Mitigating Group Bias in Federated Learning for Heterogeneous Devices

Khotso Selialia
University of Massachusetts Amherst
Amherst, Massachusetts, USA

Yasra Chandio
University of Massachusetts Amherst
Amherst, Massachusetts, USA

Fatima M. Anwar
University of Massachusetts Amherst
Amherst, Massachusetts, USA

Abstract

Federated learning is emerging as a privacy-preserving model training approach in distributed edge applications. As such, most edge deployments are heterogeneous in nature, i.e., their sensing capabilities and environments vary across deployments. This edge heterogeneity violates the independence and identical distribution (IID) property of local data across clients. It produces biased global models, i.e., models that contribute to unfair decision-making and discrimination against a particular community or a group. Existing bias mitigation techniques only focus on bias generated from label heterogeneity in non-IID data without accounting for domain variations due to feature heterogeneity.

Our work proposes a group-fair FL framework that minimizes group-bias while preserving privacy. Our main idea is to leverage average conditional probabilities to compute a cross-domain group *importance weights* derived from heterogeneous training data to optimize the performance of the worst-performing group using a modified multiplicative weights update method. Additionally, we propose regularization techniques to minimize the difference between the worst and best-performing groups while ensuring through our thresholding mechanism to strike a balance between bias reduction and group performance degradation. Our evaluation of image classification benchmarks assesses the fair decision-making of our framework in real-world settings.

CCS Concepts: • **Computing methodologies** → *Machine learning*; **Distributed computing methodologies**.

Keywords: Federated Learning, Algorithmic Fairness, Group Fairness

ACM Reference Format:
Khotso Selialia, Yasra Chandio, and Fatima M. Anwar. 2024. Mitigating Group Bias in Federated Learning for Heterogeneous Devices. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3630106.3658954

ACM FAccT '24, June 3–6, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658954

1 Introduction

Federated learning (FL) is a privacy-preserving machine learning (ML) technique wherein local models are trained on decentralized edge devices (*clients*) and subsequently aggregated at the server to form a *global model*. This approach alleviates the need for raw data transfers and ensures data privacy, making it particularly well-suited for applications with privacy sensitivities, such as medical diagnosis [21, 38, 59], next-character prediction[67], activity recognition [17, 56, 66], and human emotion recognition [14, 46, 72], where preserving data security is imperative. Despite its merits, there is a growing concern regarding FL models, as they exhibit exceptional performance for certain groups while simultaneously underperforming for others (e.g., providing accurate image captioning for pristine group images than noisy group images as shown in Figure 1). A group categorizes data based on attributes such as race, gender, class, or label [7].

Group biases and discriminatory practices threaten societal well-being, undermining public confidence in ML models and their applications [7]. Research shows racial bias in electronic health records, especially in medical analysis, potentially causing treatment disparities for minority groups [68]. Biased models often result from label heterogeneity in non-IID data across clients, as discussed in works like [52, 57], arising from diverse label distributions tied to data collection device environments. For example, certain geo-regions may have varying label distributions, reducing training data volume for specific groups [8, 30].

Our work highlights *feature noise heterogeneity* as a significant source of group bias in FL models, stemming from varied noise-influenced features due to domain differences, especially in heterogeneous devices [48]. Heterogeneity leads to distinct feature distributions in local client data. For example, low-quality sensors on some devices introduce distortion like Gaussian noise, resulting in different feature distributions compared to high-quality sensor devices [47]. This inherent feature noise causes shifts in group data moments, which are statistical properties such as mean and variance within a group in a dataset [35], influencing biased model outcomes.

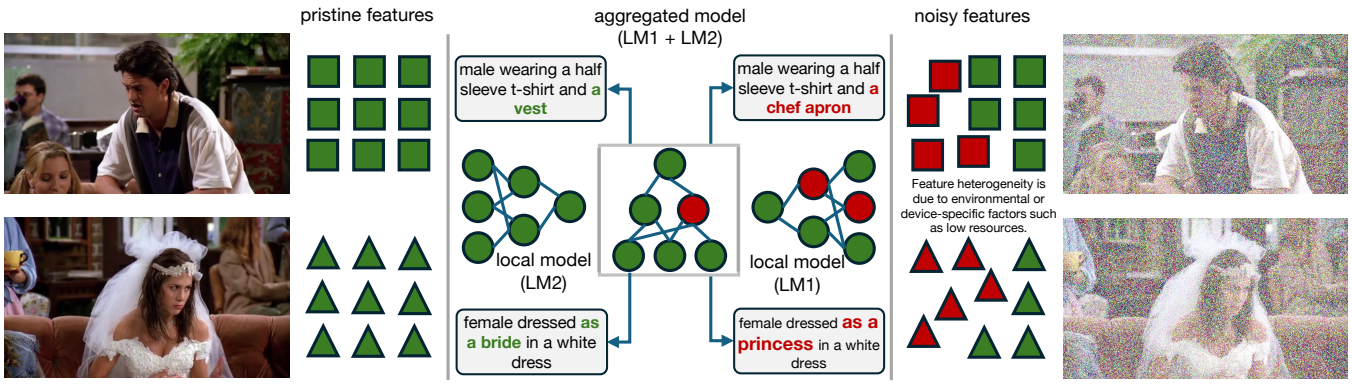


Figure 1. Illustrating the adverse effects of *feature heterogeneity (noise)* and its bias impact on image classification data [42] on an example language model (LM) in FL settings. The global LM, engaging in image captioning based on features from multiple clients, shows higher performance for images without distortions compared to those with a shift in feature distributions. This emphasizes the intricate interplay of feature heterogeneity and bias in FL, highlighting the influence of heterogeneous client datasets on the model's outcome.

Previous FL research introduces Disparate Learning Processes (DLPs) to tackle bias and fairness issues. Examples of DLPs include *in-processing* methods like [9, 11, 12, 15, 16, 18, 23, 31, 43, 44, 52, 57, 61, 73, 74, 76] and *Robustness and generalization* strategies such as [34, 41]. In-processing techniques modify learning to include group fairness constraints, while robustness and generalization enhance model resilience in diverse data settings. However, DLPs don't ensure fairness in settings with feature heterogeneity, especially due to feature noise, as they don't address misaligned moments in feature distributions [35]. For DLPs that use "reweighting" with *importance weights* to adjust the model's objective function, their effectiveness relies on suitable importance weight selection [6]. Importance weights prioritize specific groups or features during training to mitigate biases and enhance fairness [6]. If not chosen carefully or not aligned with genuine sources of bias, these weights can lead to continued unfairness [6]. We propose using weights derived from noisy feature data for more efficient debiasing in FL models affected by feature noise. This work introduces learnable importance weights from heterogeneous data features to enhance fairness in training, utilizing the *multiplicative weight update (MW)* method [3] for better fairness based on feature characteristics, especially considering data characteristics with feature noise. Our approach is inspired by insights from social science, particularly addressing discrimination as a health disparity determinant [36]. By incorporating learnable importance weights, we aim to mitigate biases across demographic groups, contributing to a more equitable FL framework.

The efficacy of *importance weighting* diminishes due to exploding weight norms from the empirical risk scaling with importance weights, especially in large models, risking overfitting [6]. To tackle this, we propose using neural network regularization techniques [55] in *Multiplicative Weight update with Regularization (MWR)* to mitigate *group bias*. Additionally, methods using *importance weighting* may introduce unfairness by overly emphasizing poorly-performing groups, potentially reducing the performance of better-performing groups to minimize overall variability [13]. To address this issue, we present a heuristic approach for deriving *importance weights* that mitigate group bias while maintaining a performance threshold for better-performing groups, preventing their performance from dropping below a desirable level. We summarize our contributions below:

- **Enabling Privacy-preserving Group Fairness**: We highlight the notion of group fairness across clients in FL settings and propose a *Multiplicative Weight (MW)* update method to mitigate bias due to feature heterogeneity. Our approach requires an estimate of the global group importance weights, which we compute as a mixture of cross-domain likelihood estimates of heterogeneous local data across clients in a privacy-preserving manner.
- **Ensuring Optimality through Regularization**: We extend our approach by incorporating the L1 regularization technique to increase its effectiveness in mitigating group bias, which we call *MWR*. It combats diverging weight norms that fail to converge to a model that optimizes worst group performance.
- **Satisfying Worst- and Best-group Performance**: We ensure that *MWR* optimizes the performance of the worst-performing group while also keeping the performance of the best-performing group above a desirable threshold.
- **Implementation and Evaluation**: We implement and evaluate the *MWR* method against existing bias-mitigation techniques on commonly used state-of-the-art image classification FL benchmark datasets (CIFAR10 [37], MNIST [40],

FashionMNIST [71], USPS [32], SynthDigits [24], and MNIST-M [24]). Our findings show that *MWR* outperforms baseline methods, boosting the accuracy of the worst group's performance up to 41% without substantially degrading the best group's performance.

2 Background and Related Work

2.1 Bias in Machine Learning.

Bias in ML refers to a model favoring specific individuals or groups, leading to unfair outcomes [51]. Common sources of bias in centralized learning include prejudice, underestimation, and negative legacy [1, 8, 49]. Techniques such as pre-processing, in-processing, and post-processing [22, 26, 33] have effectively mitigated centralized learning bias. However, applying centralized learning techniques in FL is challenging due to privacy concerns, requiring access to features across clients and risking data privacy compromise.

2.2 Bias Metrics

In FL, *group bias* is assessed through three dimensions: 1) aiming for equal opportunities by evaluating the performance discrepancy in True Positive Rates (*TPR*) between groups [58, 69]; 2) optimizing the Worst-case TPR (*WTPR*) for each group [50, 58]; 3) minimizing the standard deviation of TPR (*TPSD*) to ensure fairness across groups [58, 73]. The choice of TPR as a performance metric of in assessing group group fairness aligns our approach with recent advancements in bias mitigation literature [58]. This decision stems from recognizing the critical importance of fairly detecting true positives, which cannot be addressed solely by relying on accuracy. While our primary focus is on achieving fairness with a *minimax* property (optimizing *WTPR* outcome within each group), we evaluate using various fairness metrics to ensure versatility and broad support.

2.3 Bias Mitigation

The bias mitigation work falls mainly into four categories, including: 1) *Client-fairness* techniques [12, 31, 43, 44, 52, 61], 2) *Group-fairness* techniques [9, 11, 15, 16, 18, 23, 57, 73, 74, 76], 3) *Collaborative Fairness* techniques [19, 48, 54, 75], and 4) *Robustness and Generalization* techniques [34, 41, 60].

Client fairness targets the development of algorithms leading to models that exhibit similar performance across different clients [44]. On the other hand, ***group fairness*** requires the model to perform similarly on different demographic groups [73]. Many state-of-the-art fairness techniques in FL, focusing on *client fairness* and *group fairness*, use *in-processing* methods to modify the learning process or objective function by incorporating fairness constraints [73]. In-processing involves assigning weights to the objective function from different clients or groups during training to balance the influence of the model on different groups or clients. For instance, AFL [52] optimizes the combination of worst-weighted losses from local clients, proving resilient to data with an unknown distribution. q-FFL [44] reweights loss functions to give higher weights to devices with poorer performance, addressing challenges in fair resource allocation in computer networks. TERM handles outliers and class imbalance by tilting the loss function with a designated tilting factor [43]. GIFAIR-FL [73] introduces a regularization term, regarded as loss function weighting, to guide the optimizer towards group fair solutions. Despite the benefits, in-processing techniques face challenges, particularly sensitivity to outliers and dependence on the choice of reweighting schemes. If importance weights do not align well with data characteristics, outliers introduced by noise can have a significant impact, leading to biases. Feature noise may cause alterations in the distribution of features among groups, inducing discrepancies and bias in statistical properties.

Collaborative Fairness methodologies propose compensating each client's performance based on their contribution to learning the global model, intending to align rewards with individual client input. This approach entails providing more rewards to highly contributing clients, thereby encouraging active participation in FL. Conversely, offering lower rewards helps prevent free-riders, ensuring a fair distribution of incentives [48]. It is important to note that while we discuss *Collaborative Fairness*, here does not specifically address mitigating group bias in FL, as these techniques do not inherently focus on improving group performances.

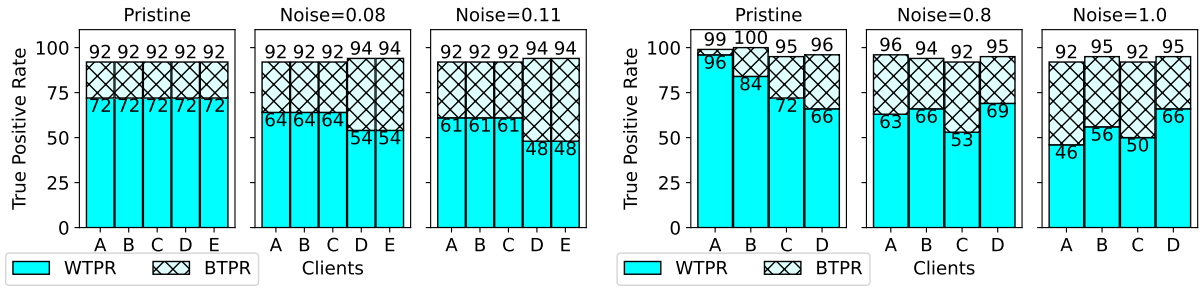
Robustness and Generalization techniques address distributional shifts in user data. For instance, FedRobust [60] trains a model to handle worst-case affine shifts, assuming that each client can express its data distribution as an affine transformation of a global distribution, focusing on group fairness. However, FedRobust requires sufficient data for each client to estimate the local worst-case shift, impacting global model performance when this condition is unmet. FedNTD tackles *catastrophic forgetting* distillation [29] but may not fully handle bias from feature noise. SCAFFOLD [34] addresses client drift in heterogeneous data by estimating update directions. However, SCAFFOLD may not correct moments in noisy feature distributions. In contrast, we use importance weights from noisy features to prioritize disadvantaged groups during training, enhancing fairness by indirectly correcting misaligned moments.

3 Preliminary Study

This section analyzes *group-bias* arising from heterogeneous feature distributions within local data across clients. The study utilizes Federated Averaging (FedAvg [45]), a widely adopted aggregation method for training global models in FL.

3.1 Experimental Setup

Applications and Datasets. Our study analyzes *group-bias* across $K \in \{4, 5\}$ clients (computers that simulate the FL



(a) CIFAR10 dataset **(b)** DIGITS dataset

Figure 2. Varied noise levels in CIFAR10 and DIGITS datasets. The notation "Noise = x" denotes the introduction of Gaussian noise with variance x," specifically applied to clients D and E in CIFAR10 and clients A and B in DIGITS.

environment, mirroring real-world heterogeneous data collection devices following recent works in FL [30, 52, 73]) using two deep learning models and two datasets. We employ the ResNet model [28] for CIFAR10 [37] image classification and a Convolutional Neural Network (CNN) on the DIGITS classification dataset, which comprises data from diverse sources with feature shifts. The goal is to replicate real-world FL scenarios with varied client data. We construct the DIGITS dataset by combining data from SynthDigits [24], MNIST-M [24], and MNIST [4].

We select these datasets to compare *group-bias* with existing bias mitigation techniques in FL. Each dataset is evenly distributed among K clients in the FL framework, ensuring equal allocation of group data points. Clients utilize replicated versions of the original benchmark test set, aligning noise feature distributions between training and test data.

We set all model parameters to match FL parameters for global model convergence under IID data settings, including label and feature noise homogeneity. Client settings include a mini-batch size of 128, a learning rate of 0.01, and 40 (for CIFAR10) and 12 (for DIGITS) training rounds.

Heterogeneous Feature Distributions. We add noise to mimic real-world distorted images that fail to share the same feature distribution with the pristine training images [25, 62, 64]. In particular, we add Gaussian noise with a variance greater than or equal to 0.03, consistent with the real-world deployments[48]. We create two different distortion levels in each dataset across K clients. For the CIFAR10, three advantaged clients (A, B, C) lack distortions, while the other two disadvantaged clients (D, E) host data with Gaussian noise of variance $var \in \{0.03, 0.07, 0.11, 0.3, 0.4, 0.8, 1.0\}$. For the DIGITS dataset, two advantaged clients (C, D) lack distortions, while the other two disadvantaged clients (A, B) host data with Gaussian noise.

3.2 Key Findings

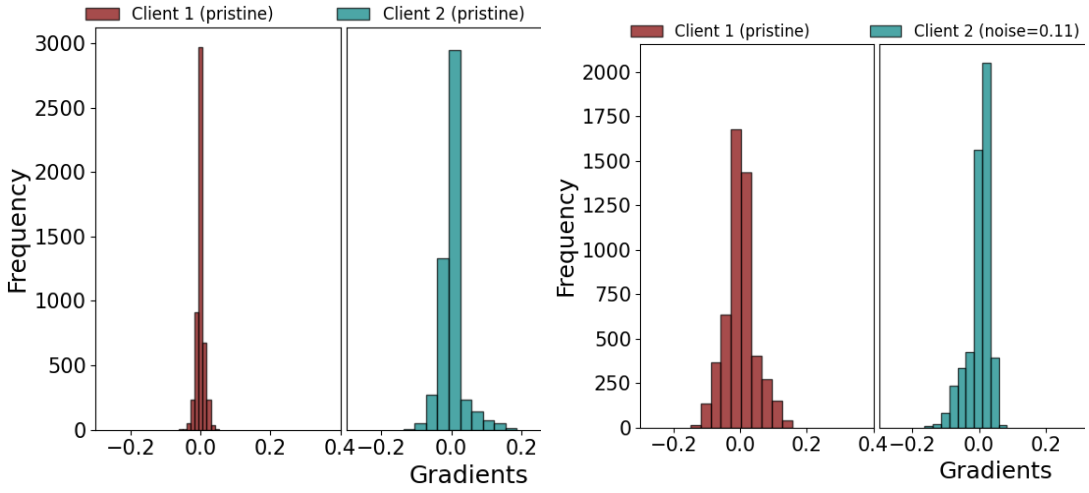
Non-IID Study. We study the FL model's unfairness by examining how the biased global model treats local groups differently for each client. We measure the TPR performance gap between the best and worst groups using each client's local test data (with a similar distortion level as the training data). Figure 2a shows *group-bias* in CIFAR10, while Figure 2b illustrates this in DIGITS. The global model's recognition of local groups varies per client, as seen in the discrepancy between their performances. Increasing Gaussian noise on a client amplifies this difference, indicating that *heterogeneous local features across clients contribute to group bias.*

Limitation of Federated Averaging. We empirically investigate how heterogeneous local data distributions affect local model gradients. Post-convergence, we extract gradients from the last linear layer of each local model across two clients. Figure 3 shows histograms of these gradients, highlighting variations across clients with heterogeneous features (3b) compared to more consistent distributions in clients with homogeneous features (3a). In 3a, a Spearman correlation [53] of 0.46 indicates strong correlation and uniformity among clients with IID features. Conversely, in Figure 3b, clients with non-IID features show a correlation of -0.14, suggesting dissimilarity.

Our non-IID study underscores the challenges in conventional FedAvg schemes, revealing consistently unfair model behavior across distinct applications and datasets. This problem emphasizes the need for bias mitigation methods to alleviate adverse outcomes, including performance degradation in critical applications like medical contexts and the inability to adapt to dynamic heterogeneous environments.

4 Methodology

The primary objective of our work is to address group bias resulting from feature heterogeneity across clients, all while preventing the leakage of sensitive data. In this section, we formally define our problem and then present our approach to mitigate group bias without substantially degrading the best group performance.



(a) Correlation=0.46 **(b)** Correlation=-0.14

Figure 3. Gradient distribution in a fully connected layer on the CIFAR10 dataset. The red and blue bars depict the local gradient distribution on client 1 and client 2, respectively. In (a), the distribution of local gradients is demonstrated across the two clients in IID settings. In (b), the distribution is shown in non-IID settings, with the introduction of Gaussian noise with variance x (noise = x) on non-IID clients.

Algorithm 1 MW group-fairness in Federated Learning

Input: $(\mathbf{x}_i, y_i, g_j, c_k)$, global fairness learning rate η_μ, iteration count T, model class H.
Let $\epsilon_{g_j} \longleftarrow \frac{1}{|\mathbf{G}|} \sum_{(\mathbf{x},y)\in g_j} \mathscr{L}(h_\theta(\mathbf{x}), y)$; (for each c_k)
Initialize $\lambda_{g_j} \longleftarrow P(\mathbf{G} = g_j)$ and θ randomly.
for $t = 1$ `to` T **do**
  for each client $c_k \in \mathbf{C}$ **do**
    Compute $w_{g_j}^t \longleftarrow \frac{\lambda_{g_j}}{P(\mathbf{G}=g_j)}$
    Find $h_{c_k} \longleftarrow \arg\min_{h\in H} \sum_g w_{g_j}^t \cdot \epsilon_{g_j}(h_{c_k})$; (for $h_{c_k} \in H$)
    Update $\lambda_{g_j} \longleftarrow \lambda_{g_j} \cdot \exp(-\eta_\mu \cdot \epsilon_{g_j}(h_{c_k}))$; (Multiplicative Weight Update)
    Send h_{c_k} $(\boldsymbol{\theta}_{c_k})$ to the server.
Server **computes:** $\theta \longleftarrow \sum_{c_k\in C} \frac{n_{c_k}}{n} \boldsymbol{\theta}_t^c$; (FedAvg: n_{c_k}–number of data points at client c_k; n–total data points in FL)
Output: Uniform distribution over the set of models $h_1, ..., h_T$ with parameters $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_T$, respectively

4.1 Problem Statement and Workflow

Our configuration assumes a 4-tuple:

$$(\mathbf{x}_{1\le i\le|\mathbf{X}|}, y_{1\le i\le|\mathbf{Y}|}, g_{1\le j\le|\mathbf{G}|} c_{1\le k\le|\mathbf{C}|})$$

drawn from distribution $P(\mathbf{X}, \mathbf{Y}, \mathbf{G}, \mathbf{C})$. Here, $\mathbf{x}_i \in \mathbf{X}$ represents training images from a total of $|\mathbf{X}|$ images, $y_i \in \mathbf{Y}$ corresponds to $|\mathbf{Y}|$ targets, $g_j \in \mathbf{G}$ denotes group membership (from $|\mathbf{G}|$ groups) of $\mathbf{x}_i$, and c_k is the client on which $(\mathbf{x}_i, y_i)$ resides out of $|\mathbf{C}|$ clients. Our primary goal is to derive a global model $h_{\boldsymbol{\theta}}$ (with parameters $\boldsymbol{\theta}$) that mitigates *group bias* for each client, with following objective:

$$\mathscr{L}(h_{\boldsymbol{\theta}}) = \arg\min_h \frac{1}{|\mathbf{C}|} \sum_{c_i=1}^{|\mathbf{C}|} \mathscr{L}_{c_k}(h_{\boldsymbol{\theta}}(\mathbf{x}_{i,k}), y_{i,k}) \quad (1)$$

In equation 1 $\mathscr{L}c_k(h_{\boldsymbol{\theta}}(\mathbf{x}_{i,k}), y_{i,k})$, the empirical risk of client c_k combines group empirical risks $\ell_{g,j}(h(\mathbf{x}_{i,k}), y_{i,k})$ with *group importance* w_{g_j}. Importance is based on the ratio $\frac{q(g_j|\mathbf{x}_i)}{p(g_j|\mathbf{x}_i)}$, where $p(g_j|\mathbf{x}_i)$ and $q(g_j|\mathbf{x}_i)$ represent training and test distributions in $D = \cup D_{c_k}$ (global dataset as a union of local datasets D_{c_k}), respectively. We compute w_{g_j} as an aggregation of all per-client local group importance weights $w_{g_{j,k}}$, $\forall g_j \in \mathbf{G}, c_k \in \mathbf{C}$. Each w_{g_j} obtained from a multi-class logistic linear regression probabilistic model [27] is used to train a local model, $h_{\theta_{c_k}}$, minimizing the empirical risk of the *worst-performing* group. On the server side, $h_{\boldsymbol{\theta}_{c_k}}$ from all clients is received and aggregated into a global model $h_{\boldsymbol{\theta}}$.

Workflow. We illustrate the end-to-end workflow for training with the proposed approach in Figure 4.

❶ In our setup, the server selects all the available clients in each round to avoid the effect of client sampling bias [10, 70, 77]. Then, the server distributes copies of the global model to the clients.

❷-❹ Each client computes the *mixture of group likelihoods*, denoted as $p(g_j|\mathbf{x}_i)$ (specifically, $p(g_j|\mathbf{x}_{i,k})$). In § 4.2, we outline the privacy-preserving computation details of this denominator, occurring once at the beginning of FL. After each round, clients communicate the local model and local $p(g_j|\mathbf{x}_{i,k})$ for all groups (only in the first round) to the server.

❺ After clients submit their local models and local $p(g_j|\mathbf{x}_{i,k})$, the server uses FedAvg to aggregate the local models and generate an updated global model. Additionally, the server computes a *mixture of group likelihoods* for all groups using

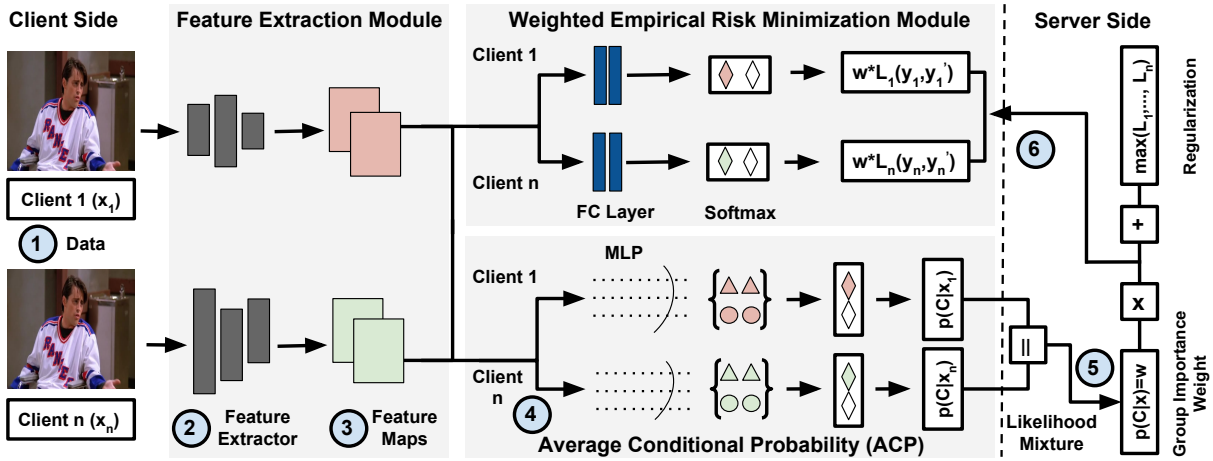


Figure 4. ***Overview of the proposed approach.***

local likelihoods (emphasizing that this computation occurs once at the beginning of FL).

❻ Each client performs local training after distributing updated global model copies and a *mixture of likelihoods* for all groups. The training involves using our approach *MWR* to adjust group importance weights based on the *mixture of likelihoods* for all groups (§ 4.3).

❼ Each client computes the performance threshold for the best group and compares it with the best group performance to evaluate *MWR*'s effectiveness in mitigating group bias without compromising the best group performance (§ 4.5).

4.2 Enabling Privacy-preserving Group Fairness

Our approach centers on weighting empirical risks with group importance weights, w_{g_j}, as shown in Equation 1. Calculating these weights is straightforward in centralized learning [20], where a global data view is available. However, In FL, lacking this global view is not trivial. We must estimate w_{g_j} while safeguarding client data privacy. Our solution addresses this by approximating the denominator of w_{g_j} ($p(\mathbf{G} = g_j|\mathbf{X})$) through a process involving a mixture of group likelihoods across clients. Suppose $\mathbf{G} = 1, ..., j$ represents groups across clients in FL. Each client c_k employs a multiclass logistic linear regression probabilistic model [2] to predict the likelihood of an input sample $\mathbf{x}_{i,k}$ belonging to a specific group g_j. The model is defined as $p(\mathbf{G} = g_j|\mathbf{X} = \mathbf{x}_{i,k}) = \prod_j^J f_{\theta,j}(\mathbf{x}_i)^{[g_j=j]}$, where $f_{\theta,j}(\mathbf{x}_{i,k})^{[g_j=j]}$ is a multinomial probability mass function [39]. Each client uses the softmax function $f_{\theta,j}(\mathbf{x}_{i,k})^{[g_j=j]} = \prod_j^J \frac{\exp(\mathbf{x}_{i,k}\theta_{c_k})}{\sum_j^J \exp(\mathbf{x}_{i,k}\theta_{c_k})}$ to obtain group membership probabilities ensuring that these probabilities are positive and sum up to one. Clients share their group likelihood estimates with the server. The server then computes each group's global average likelihood using per-client group average likelihood estimates and the *law of total probability*. For an event space $\{c_1, c_2, ..., c_{|\mathbf{C}|}\}$ with $P(c_k) \geq 0 \quad \forall k$,

$$p(\mathbf{G} = g_j|\mathbf{X}) = \sum_{j=1}^{|\mathbf{C}|} p(\mathbf{G} = g_j|c_k, \mathbf{x}_i)p(c_k). \tag{2}$$

Here $p(\mathbf{G} = g_j|c_k, \mathbf{x}_i)$ represents per-group likelihood estimates per client, and $p(c_k)$ is the likelihood of a client c_k. In our scenario, $p(c_k)$ is uniform for all clients participating in each training round. Utilizing the law of total probability due to independence in clients' participation in FL, the server distributes group likelihood mixtures $p(\mathbf{G} = g_j|\mathbf{X})$ to all clients. Clients use this information to compute group importance weights w_{g_j}, updated using *MWR* in each round based on $p(\mathbf{G} = g_j|\mathbf{X})$. To ensure data privacy, clients and the server share required information ($p(\mathbf{G} = g_j|c_k, \mathbf{x}_i)$) by revealing differentially private likelihood estimates.

To solve the group bias problem, we modify the *MW* algorithm and transform it into a constrained optimization problem to improve the performance of the the worst-performing group. Algorithm 1 details the workings of the *MW* algorithm. We assign each client with groups and a set of $|\mathbf{G}|$ classes for the underlying application during the local learning process. The optimization constraints comprise decisions made by both the local and global models for groups assigned to clients, ensuring fairness in group classification. Using image features in the training dataset, we validate constraint satisfaction in each local training iteration and identify suitable groups. We then associate decisions made by each local model with a group empirical risk that quantifies how well a decision made by the local model satisfies the constraints. Over time, we minimize the overall risk of the global model by ensuring that each local model incurs a low per-group risk. This involves tracking the global weight for each group and randomly selecting groups with a probability proportional to their importance weights w_{g_j}. In each iteration, we update w_{g_j} using the *MW* algorithm, multiplying their numerator $q(\mathbf{G} = g_j|\mathbf{G})$ with factors dependent on the risk

of the associated group decision. This update is performed while maintaining the denominator $p(\mathbf{G} = g_j|\mathbf{G})$ fixed as in $\frac{\lambda \cdot \exp(\eta \cdot \ell_{g_j}(h))}{q(\mathbf{G}=g_j|\mathbf{X})}$, which penalizes costly group decisions.

4.3 Ensuring Optimality through Regularization

The *MW* algorithm maximizes worst-group performance by scaling the empirical risk and deep neural network weights. However, the weight magnitude does not ensure optimal risk function convergence [6]. In our setup, model parameters θ are trained with cross-entropy loss and stochastic gradient descent (SGD) [5] optimization, converging toward the solution of the hard-margin support vector machine[1] in the direction $\frac{\theta_t}{||\theta_t||}$ [65]. Introducing weight to the loss function may introduce inconsistencies in the margin. Instead of directly applying importance weighting to the empirical risk, we aim to minimize the following objective for each client k: $\sum_{c_k=1}^{|\mathbf{C}|} \mathscr{L}_{c_k}(h_\theta(\mathbf{x}_{i,k}), y_{i,k}) + \frac{\lambda}{m} \sum_{j=1}^{m} \|\theta_{j,c_k}\|$.

Since the optimization problem with *importance weighting* is vulnerable to scaling weights and biases, we introduce regularization to the norm of θ_{c_k} to increase the margin and mitigate the risk of its enlargement due to scaling, forming the basis of our ***Multiplicative Weight update with Regularization (MWR)*** algorithm.

4.4 Bias Mitigation without Degrading High-Performing Groups

While *MWR* ensures group fairness, *importance weighting* approaches may exhibit unfairness by disproportionately focusing on the worst-performing groups, potentially degrading the performance of the best-performing groups in an attempt to reduce the variance in estimating their contributions to the overall performance [13]. Practically, an algorithm for bias mitigation should achieve fairness without significantly degrading the performance of best-performing groups. To address this, we propose a heuristic approach to reweighing the likelihood (*group importance weights*) associated with each data point belonging to group $\mathbf{G} = g_j$ in the dataset. Suppose we have a set of unnormalized importance weights $w_1, w_2, ..., w_n$ corresponding to n data points in a dataset, where each data point has an associated importance weight, we normalize these weights for each group by computing $\hat{w_1}, \hat{w_2}, ..., \hat{w_{|\mathbf{G}|}}$ using:

$$\hat{w_{g_j}} = \frac{\sum_{i=1}^{n} w_i \mathbb{I}(\mathbf{G} = g_j)}{\sum_{i=1}^{n} w_i} \tag{3}$$

The rationale behind Equation3 is to distribute emphasis evenly among different groups, preventing a scenario where a single group dominates the estimation due to an excessively high importance weight. Through weight normalization, we ensure that each group's contribution aligns more closely with its true importance or representation within the dataset.

[1]A linear classification algorithm that seeks a hyperplane with a strict margin, allowing no misclassification in the training data. [63]

4.5 Satisfying Performance Thresholds

Finally, we establish a performance threshold for the best true positive rate (BTPR) to mitigate group bias without significantly compromising the BTPR . We denote BTPR for a client c_i as TPR_{best,c_i} and WTPR as TPR_{worst,c_i}. We define the threshold for the best TPR as $TPR_{threshold}$. Our fairness enforcement objective aims to minimize the gap between the best and worst-performing groups while maintaining a specified level of TPR performance, as follows:

$$TPR_{best,c_i} - TPR_{threshold} \leq \eta_\mu \times (TPR_{best,c_i} - TPR_{worst,c_i}) \tag{4}$$

$$TPR_{threshold} \geq TPR_{best,c_i} - \eta_\mu \times (TPR_{best,c_i} - TPR_{worst,c_i}) \tag{5}$$

Here η_μ is a parameter governing the trade-off between group fairness and performance. Inequality in 4 scales the difference between BTPR and WTPR by η_μ and compares it to the difference between the BTPR and the threshold. For each client, we rearrange the inequality in 4 to obtain the minimum BTPR threshold as expressed in equation 5.

5 Evaluation

This section evaluates our *MWR* group-bias mitigation technique on four image classification datasets (CIFAR10, DIGITS, MNIST, and FashionMNIST). We benchmark our approach against standard bias mitigation techniques in FL.

5.1 Experiment Testbed

Our evaluation setup uses the same number of clients, data partitioning scheme, and other learning components (such as learning rate, train/test split, batch size, epochs, rounds) described in §3.1 unless stated otherwise.

Baseline. We evaluate our approach across four key categories, scrutinizing both bias reduction and overall model performance. The *FL baseline category* (FedAvg) represents a conventional learning scheme in FL. In the *FL bias-reduction category*, we include methods such as AFL[52], TERM[43], and GIFAIR-FL [73]. These methods employ empirical risk reweighting to mitigate bias and adapt the global model to diverse local data distributions. The *FL heterogeneity category* (FedNTD [41]) specifically addresses performance loss in FL models arising from data heterogeneity by managing global model memory loss. In the *FL robustness category* (SCAFFOLD [34]), the focus is on enhancing the resilience of FL models against outliers and noisy data, thereby mitigating the impact of irregularities in specific device local datasets. To ensure a fair evaluation across all baselines, we meticulously calibrate hyperparameters across datasets, guaranteeing the convergence of the global model.

Hyperparameter Tuning for MWR. We use the same experimental setup as FedAvg, AFL, FedNTD, TERM, GIFAIR-FL, and SCAFFOLD. However, to apply *MWR* update algorithm per-group loss, we set the value of η_μ (see Algorithm 1) to different values in the set $\{0.01, 0.02, 0.001, 0.009, 0.0001\}$ based on the level of Gaussian noise in data partitions. Finally,

MWR uses an $L1$ regularization parameter of 0.00001 for all datasets.

5.2 Efficacy and Robustness Analysis

We now assess the efficacy and robustness of our *MWR* group-bias mitigation technique with the baselines.

5.2.1 Effect on Group Bias. We assess the efficacy of *MWR*'s group-bias mitigation through: (i) evaluating the *best- and worst-group performance (TPR)*, (ii) analyzing the *TPR group variance per client*, and (iii) examining the *TPR discrepancy per client*. This evaluation is conducted on four datasets, incorporating *low-grade distortion* to simulate prevalent real-world heterogeneity [30].

Table1 presents the TPR, TPRSD, WTPR, and BTPR perfromance scores across various bias mitigation techniques and datasets. Notably, among these techniques, *MWR* stands out by achieving a significantly fairer outcomes for groups. We can see that our algorithm substantially decreases TPRSD across most clients while maintaining a consistently high TPR. Importance weighting, especially when derived from features characteristics, is powerful in mitigating biases caused by feature noise. If the bias is primarily driven by certain features, assigning appropriate weights to these features can help the model focus on relevant information and reduce the impact of noisy features, resulting in more consistent and equitable predictions.

Although AFL and FedNTD occasionally outperform *MWR* in some instances concerning the TPRSD metric as can be seen in DIGITS dataset's client4 and MNIST dataset's clients4 and 5, the differences between the results are marginal. Importance weighting is sensitive to distribution shifts in the feature space. If there are instances where the distribution shifts significantly, the importance weights may not be as effective. On the other hand, techniques such as FedNTD, through knowledge distillation, seem to be more robust to feature noise as it involves transferring knowledge from a more complex model (teacher) to a simpler one (student), potentially leading to better generalization and lower standard deviation in true positive rates across groups. Additionally, it becomes evident from Table 1 that *MWR* results in an increased WTPR for the group with the smallest TPR, accompanied by the smallest TPRD among the evaluated bias mitigation techniques.

Importance weights derived from image features captures the distinctive characteristics of different groups more effectively than other methods. This adaptability is crucial in mitigating bias since it tailors the mitigation strategy to the specific features and challenges present in each group. Despite TERM appearing to outperform our proposed method for the minimax group fairness metric (WTPR) in CIFAR10 dataset's clients 1, 2, and 3, this can be understood as a consequence of the reduction in TPR among privileged clients lacking local data with distortions. This reduction elevates the lower TPR among disadvantaged clients affected by distortions Importantly, the differences between the results are marginal, indicating a closely competitive performance between the methods despite this disparity while elevating the *group-fairness* among clients.

Takeaway: *MWR ensures fairness across groups and maintains predictive accuracy by using importance weights that prioritize the worst-performing groups. Its key strength lies in maintaining fairness without sacrificing performance, achieved through even distribution of importance weights among different groups.*

5.2.2 Robustness of Bias Mitigation. In our previous analysis, we added low-grade Gaussian noise to mimic noise in edge device images [47]. To further test *MWR*'s resilience against increased feature heterogeneity, we raised noise levels in segmented datasets like CIFAR10, MNIST, DIGITS, and Fashion-MNIST to variances of 0.11, 1.10, 1.00, and 0.4, respectively. Model performance evaluation used the same fairness metrics as before. Table 2 displays TPR, TPRSD, WTPR, and BTPR scores across various bias mitigation techniques and datasets, exploring high-grade distortion scenarios in local data. Consistent with our earlier findings, *MWR* delivers significantly fairer outcomes across diverse groups. The table shows *MWR* reduces TPRSD across most devices while maintaining high TPR. Compared with Table 1, *MWR* increases WTPR for the lowest TPR group, resulting in minimal TPRD among bias mitigation techniques. This enhancement in WTPR for disadvantaged groups minimally affects high-performing groups' performance.

Although some bias mitigation techniques may slightly outperform in TPRSD and WTPR fairness metrics, this often occurs at the expense of decreased TPR in privileged clients not affected by distortions. However, this decrease compensates for an increase in lower TPR among disadvantaged clients. Despite these differences, the results remain closely competitive among methods, indicating similar performance despite disparity, while simultaneously improving group fairness among clients.

Takeaway. *our robustness analysis suggests that* MWR *stands out as a robust and fair approach even in scenarios with high-grade heterogeneity, showcasing its effectiveness in mitigating bias across diverse datasets and client groups.*

5.3 Privacy Analysis

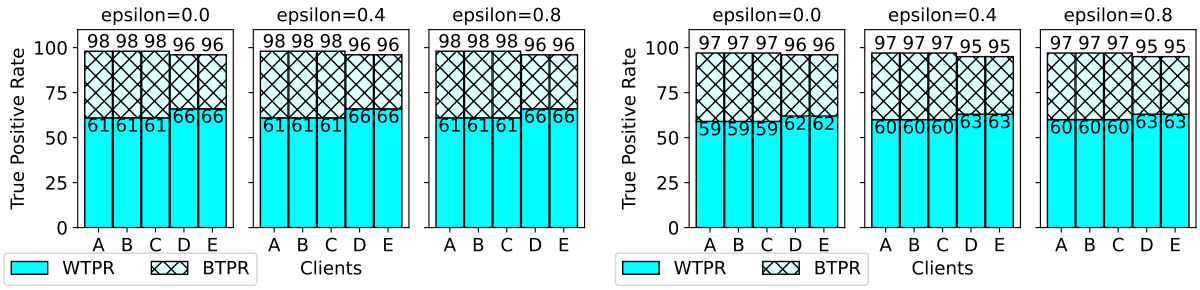
This section explores how differential privacy affects group fairness and performance in *MWR*, particularly in scenarios where local group probability distributions $p(\mathbf{G} = g_i|\mathbf{x}_{i,k})$ are shared with the server to compute *importance weights*. Differential privacy is crucial for preserving privacy in client metadata, preventing disclosure of sensitive details like group selection probabilities.

Algorithms	Datasets																			
	CIFAR10						DIGITS				Fashion-MNIST					MNIST				
	Client #	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
FedAvg [45]	TPRD ↓	28	28	28	40	40	33	28	39	26	48	48	48	55	55	2	2	2	18	18
	TPRSD ↓	9.13	9.13	9.13	13.29	13.29	9.01	9.91	13.53	6.1	14.19	14.19	14.19	16.1	16.1	0.6	0.6	0.6	5.29	5.29
	WTPR ↑	64	64	64	54	54	63	66	53	69	47	47	47	40	40	98	98	98	74	74
	BTPR ↑	92	92	92	94	94	96	94	92	95	95	95	95	95	95	100	100	100	92	92
AFL [52]	TPRD ↓	29	29	29	36	36	36	33	37	25	48	48	48	55	55	2	2	2	18	18
	TPRSD ↓	8.82	8.82	8.82	10.3	10.3	9.91	12.04	12.74	**5.18**⊙	14.19	14.19	14.19	16.05	16.05	0.6	0.6	0.6	5.03	5.03
	WTPR ↑	62	62	62	56	56	59	60	55	70	47	47	47	40	40	98	98	98	75	75
	BTPR ↑	91	91	91	92	92	95	93	92	95	95	95	95	95	95	100	100	100	93	93
FedNTD [41]	TPRD ↓	26	26	26	36	36	27	28	**33** ∗	28	46	46	46	50	50	2	2	2	17	17
	TPRSD ↓	8.38	8.38	8.38	11.51	11.51	8.02	7.88	12.54	7.71	13.83	13.83	13.83	14.96	14.96	0.6	0.6	0.6	**4.24** ⊙	**4.24** ⊙
	WTPR ↑	66	66	66	57	57	66	65	56	64	49	49	49	45	45	97	97	97	76	76
	BTPR ↑	92	92	92	93	93	93	93	89	92	95	95	95	95	95	99	99	99	93	93
TERM [43]	TPRD ↓	26	26	26	34	34	34	33	36	24	48	48	48	55	55	2	2	2	19	19
	TPRSD ↓	8.02	8.02	8.02	11.32	11.32	9.41	11.32	13.03	5.9	14.06	14.06	14.06	16.11	16.11	0.6	0.6	0.6	5.41	5.41
	WTPR ↑	**69**•	**69**•	**69**•	61	61	61	61	54	70	47	47	47	40	40	98	98	98	74	74
	BTPR ↑	95	95	95	95	95	95	94	90	94	95	95	95	95	95	100	100	100	93	93
GIFAIR-FL [73]	TPRD ↓	**24** ∗	**24**∗	**24**∗	36	36	26	30	48	39	43	43	43	50	50	2	2	2	16	16
	TPRSD ↓	8.47	8.47	8.47	11.17	11.17	7.82	8.55	15.63	13.16	12.82	12.82	12.82	14.48	14.48	0.6	0.6	0.6	5.47	5.47
	WTPR ↑	68	68	68	56	56	68	64	44	52	53	53	53	46	46	98	98	98	76	76
	BTPR ↑	92	92	92	92	92	94	94	92	91	96	96	96	96	96	100	100	100	92	92
SCAFFOLD [34]	TPRD ↓	29	29	29	65	65	60	64	84	73	50	50	50	60	60	2	2	2	25	25
	TPRSD ↓	10.19	10.19	10.19	20.42	20.42	18.02	20.94	26.35	24.64	14.63	14.63	14.63	17.24	17.24	1.36	1.36	1.36	6.57	6.57
	WTPR ↑	63	63	63	32	32	37	28	12	20	46	46	46	35	35	97	97	97	70	70
	BTPR ↑	92	92	92	97	97	97	92	96	93	96	96	96	95	95	99	99	99	95	95
MWR	TPRD ↓	25	25	25	**30** ∗	**30** ∗	**21** ∗	**19** ∗	39	**23** ∗	**37** ∗	**37** ∗	**37** ∗	**30** ∗	**30** ∗	**1** ∗	**1** ∗	**1** ∗	**13** ∗	**13** ∗
	TPRSD ↓	**7.94**⊙	**7.94**⊙	**7.94**⊙	**10.05**⊙	**10.05**⊙	**5.79**⊙	**5.86**⊙	**11.79**⊙	5.9	**11.02**⊙	**11.02**⊙	**11.02**⊙	**11.17**⊙	**11.17**⊙	**0.4**⊙	**0.4**⊙	**0.4**⊙	4.83	4.83
	WTPR ↑	68	68	68	**63**•	**63**•	**77**•	**77**•	**58**•	**73**•	**61**•	**61**•	**61**•	**66**•	**66**•	**99**•	**99**•	**99**•	**80**•	**80**•
	BTPR-threshold	92.5	92.5	92.5	92.4	92.4	97.8	95.8	96.6	95.7	97.6	97.6	97.6	95.7	95.7	99.9	99.9	99.9	92.9	92.9
	BTPR ↑	**93**▷	**93**▷	**93**▷	**93**▷	**93**▷	**98**▷	**96**▷	**97**▷	**96**▷	**98**▷	**98**▷	**98**▷	**96**▷	**96**▷	**100**▷	**100**▷	**100**▷	**93**▷	**93**▷

Table 1. Performance evaluation of bias mitigation techniques across various datasets and benchmark models under low-grade noise. Symbols used: ↑indicates that higher values are more desirable, while ↓ indicates that lower values are more desirable. For each client across each benchmarks in a particular dataset ∗ signifies the best TPRD; ⊙ designates the best TPRSD; • represents the best WTPR; and ▷ indicates the best BTPR . (Note: On DIGITS dataset,training involves only 4 clients, reflecting its composition of merely 4 heterogeneous datasets.)

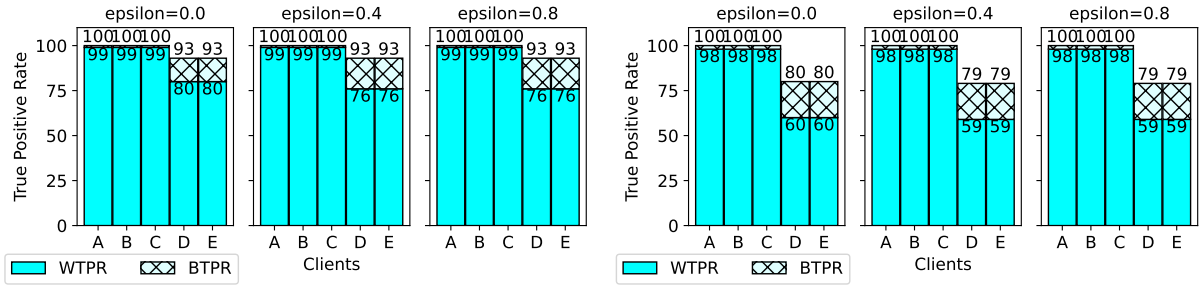
Algorithms	Datasets																			
	CIFAR10						DIGITS				Fashion-MNIST					MNIST				
	Client #	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
FedAvg [45]	TPRD ↓	31	31	31	46	46	46	39	42	29	48	48	48	59	59	2	2	2	29	29
	TPRSD ↓	9.91	9.91	9.91	14.7	14.7	13.39	12.48	13.85	6.87	14.39	14.39	14.39	17.05	17.05	0.78	0.78	0.78	8.45	8.45
	WTPR ↑	61	61	61	48	48	46	56	50	**66**•	46	46	46	35	35	97	97	97	51	51
	BTPR ↑	92	92	92	94	94	92	95	92	95	94	94	94	94	94	99	99	99	80	80
AFL [52]	TPRD ↓	33	33	33	44	44	47	43	40	29	49	49	49	59	59	2	2	2	28	28
	TPRSD ↓	9.36	9.36	9.36	14.13	14.13	13.6	14.01	**13.35** ⊙	6.9	14.65	14.65	14.65	17.11	17.11	0.7	0.7	0.7	7.63	7.63
	WTPR ↑	61	61	61	49	49	43	49	52	66	45	45	45	35	35	97	97	97	52	52
	BTPR ↑	94	94	94	93	93	90	92	92	95	94	94	94	94	94	99	99	99	80	80
FedNTD [41]	TPRD ↓	26	26	26	56	56	35	29	**37**	**27**	46	46	46	50	50	2	2	2	25	25
	TPRSD ↓	8.91	8.91	8.91	16.69	16.69	10.65	9.03	13.64	8.76	13.83	13.83	13.83	15.01	15.01	0.74	0.74	0.74	6.77	6.77
	WTPR ↑	65	65	65	40	40	54	59	51	64	49	49	49	45	45	97	97	97	56	56
	BTPR ↑	91	91	91	96	96	89	88	88	91	95	95	95	95	95	99	99	99	81	81
TERM [43]	TPRD ↓	**23** ∗	**23** ∗	**23** ∗	40	40	47	40	43	30	48	48	48	59	59	2	2	2	30	30
	TPRSD ↓	7.9	7.9	7.9	13.41	13.41	13.72	13.07	14.13	**5.87** ⊙	14.39	14.39	14.39	17.08	17.08	0.78	0.78	0.78	8.64	8.64
	WTPR ↑	**69**•	**69**•	**69**•	53	53	44	55	49	65	46	46	46	35	35	97	97	97	51	51
	BTPR ↑	92	92	92	93	93	91	95	92	95	94	94	94	94	94	99	99	99	81	81
GIFAIR-FL [73]	TPRD ↓	30	30	30	53	53	32	37	48	40	45	45	45	53	53	2	2	2	27	27
	TPRSD ↓	**8.16** ⊙	**8.16** ⊙	**8.16** ⊙	14.92	14.92	10.13	10.18	15.69	13.06	13.4	13.4	13.4	15.46	15.46	0.66	0.66	0.66	7.64	7.64
	WTPR ↑	63	63	63	42	42	56	56	43	51	51	51	51	42	42	98	98	98	54	54
	BTPR ↑	93	93	93	95	95	88	93	91	91	96	96	96	95	95	100	100	100	81	81
SCAFFOLD [34]	TPRD ↓	38	38	38	94	94	47	60	84	74	51	51	51	63	63	5	5	5	54	54
	TPRSD ↓	13.21	13.21	13.21	26.53	26.53	14.73	18.13	27.46	23.65	14.77	14.77	14.77	18.27	18.27	1.32	1.32	1.32	14.06	14.06
	WTPR ↑	57	57	57	5	5	48	35	10	22	45	45	45	31	31	95	95	95	29	29
	BTPR ↑	95	95	95	99	99	95	95	94	96	96	96	96	94	94	100	100	100	83	83
MWR	TPRD ↓	29	29	29	**33** ∗	**33** ∗	**28** ∗	**24** ∗	44	29	**38** ∗	**38** ∗	**38** ∗	**34** ∗	**34** ∗	**2** ∗	**2** ∗	**2** ∗	**20** ∗	**20** ∗
	TPRSD ↓	10.29	10.29	10.29	**12.27** ⊙	**12.27**⊙	**8.01**⊙	**8.09**⊙	13.76	7.98	**11.35** ⊙	**11.35**⊙	**11.35**⊙	**12.44**⊙	**12.44**⊙	**0.63**⊙	**0.63**⊙	**0.63**⊙	**6.45**⊙	**6.45**⊙
	WTPR ↑	66	66	66	**58**•	**58**•	**68**•	**69**•	**51**•	65	**59**•	**59**•	**59**•	**62**•	**62**	**98**•	**98**•	**98**•	**60**•	**90**•
	BTPR-threshold	94.7	94.7	94.7	90.6	90.6	95.7	92.7	94.6	93.7	96.6	96.6	96.6	95.6	95.6	99.9	99.9	99.9	79.8	78.8
	BTPR ↑	**95**▷	**95**▷	**95**▷	**91**▷	**91**▷	**96**▷	**93**▷	**95**▷	**94**▷	**97**▷	**97**▷	**97**▷	**96**▷	**96**▷	**100**▷	**100**▷	**100**▷	**80** ▷	**80**▷

Table 2. Performance evaluation of bias mitigation techniques across various datasets and benchmark models under low-grade noise. Symbols used: ↑indicates that higher values are more desirable, while ↓ indicates that lower values are more desirable. For each client across each benchmarks in a particular dataset ∗ signifies the best TPRD; ⊙ designates the best TPRSD; • represents the best WTPR; and ▷ indicates the best BTPR . (Note: On DIGITS dataset,training involves only 4 clients, reflecting its composition of merely 4 heterogeneous datasets.)



(a) FasionMNIST with DP in MWR ($var = 0.3$) **(b)** FasionMNIST with DP in MWR ($var = 0.4$)

Figure 5. Examining the performance trade-off in *MWR* concerning privacy and accuracy across various levels of differential privacy (DP) noise factors on FashionMNIST. In (a), a base Gaussian noise with a variance of 0.3 is introduced to all methods, while in (b), Gaussian noise with a variance of 0.4 is applied to all methods.



(a) MNIST with DP in MWR ($var = 0.8$) **(b)** MNIST with DP in MWR ($var = 1.1$)

Figure 6. Examining the performance trade-off in *MWR* concerning privacy and accuracy across various levels of differential privacy (DP) noise factors on MNIST. In (a), a base Gaussian noise with a variance of 0.8 is introduced to all methods, while in (b), Gaussian noise with a variance of 1.1 is applied to all methods.

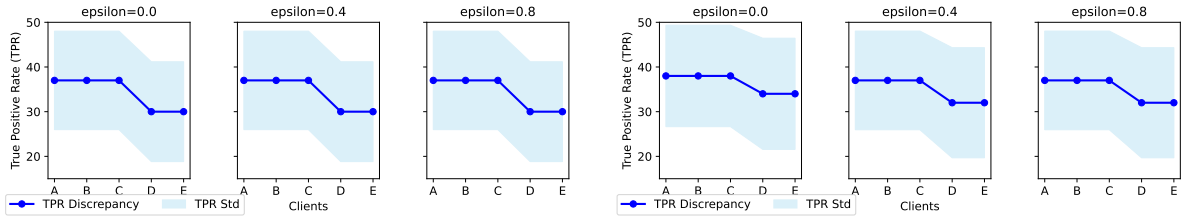
We use the MNIST and FashionMNIST datasets for our privacy budget analysis, maintaining consistency in experimental setups and various learning components as detailed in §3.1. We introduce different levels of Laplace noise, denoted by ϵ, to local probability distributions. An ϵ value of 0.00 represents perfect differential privacy in the implementation of *MWR*.

Figures 5 to 8 show the impact of varying levels of Laplace noise (ϵ) on group-fairness metrics (WTPR, TPRSD, and TPRD) and group performance (TPR) in MWR, addressing bias in local data with different levels of feature noise. In Figures 5a to 7b, we see that using a privacy budget ($\epsilon \in 0.0, 0.4, 0.8$) for metadata exchange maintains fairness metrics similar to deploying MWR without privacy ($\epsilon \longrightarrow \infty$) on MNIST and FashionMNIST. This is evident from minimal variations in WTPR, TPRSD, and TPRD across all clients (with high and low feature heterogeneity) under all privacy budgets. Moreover, the privacy budget ensures fairness while preserving the best and worst TPR performance. This aligns with the fairness guarantee of MWR, as the privacy budget values ($\epsilon \in 0.0, 0.4, 0.8$) fall within a range that provides algorithmic fairness, as noted in [1]. Our privacy analysis underscores that our method ensures client privacy through differential privacy on shared metadata without significantly affecting bias or accuracy.

Takeaway. *MWR demonstrates the feasibility of preserving sensitive information while effectively reducing group bias.*

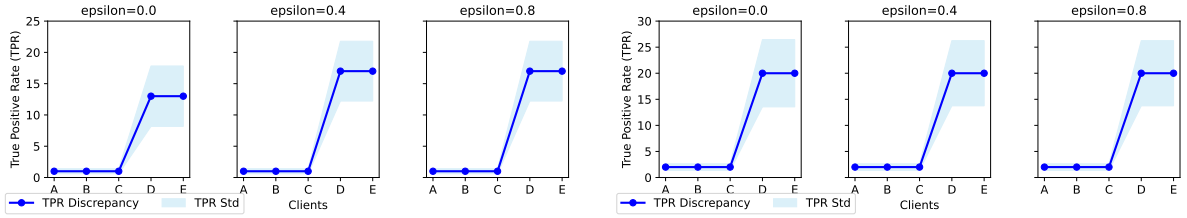
5.4 Fairness Budget Analysis

MWR incorporates a fairness budget, denoted as η_μ, to regulate importance weight adjustments for fairness. This control mechanism in *MWR* adjusts importance weights based on past group performance (group loss) for fairness metrics. We assess the impact of η_μ on group fairness metrics (WTPR, TPRSD, TPRD) using MNIST and FashionMNIST datasets, setting η_μ to different values ($-0.009, -0.003, -0.001, -0.0002$). Tables 3 and 4 show how the fairness budget η_μ affects both group fairness and group performance (TPR) with *MWR*. Increasing η_μ values improve fairness guarantees, leading to better WTPR, TPRSD, and TPRD due to faster convergence and adaptation to fairness issues. Conversely, lower η_μ values result in more gradual adjustments, slowing down the algorithm's fairness improvements. This experiment is



(a) FasionMNIST with DP in MWR ($var = 0.3$) **(b)** FasionMNIST with DP in MWR ($var = 0.3$)

Figure 7. Analyzing the privacy-bias trade-off in *MWR* across differential privacy (DP) noise levels on FashionMNIST. (a) introduces a base Gaussian noise with a variance of 0.3, and in (b), Gaussian noise with a variance of 0.4 is applied. Shaded areas represent deviation represented by TPRSD.



(a) MNIST with DP in MWR ($var = 0.8$) **(b)** MNIST with DP in MWR ($var = 1.1$)

Figure 8. Analyzing the privacy-bias trade-off in *MWR* across differential privacy (DP) noise levels on MNIST. (a) introduces a base Gaussian noise with a variance of 0.8, and in (b), Gaussian noise with a variance of 1.1 is applied. Shaded areas represent deviation represented by TPRSD.

crucial for understanding how adjusting fairness settings impacts outcomes, helping us strike a balance between fairness and the specific fairness parameter we use.

Takeaway. *Fine-tuning the fairness budget in MWR significantly shapes the degree of fairness. Higher values amplify fairness, while lower values diminish it, underscoring the pivotal role of this parameter in mitigating group bias.*

6 Conclusion and Future Work

This study explores FL group bias in decentralized, heterogeneous edge deployments, where devices capture data with diverse features often influenced by noise. Our framework, *MWR*, uses *importance weighting* and *average conditional probabilities* based on data features to improve group fairness in FL across varied local datasets. Heterogeneous features in local group data can bias FL models for minority clients, impacting specific groups on those clients. *MWR* addresses this bias by optimizing worst-performing groups without compromising the best-performing ones compared to other FL methods. While effective, *MWR* relies on group information to mitigate bias across clients, which can lead to persistent loss discrepancies under severe feature heterogeneity. Future work aims to incorporate methods for estimating and denoising data features to reduce noise without compromising data quality. *MWR* is highly adaptable and can be extended to complex applications beyond image classification. It can optimize diagnostic outcomes in healthcare datasets, handle multimodal and text-based applications like next-character prediction and image captioning, and mitigate bias in emotion prediction applications within FL settings, ensuring equitable outcomes across diverse groups.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant number 2237485.

References

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

[2] Felix Abramovich, Vadim Grinshtein, and Tomer Levy. 2021. Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory* 67, 7 (2021), 4637–4646.

[3] Sanjeev Arora, Elad Hazan, and Satyen Kale. 2012. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing* 8, 1 (2012), 121–164.

[4] Alejandro Baldominos, Yago Saez, and Pedro Isasi. 2019. A survey of handwritten character recognition with mnist and emnist. *Applied Sciences* 9, 15 (2019), 3169.

[5] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Springer, 177–186.

[6] Jonathon Byrd and Zachary Lipton. 2019. What is the effect of importance weighting in deep learning?. In *International conference on machine learning*. PMLR, 872–881.

[7] Canyu Chen, Yueqing Liang, Xiongxiao Xu, Shangyu Xie, Yuan Hong, and Kai Shu. 2022. On Fair Classification with Mostly Private Sensitive

	noise variance = 0.3										**noise variance = 0.4**									
	$\eta_\mu = -0.003$					$\eta_\mu = -0.009$					$\eta_\mu = -0.003$					$\eta_\mu = -0.009$				
Client #	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
TPRD ↓	37	37	37	**29**	**29**	37	37	37	30	30	38	38	38	**32**	**32**	38	38	38	34	34
TPRSD ↓	**10.92**	**10.92**	**10.92**	**11.11**	**11.11**	11.02	11.02	11.02	11.17	11.17	**11.21**	**11.21**	**11.21**	**12.05**	**12.05**	11.35	11.35	11.35	12.44	12.44
WTPR ↑	**61**	**61**	**61**	**68**	**68**	61	61	61	66	66	59	59	59	**63**	**63**	59	59	59	62	62
BTPR ↑	98	98	98	97	97	98	98	98	96	96	97	97	97	95	95	97	97	97	96	96

Table 3. Impact of the fairness budget η_μ on Fashion-MNIST. A base Gaussian noise with a variance of 0.3, 0.4 is introduced to MWR in (a) and (b), respectively. ↑: Higher is best, ↓: Lower is best.

	noise variance = 0.8										**noise variance = 1.1**									
	$\eta_\mu = -0.002$					$\eta_\mu = -0.001$					$\eta_\mu = -0.002$					$\eta_\mu = -0.001$				
Client #	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
TPRD ↓	1	1	1	13	13	1	1	1	13	13	2	2	2	20	20	2	2	2	20	20
TPRSD ↓	**0.3**	**0.3**	**0.3**	**4.33**	**4.33**	0.4	0.4	0.4	4.83	4.83	**0.53**	**0.53**	**0.53**	**6.39**	**6.39**	0.63	0.63	0.63	6.45	6.45
WTPR ↑	99	99	99	80	80	99	99	99	80	80	98	98	98	60	60	98	98	98	60	60
BTPR ↑	100	100	100	93	93	100	100	100	93	93	100	100	100	80	80	100	100	100	80	80

Table 4. Impact of the fairness budget η_μ on the TPR, TPRD, and WTPR on MNIST. A base Gaussian noise with a variance of 0.8, 1.1 is introduced to MWR in (a) and (b), respectively. ↑: Higher is best, ↓: Lower is best.

Attributes. *arXiv preprint arXiv:2207.08336* (2022).

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

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

[10] Gregory Francis Coppola. 2015. Iterative parameter mixing for distributed large-margin training of structured predictors for natural language processing. (2015).

[11] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 26091–26102.

[12] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. 2020. Distributionally robust federated averaging. *Advances in neural information processing systems* 33 (2020), 15111–15122.

[13] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. 2021. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 66–76.

[14] Sidney D'Mello, Rosalind W. Picard, and Arthur Graesser. 2007. Toward an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems* 22, 4 (2007), 53–61. https://doi.org/10.1109/MIS.2007.79

[15] Wei Du and Xintao Wu. 2021. Robust fairness-aware learning under sample selection bias. *arXiv preprint arXiv:2105.11570* (2021).

[16] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. 2021. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 181–189.

[17] Sannara Ek, François Portet, Philippe Lalanda, and German Vega. 2020. Evaluation of federated learning aggregation algorithms: application to human activity recognition. In *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers*. 638–643.

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

[19] Zhenan Fan, Huang Fang, Zirui Zhou, Jian Pei, Michael P Friedlander, Changxin Liu, and Yong Zhang. 2022. Improving fairness for data valuation in horizontal federated learning. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 2440–2453.

[20] Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. 2021. Learning bounds for open-set learning. In *International conference on machine learning*. PMLR, 3122–3132.

[21] Ines Feki, Sourour Ammar, Yousri Kessentini, and Khan Muhammad. 2021. Federated learning for COVID-19 screening from Chest X-ray images. *Applied Soft Computing* 106 (2021), 107330.

[22] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[23] Borja Rodríguez Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. 2021. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS 2021 Workshop Privacy in Machine Learning*.

[24] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[25] Sanjukta Ghosh, Rohan Shet, Peter Amon, Andreas Hutter, and André Kaup. 2018. Robustness of deep convolutional neural networks for image degradations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2916–2920.

[26] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[27] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[29] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[30] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. 2020. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*. PMLR, 4387–4398.

[31] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. 2020. Fedmgda+: Federated learning meets multi-objective optimization. *arXiv preprint arXiv:2006.11489* (2020).

[32] Jonathan J. Hull. 1994. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence* 16, 5 (1994), 550–554.

[33] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 35–50.

[34] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*. PMLR, 5132–5143.

[35] Fereshte Khani and Percy Liang. 2020. Feature noise induces loss discrepancy across groups. In *International Conference on Machine Learning*. PMLR, 5209–5219.

[36] Nancy Krieger. 2012. Methods for the scientific study of discrimination and health: an ecosocial approach. *American journal of public health* 102, 5 (2012), 936–944.

[37] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[38] Hanchao Ku, Willy Susilo, Yudi Zhang, Wenfen Liu, and Mingwu Zhang. 2022. Privacy-Preserving federated learning in medical diagnosis with homomorphic re-Encryption. *Computer Standards & Interfaces* 80 (2022), 103583.

[39] Sadan Kulturel-Konak, Alice E Smith, and Bryan A Norman. 2006. Multi-objective tabu search using a multinomial probability mass function. *European Journal of Operational Research* 169, 3 (2006), 918–931.

[40] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[41] Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. 2022. Preservation of the global knowledge by not-true distillation in federated learning. *Advances in Neural Information Processing Systems* 35 (2022), 38461–38474.

[42] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10143–10152.

[43] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. 2020. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162* (2020).

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

[45] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189* (2019).

[46] Christine Lisetti, Fatma Nasoz, Cynthia Lerouge, Onur Ozyer, and Kaye Alvarez. 2003. Developing multimodal intelligent affective interfaces for tele-home health care. *Int. J. Hum.-Comput. Stud.* 59 (07 2003), 245–255. https://doi.org/10.1016/S1071-5819(03)00051-X

[47] Zida Liu, Guohao Lan, Jovan Stojkovic, Yunfan Zhang, Carlee Joe-Wong, and Maria Gorlatova. 2020. CollabAR: Edge-assisted collaborative image recognition for mobile augmented reality. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 301–312.

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

[49] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*. 349–358.

[50] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*. PMLR, 6755–6764.

[51] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

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

[53] Leann Myers and Maria J Sirois. 2004. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences* 12 (2004).

[54] Lokesh Nagalapatti and Ramasuri Narayanam. 2021. Game of gradients: Mitigating irrelevant clients in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9046–9054.

[55] Ismoilov Nusrat and Sung-Bong Jang. 2018. A comparison of regularization techniques in deep neural networks. *Symmetry* 10, 11 (2018), 648.

[56] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. 2021. Clusterfl: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 54–66.

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

[58] Raphael Poulain, Mirza Farhan Bin Tarek, and Rahmatollah Beheshti. 2023. Improving Fairness in AI Models on Electronic Health Records: The Case for Federated Learning Methods. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1599–1608.

[59] Adnan Qayyum, Kashif Ahmad, Muhammad Ahtazaz Ahsan, Ala Al-Fuqaha, and Junaid Qadir. 2022. Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. *IEEE Open Journal of the Computer Society* 3 (2022), 172–184.

[60] Amirhossein Reisizadeh, Farzan Farnia, Ramtin Pedarsani, and Ali Jadbabaie. 2020. Robust federated learning: The case of affine distribution shifts. *Advances in Neural Information Processing Systems* 33 (2020), 21554–21565.

[61] Jae Ro, Mingqing Chen, Rajiv Mathews, Mehryar Mohri, and Ananda Theertha Suresh. 2021. Communication-efficient agnostic federated averaging. *arXiv preprint arXiv:2104.02748* (2021).

[62] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer, 213–226.

[63] Houssem Sifaou, Abla Kammoun, and Mohamed-Slim Alouini. 2019. Phase transition in the hard-margin support vector machines. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 415–419.

[64] Congzheng Song, Filip Granqvist, and Kunal Talwar. 2022. FLAIR: Federated Learning Annotated Image Repository. *arXiv preprint arXiv:2207.08869* (2022).

[65] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2018. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research* 19, 1 (2018), 2822–2878.

[66] Konstantin Sozinov, Vladimir Vlassov, and Sarunas Girdzijauskas. 2018. Human activity recognition using federated learning. In *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social*

Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom). IEEE, 1103–1111.

[67] Jingwei Sun, Ang Li, Lin Duan, Samiul Alam, Xuliang Deng, Xin Guo, Haiming Wang, Maria Gorlatova, Mi Zhang, Hai Li, et al. 2022. FedSEA: A Semi-Asynchronous Federated Learning Framework for Extremely Heterogeneous Devices. (2022).

[68] Michael Sun, Tomasz Oliwa, Monica E Peek, and Elizabeth L Tung. 2022. Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record: Study examines racial bias in the patient descriptors used in the electronic health record. *Health Affairs* 41, 2 (2022), 203–211.

[69] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2021. Modeling techniques for machine learning fairness: A survey. *arXiv preprint arXiv:2111.03015* (2021).

[70] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K Leung, Christian Makaya, Ting He, and Kevin Chan. 2019. Adaptive federated learning in resource constrained edge computing systems. *IEEE journal on selected areas in communications* 37, 6 (2019), 1205–1221.

[71] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[72] Georgios N. Yannakakis and Julian Togelius. 2011. Experience-Driven Procedural Content Generation. *IEEE Transactions on Affective Computing* 2, 3 (2011), 147–161. https://doi.org/10.1109/T-AFFC.2011.6

[73] Xubo Yue, Maher Nouiehed, and RA Kontar. 2021. Gifair-fl: An approach for group and individual fairness in federated learning. *arXiv preprint arXiv:2108.02741* (2021).

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

[75] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. 2020. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 1051–1060.

[76] Fengda Zhang, Kun Kuang, Yuxuan Liu, Long Chen, Chao Wu, Fei Wu, Jiaxun Lu, Yunfeng Shao, and Jun Xiao. 2021. Unified group fairness on federated learning. *arXiv preprint arXiv:2111.04986* (2021).

[77] Fan Zhou and Guojing Cong. 2017. On the convergence properties of a K-step averaging stochastic gradient descent algorithm for nonconvex optimization. *arXiv preprint arXiv:1708.01012* (2017).