Chatterjee, C.; and Deb, P. 2022. Data-Centric Client Selection for Federated Learning over Distributed Edge Networks. *IEEE Transactions on Parallel and Distributed Systems*, 34: 675–686.

Shi, Y.; Yu, H.; and Leung, C. 2023. Towards Fairness-Aware Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 37: 1–17.

Tang, B.; and He, H. 2015. Kerneladasyn: Kernel based adaptive synthetic data generation for imbalanced learning. In *IEEE Congress on Evolutionary Computation (CEC-14)*, 664–671. Sendai, Japan: IEEE.

Tang, M.; Peng, F.; and Wong, V. 2024. A Blockchain-Empowered Incentive Mechanism for Cross-Silo Federated Learning. *IEEE Transactions on Mobile Computing*, 40: 1–13.

Wang, X.; Zhao, Y.; Qiu, C.; Liu, Z.; Nie, J.; and Leung, V. 2022. InFEDge: A Blockchain-Based Incentive Mechanism in Hierarchical Federated Learning for End-Edge-Cloud Communications. *IEEE Journal on Selected Areas in Communications*, 40: 3325–3342.

Wang, Z.; Hu, Q.; Li, R.; Xu, M.; and Xiong, Z. 2023. Incentive Mechanism Design for Joint Resource Allocation in Blockchain-based Federated Learning. *IEEE Transactions on Parallel and Distributed Systems*, 34: 1536–1547.

Xiao, M.; Xu, Y.; Zhou, J.; Wu, J.; Zhang, S.; and Zheng, J. 2023. AoI-aware Incentive Mechanism for Mobile Crowdsensing using Stackelberg Game. In *IEEE International Conference on Computer Communications (INFOCOM-42)*, 1–10. Istanbul, Turkey: IEEE.

Y. Saputra, D. H. T. V. E. D. S. C., D. Nguyen. 2022. Federated Learning Meets Contract Theory: Economic-efficiency Framework for Electric Vehicle Networks. *IEEE Transactions on Mobile Computing*, 21: 2803 – 2817.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Transactions on Knowledge and Data Engineering*, 35: 1–20.

Zhang, T.; Lam, K.; Zhao, J.; Li, F.; Han, H.; and Jamil, N. 2023. Enhancing Federated Learning with spectrum allocation optimization and device selection. *IEEE Transactions on Networking*, 31: 1981–1996.

Zhao, Y.; Gong, X.; and Mao, S. 2023. Truthful Incentive Mechanism for Federated Learning with Crowdsourced Data Labeling. In *IEEE International Conference on Computer Communications (INFOCOM-42)*, 1–10. Istanbul, Turkey.