LipFed: Mitigating Individual Subgroup Bias in Federated Learning with Lipschitz Constraints

KHOTSO SELIALIA, University of Massachusetts Amherst, USA
YASRA CHANDIO, University of Massachusetts Amherst, USA
JIMI OKE, University of Massachusetts Amherst, USA
FATIMA M. ANWAR, University of Massachusetts Amherst, USA

Federated learning (FL) trains decentralized machine learning models while preserving privacy. However, FL models are biased, leading to unfair model outcomes towards subgroups with intersecting attributes. However, FL models are biased, leading to unfair model outcomes towards subgroups with intersecting attributes. To address this, we propose LipFed, a subgroup bias mitigation technique that leverages Lipschitz-based fairness constraints to mitigate subgroup bias in FL. We evaluate LipFed's efficacy in achieving subgroup fairness across clients while preserving model utility. Our experiments on benchmark datasets and real-world datasets demonstrate that LipFed effectively mitigates subgroup bias without significantly compromising group fairness or model performance.

ACM Reference Format:
Khotso Selialia, Yasra Chandio, Jimi Oke, and Fatima M. Anwar. 2025. LipFed: Mitigating Individual Subgroup Bias in Federated Learning with Lipschitz Constraints. 1, 1 (January 2025), 31 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Federated learning (FL) trains a global model using private data from decentralized edge devices or clients without the need to collect this data centrally, thereby promoting collaborative learning while preserving data privacy [40]. FL is particularly well-suited for privacy-sensitive applications such as medical diagnosis [12, 29], gender prediction [28], next-character prediction [55], and activity recognition [10, 42, 54]. Despite the advantages of collaborative learning and privacy preservation, FL inevitably absorbs undesired biases from the statistically heterogeneous data of its clients [1]. For instance, a crime detection FL algorithm may base its predictions of crime suspects on skin color [4], leading to incorrect assumptions about who should be incarcerated [46]. If these biases remain unchecked, they can erode user trust and negatively impact experiences, thereby affecting the adoption and acceptance of FL.

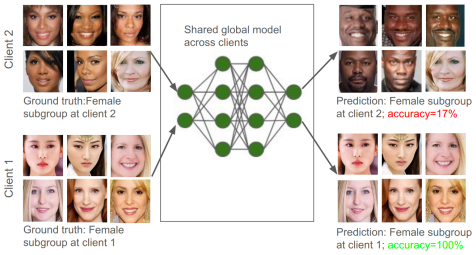


Fig. 1. Subgroup Bias in FL. The global model achieves 100% accuracy on Client 1's diverse subgroup but only 17% on Client 2's predominantly black women subgroup, highlighting bias from uneven data distribution across clients

Authors' addresses: Khotso Selialia, University of Massachusetts Amherst, Amherst, Massachusetts, USA, ; Yasra Chandio, University of Massachusetts Amherst, Amherst, Massachusetts, USA, ; Jimi Oke, University of Massachusetts Amherst, Amherst, Massachusetts, USA, ; Fatima M. Anwar, University of Massachusetts Amherst, Amherst, Massachusetts, USA, .

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

Recent FL research has focused on addressing bias, targeting *client bias* [6, 18, 22, 33, 35, 41] and *group bias* [5, 43, 62]. Client bias techniques aim to ensure similar model performance across clients [43], with approaches like [6, 40, 41] optimizing the worst-performing client's performance through importance weighting. In contrast, group fairness techniques (*fairness across multiple sensitive attributes* [57]) enforce fairness constraints for individual attributes like race, gender, or label [3, 43]. While group bias methods have progressed in developing fair FL algorithms; they do not always take into account how complicated it is to ensure that intersectional groups (*subgroups* [26]) are treated fairly. For example, subgroups like *black women* and *white women* (intersection of race and gender groups) share the *woman* attribute but may experience disparate treatment due to their distribution across decentralized clients. This oversight can lead to what we identify as a fairness gerrymandering problem in FL, where seemingly fair metrics at the group level obscure biases at the distributed intersectional subgroup level. In this paper, we call this idea of bias *individual subgroup bias*. This phenomenon is illustrated in **Example 1** and shown in Figure 1.

Example 1. *In a hypothetical scenario with race (black and white) and gender (male and female) as groups, consider a classifier predicting positive outcomes only for black men or white women. This classifier appears fair across groups, predicting positively for both men and women by* 50% *and both black and white groups by* 50% *However, examining subgroups like black and white women violates statistical parity fairness. For instance, black women may be disproportionately labeled unfavorably in comparison with white women, causing an unfair disadvantage for this intersectional subgroup (black women) . This example demonstrates Simpson's Paradox [44] in fairness evaluation, where seemingly fair techniques for groups become unfair for their fine-grained subgroups.*

In FL, *individual subgroup bias* arises because subgroup data across clients, despite sharing similar attributes, fails to be *independent and identically distributed (IID)*. This deviation from IID-ness, influenced by diverse feature distributions among individual subgroups due to geographical locations, weather conditions, and variations in data collection device specifications, leads to significant data heterogeneity [20, 39]. For instance, images of *black and white* female faces may differ globally due to skin color and varying data collection device quality. Addressing individual subgroup bias challenges is both a technical and social imperative [13, 45]. Individual fairness is crucial to addressing the ethical need to treat individuals with equity and respect for their unique attributes, consistent decisions or outcomes among similar individuals, and ensuring commitment to justice that acknowledges individual differences while preventing bias. Maintaining fairness treats similar individuals equitably and is vital for building trustworthy systems, particularly in domains like employment and credit, where biases can significantly impact lives and society. Based on these individual fairness requirements and to bridge the technical challenges of non-IID data in FL with the social need to develop algorithms that reflect and respect diverse global communities, we ask the following research question: ***How can FL models achieve individual subgroup fairness, particularly across subgroups with similar attributes distributed across multiple clients, without compromising overall group fairness and model utility?***

To answer our research question, in this paper, we propose *Lipschitz Fair Federated Learning (`LipFed`)* to mitigate individual subgroup bias in FL. Inspired by the principles outlined in [9], which advocate for treating similar individuals similarly by adhering to a Lipschitz condition for a task-relevant metric, our novel framework `LipFed` applies the Lipschitz property to ensure equitable outcomes for subgroups with similar attributes across distributed clients. `LipFed` adapts Lipschitz constraints to address the unique challenges of FL (§4.1). By applying a robust distance metric, our framework quantifies subgroup similarities and their model performance across clients (§4.2.1). This ensures that small differences in input characteristics in similar subgroups do not cause disproportionate performance disparities, effectively addressing decentralized data complexities and promoting individual subgroup fairness in FL (§4.2.2).

Contributions. In summary, we make the following contributions:

(1) We identify the *individual subgroup bias* problem in FL (§3), focusing on bias at the individual subgroup level rather than statistical bias across fixed demographic groups, addressing intersectional biases more comprehensively, ensuring fairness, and reducing discrimination based on intersecting attributes.
(2) We propose `LipFed` (§4), which uses the Lipschitz property to train fair models for individual subgroups in FL. `LipFed` makes sure that small changes in sensitive features cause small changes in model predictions, guaranteeing fair results to individual subgroups distributed across clients.
(3) We conduct theoretical analysis and establish precise bounds for subgroup and statistical performance. By providing clear performance bounds (§5), our work promotes a more transparent and accountable approach to addressing individual subgroup and statistical fairness challenges, fostering trust and reliability in FL.
(4) We apply the `LipFed` across six datasets (§6), reducing individual subgroup bias by up to 49% without substantially degrading model utility, though with some trade-offs in statistical fairness, clarified through our theoretical analysis (§5.2). `LipFed` also improves other existing FL methods by up to 25% in mitigating global subgroup bias.

2 BACKGROUND AND RELATED WORK

2.1 Fairness in Machine Learning

Machine learning algorithms aimed at achieving fair models are typically classified into distinct categories such as: *individual fairness*[9] and *group fairness*[16, 23, 24, 38, 63].

Individual Fairness emphasizes the principle that similar individuals should be treated similarly in decision-making processes[9]. Individual bias, which unfairly discriminates against individuals based on sensitive attributes, is the origin of this approach. The challenge lies in developing machine learning models that improve individual fairness by ensuring that similar individuals receive similar outcomes. Solutions involve developing fairness metrics that quantitatively describe the similarity between individuals and implementing algorithms that optimize these metrics while maintaining utility. This often includes using Lipschitz conditions to ensure that the difference in treatment between any two individuals does not exceed their distance according to the fairness metric. Methods focusing on individual fairness generally show a reduction in individual biases, contributing to more equitable machine learning deployment.

Group Fairness ensures that algorithms are equitable across different groups, typically defined by sensitive attributes such as race or gender[16, 23, 24, 38, 63]. The core challenge here is the unintended bias that may arise, leading to disparities in decisions that affect employment, law enforcement, and loan approval processes[1]. Researchers aim to develop models that do not disproportionately favor any group based on these attributes. Common approaches include adjusting the training data or the model itself to balance treatment and error rates across groups. Techniques such as re-weighting training examples, modifying objective functions to include fairness constraints, or using post-processing adjustments to balance decision thresholds are typically employed. The outcomes of these methods generally demonstrate a reduction in bias, leading to algorithms that perform uniformly across different demographics.

2.2 Fairness in Federated Learning

In this section, we review methods in FL fairness, focusing on those related to individual subgroup fairness. While subgroup fairness is recognized in centralized learning [9], we address the unique challenges of FL and highlight the limitations of existing approaches. FL algorithms aimed at achieving fair models are typically classified into three distinct categories, including *client-fairness*[6, 18, 22, 33, 35, 41], *group-fairness*[5, 43, 52, 62], and *robustness techniques* [25, 32].

Client Fairness. Ensuring fairness among clients in FL is essential to mitigate biases from non-IID data distributions across devices. Techniques such as Federated Fair Averaging (FedFV) [59] adjust gradient directions and magnitudes to

balance model *average performance* based on client contributions [43], while GIFair-FL [61] dynamically modifies model updates with a fairness-aware aggregator to reduce *average loss*. FjORD [18] employs ordered dropout to customize model sizes to client capacities, enhancing both fairness and accuracy. Additionally, Agnostic Federated Learning (AFL) [41] tailors the global model to any client distribution mix, q-FFL [35] reweights losses to favor lower-performing devices, and Tilted Empirical Risk Minimization (TERM) [33] fine-tunes outlier impact and class representation, collectively improving *average performance* in diverse environments.

Group Fairness. Recent advancements in FL emphasize addressing group fairness and biases against protected groups. FairFed [11] uses fairness-aware aggregation and local debiasing to enhance group fairness under heterogeneous data conditions. FedMinMax [43] employs alternating optimization for minimax fairness across demographic groups, showing competitive performance. FCFL [5] combines algorithmic fairness and performance consistency, achieving Pareto optimality via gradient-based methods and outperforming existing models in fairness and utility.

Robustness in FL [31, 34] address data heterogeneity and forgetting in FL, where a global model is trained collaboratively without direct access to clients' data. Drawing parallels to continual learning, they argue that forgetting impedes FL convergence. The global model forgets previous knowledge, and local training induces forgetting outside the local distribution. To mitigate this, they propose Federated Not-True Distillation (FedNTD) to preserve global knowledge on non-local classes by reducing forgetting and achieves state-of-the-art performance, improving prediction consistency.

2.3 Limitations of Existing Techniques

While valuable, current bias mitigation techniques in both centralized and FL settings have notable limitations. Firstly, centralized bias mitigation techniques are designed for centralized settings and thus cannot be directly applied to FL settings. This limitation is due to their reliance on raw data sharing, which could lead to sensitive data leakage and thus fail to preserve privacy in FL. Secondly, existing bias mitigation approaches in FL mostly concentrate on client and group fairness, often neglecting to ensure fairness for individual subgroups with overlapping characteristics. According to Simpson's Paradox [44], techniques that appear fair at the group level may still result in unfair outcomes for more finely defined subgroups. The subsequent section will explore these deviations and their implications through empirical studies.

3 PRELIMINARY STUDY AND PROBLEM FORMULATION

This section defines formal definitions of FL and the problem of *individual subgroup fairness* addressed in this paper, establishing the study's framework. Specifically, this section covers the local data heterogeneity of decentralized FL clients, how FL learns from such heterogeneous data across clients, and individual subgroup fairness in FL. In this section, the key question we aim to answer through an empirical study is: ***what is the effect of data heterogeneity on individual subgroup fairness across clients in FL?***

3.1 Preliminaries

Federated Learning (FL) trains a global model using a server and K decentralized clients, ensuring privacy by not sharing their local data. Each client $k \in K$ has its private local dataset $\mathscr{D}_k = \{\boldsymbol{X}_k, \boldsymbol{Y}_k\}$, with N_k tuples $\{(\boldsymbol{x}_k^n \in \boldsymbol{X}_k, y_k^n \in \boldsymbol{Y}_k)\}_{n=1}^{N_k}$ representing input and output spaces. These private datasets can be grouped by attributes like race, gender, or label [3]. The local group dataset on client k is $\mathscr{D}_{g,k} = \{\boldsymbol{X}_k^g, \boldsymbol{Y}_k^g\}^{N_{g,k}}$ with $N_{g,k} \leq N_k$ samples where $g \in G$ indicates group membership. In ideal IID scenarios, clients sample $\mathscr{D}_{g,k}$ independently from a global distribution $f_g(\boldsymbol{X})$. However, real-world FL scenarios often feature non-IID/heterogeneous group data due to factors such as *inter-partition decorrelation* [20, 37], which occurs when clients fail to share standard features, resulting in decorrelated local group data across clients.

Subgroups. FL indirectly aggregates non-IID local group data from decentralized clients into a unified dataset, $\mathscr{D} = \bigcup_{k=1}^{K} \mathscr{D}_k$. Each client's data $\mathscr{D}_k$ is partitioned into local groups $\mathscr{D}_{g,k}$ based on attributes such as gender, race, or label. These local groups can overlap per client, forming *subgroups* defined by the combinations of attributes relevant to the analysis or application, such as intersections of race and gender (Black women, Asian women). This local definition of subgroups helps in tailoring models to respect and address fine-grained, context-driven fairness concerns that may arise within a single client's data scope.

Individual Subgroups. Expanding the local subgroup concept to a global scale in FL, we introduce the notion of a *individual subgroup*. In this work, we define *individual subgroups* as the aggregation of subgroups with similarities across different clients that, while distinct, share underlying attributes. For instance, the individual subgroups that share "woman" as an attribute might include local subgroups like "black women" on client 1 and "white women" on client 2. This framework ensures that similar subgroups, regardless of their global context, are treated similarly, following the principles of fairness outlined in [9]. By considering individual subgroups, we can enforce fairness constraints that are globally equitable, improving the model's ability to be fair to diverse populations while addressing differences between subgroups across clients. This approach prevents fairness measures from being biased by ignoring subgroups defined by intersecting attributes and lays the foundation for fairness-aware algorithms in FL environments.

FL uses the unified dataset $\mathscr{D}$ to learn an optimal global model h^* (with global parameters $\boldsymbol{\theta}$) from a class of hypotheses H that map input features $\boldsymbol{x}_k^n$ to outputs y_k^n. The optimal model minimizes the *empirical risk* objective with R_k as the empirical risk for client k with local parameters $\boldsymbol{\theta}_k$ as:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left\{ R(\cdot;\boldsymbol{\theta}) = \sum_{k=1}^{K} \left(\frac{N_k}{\sum_{k=1}^{K} N_k} \right) R_k(h_{\boldsymbol{\theta}}(\boldsymbol{X}_k), \boldsymbol{Y}_k) \right\} \tag{1}$$

Individual Subgroup Fairness in FL. Many FL works aim to achieve a modified formulation of Equation 1 for group-fair model parameters [33, 41, 61], often overlooking individual subgroup fairness. Suppose that there are n_g individual subgroups $\{g_j\}_{j=1}^{n_g}$ (e.g,. "black women", "white women", etc.) distributed across clients. Let the performance measures of models h_1 and model h_2 for these subgroups be represented as true positive rates (TPR) be $\{a_1^{g_j}\}_{j=1}^{n_g}$ and $\{a_2^{g_j}\}_{j=1}^{n_g}$, respectively. Model h_1 is more subgroup fair than model h_2 if $Var_{h_1}(\{a_1^{g_j}\}_{j=1}^{n_g}) < Var_{h_2}(\{a_2^{g_j}\}_{j=1}^{n_g})$, where Var_h is the performance variance used in recent FL fairness works [61]. Higher performance variance indicates greater variation in individual subgroup performance metrics, indicating potential bias. Performance is measured using $TPR_g = \frac{TP_g}{TP_g + FN_g}$ from fairness-aware optimization in FL [47] where TP_g counts true positives (correctly classified instances) and FN_g counts false negatives (incorrectly classified instances) for subgroup g_j (for more details on subgroup fairness see §B).

3.2 Experimental Setup

To examine the impact of non-IID data on individual subgroup bias in FL, we conduct experiments on classification taks using FedAvg to aggregate local models. We use four deep learning models across six datasets (two benchmarks, two real-world, and two fairness-based and large-scale), partitioned based on non-IID features across $K = \{5, 10\}$ clients. For model setup, ResNet [17] is applied to FER2013 [15] for emotion recognition (grouping seven emotions [43]), LeNet [30] for MNIST [2] (with each digit as a group), VGGNet [7] for FashionMNIST [60] (with each product as a group), ResNet for UTK [51] (for gender prediction), and Logistic Regression [19] for two ACS datasets [8] (for income: ASCI and employment prediction: ASCE). For ASCI, data is distributed by state to form two groups (Income True/False), with the state acting as an implicit sensitive attribute. For ASCE, data is filtered for individuals aged 16 to 90, forming employed/unemployed groups (see §G.3). *Note: Though our experiments involve a limited number of datasets and clients,*

the theoretical guarantees in §5 ensure that `LipFed`*'s fairness and utility-scale are reliable for the scope of the academic paper. These guarantees validate the robustness of our approach, even in broader FL settings.*

Data Partitions. Benchmark and real-world datasets are partitioned across clients using a Dirichlet distribution [21, 57]. For the income and employment tasks, data is naturally partitioned across approximately 50 clients, allowing us to validate the scalability of our approach in more complex settings (more details about experimental setup can be found in G.

Heterogeneous Feature Distributions. We simulate feature heterogeneity across clients by generating Gaussian noisy images $\tilde{I}(x, y)$ from pristine images $I(x, y)$ (with each client hosting images of unique noise level) according to the expression ($\tilde{I}(x, y) = I(x, y) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$). We control the noise intensity through variance σ^2, with $\sigma \geq 0.03$ mimicking real-world conditions [14, 39, 49, 53]. Concurrently, the ACS fairness dataset, partitioned by state, captures unique demographic landscapes reflecting inherent feature heterogeneity in socio-economic factors like age, education, race, and occupation, where average income, education levels, and employment rates vary significantly across states.

Evaluation Metrics. The primary objective of bias mitigation in FL is to minimize variations in model performance outcomes for individual subgroups that, despite sharing similar attributes, are distributed across different clients while maintaining competitive utility; in doing so, we assess three key metrics (additional discussion in §H):

- *Individual subgroup bias metrics* measure performance variance across individual subgroups that, despite sharing similar attributes, are distributed across different clients. We compute worst individual subgroup variance $\{Var_h(\{a^{g_j}\}_{k=1}^{n_g})\}$, where a^{g_j} is the model's performance on a subgroup. Low variance values (approaching zero) indicate low individual subgroup bias.
- *Group bias metrics* measure performance variance across groups. We compute worst group variance $\{Var_h(\{a^g\}_{g=1}^{G})\}$, where a^g is the model's performance on a local group at client k. Low variance values (approaching zero) indicate low group bias.
- *Utility metrics* measure overall model performance across clients. Utility is assessed using the average accuracy across all clients.

3.3 nonIID Study: Results Overview

This section addresses the impact of data heterogeneity on individual subgroup fairness across clients in FL. We summarize our findings based on the *individual subgroup bias* and *utility metrics* in Figure 2.

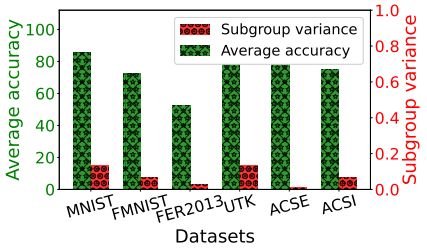


Fig. 2. Variation in TPR among subgroups and average model utility across clients.

Observation: The preliminary study results depicted in Figure 2 reveal a variation in the performance of the global model across individual subgroups that, despite sharing similar attributes, are distributed across different clients, highlighting the impact of feature distribution heterogeneity on individual subgroup fairness. Specifically, across all datasets, the individual subgroup bias metrics exhibit values consistently greater than zero, indicating non-negligible performance variation across subgroups. The magnitude of individual subgroup bias varies across datasets, with MNIST exhibiting the highest levels. Our intuition behind this pronounced disparity is attributed to the dataset's inherent simplicity, which makes the model more sensitive to performance degradation across individual subgroups due to feature distribution variations. Despite high average accuracy across all datasets, individual subgroup bias persists, demonstrating a decoupling between utility and fairness. This problem demonstrates the inadequacy of utility-focused metrics as standalone measures of fairness. The observed results align with the theoretical relationship established in §5, where individual subgroup performance

outcomes are linked to data heterogeneity quantified by Γ. The theoretical relationship suggests that Γ, which encapsulates inter-client variability in feature distributions, plays a pivotal role in driving individual subgroup bias. These observations motivate the need for advanced fairness-aware mechanisms, such as Lipschitz-based constraints, which can mitigate global subgroup bias by enforcing stability in model predictions across individual subgroups.

Takeaway: *Non-IID subgroup data across clients causes global bias; even high-utility methods may fail. Addressing this bias is crucial for improving FL's fairness and utility.*

4 LIPFED OPTIMIZATION FRAMEWORK

In this section, we will (1) define the Lipschitz property and its importance in achieving fairness, (2) describe the challenges of formalization of Lipschitz fairness constraints in FL, and (3) introduce the foundational fairness constraints that form the basis of our proposed technique, `LipFed`, for training fair FL models across subgroup X_k^g at client k and all similar *individual subgroups* $X_{k'}^g$ (sharing similar attributes) across different clients k', as illustrated in Figure 3.

4.1 Overview and Challenges

Lipschitz Fairness. Individual-level fairness for similar entities $\boldsymbol{x}$ and $\boldsymbol{x}'$, where the similarity of these entities is quantified by the distance metric $d(\boldsymbol{x}, \boldsymbol{x}')$, can be achieved by optimizing the model to satisfy the Lipschitz property [9].

Definition 3.1 (Lipschitz property). A model $h_{\boldsymbol{\theta}} : G \to \Delta(A)$ satisfies the (D, d) Lipschitz property if for every $\boldsymbol{x}, \boldsymbol{x}' \in G \quad \exists \epsilon > 0$ such that:

$$D(h_{\boldsymbol{\theta}}(\boldsymbol{x}), h_{\boldsymbol{\theta}}(\boldsymbol{x}')) \leq \epsilon \cdot d(\boldsymbol{x}, \boldsymbol{x}'). \tag{2}$$

Here, $d : G \times G \longrightarrow \mathbb{R}$ quantifies the similarity between individuals. Without a well-defined metric, $d(\cdot)$ reflects the "best" available approximation agreed upon by society [9]. $h_{\boldsymbol{\theta}} : G \to \Delta(A)$ maps individual entities $\boldsymbol{x}, \boldsymbol{x}' \in G$ to outcomes in $\Delta(A)$ (e.g., an individual's TPR).

Lipschitz Fairness in FL: In FL, the global model is aggregated from diverse local models trained on data subsets $\mathscr{D}_k$, each potentially reflecting different subgroup characteristics. Using the Lipschitz property in `LipFed`, we can ensure that similar subgroups, defined globally across clients, are treated comparably. This involves (1) defining an appropriate similarity metric $d(\cdot)$ to capture the similarity between individual subgroups and (2) implementing global model updates to ensure the (D, d) Lipschitz condition is met across all individual subgroups (for more details see§C).

By limiting the distance in model performance between any similar individuals $\boldsymbol{x}$ and $\boldsymbol{x}'$ to the product of their similarity measure $d(\boldsymbol{x}, \boldsymbol{x}')$ and a small constant ϵ, that can enable similar performance outcomes for individual subgroups. We enable this constraint by using the Lipschitz property defined in Equation 2. The individual subgroup fairness constraint $\mathscr{C}_f(\boldsymbol{\theta})$ over each client's local *empirical risk* $R(\boldsymbol{X}_k; \boldsymbol{\theta})$ is defined as:

$$\min_{\boldsymbol{\theta}} R(X_k, \boldsymbol{\theta}) \quad \text{s.t} \quad \forall X_k^g, X_{k'}^g \in G : \mathscr{C}_f(\boldsymbol{\theta}) = D(h_{\boldsymbol{\theta}}(X_k^g), h_{\boldsymbol{\theta}}(X_{k'}^g)) = \|h_{\boldsymbol{\theta}}(X_k^g) - h_{\boldsymbol{\theta}}(X_{k'}^g)\| \leq d(X_k^g, X_{k'}^g) \tag{3}$$

Challenges in Applying Lipschitz Fairness in FL: In diverse FL edge deployments with non-IID local data, the global model aggregated via FedAvg [40] can converge to a Lipschitz-fair model towards *subgroups with similarities* across clients. But the Lipschitz condition proposed in Equation 3 requires that individuals subgroups X_k^g and $X_{k'}^g$ should have outputs $h_{\boldsymbol{\theta}}(X_k^g)$ and $h_{\boldsymbol{\theta}}(X_{k'}^g)$ with the Euclidean distance $D(h_{\boldsymbol{\theta}}(X_k^g), h_{\boldsymbol{\theta}}(X_{k'}^g))$ between $h_{\boldsymbol{\theta}}(X_k^g)$ and $h_{\boldsymbol{\theta}}(X_{k'}^g)$ at most $d(X_k^g, X_{k'}^g)$. So, applying the Lipschitz condition for individual subgroup fairness in FL poses two major challenges:

- *Lack of well-defined similarity metric:* No well-defined metric $d(\cdot)$ exists to assess the similarity between individual subgroups X_k^g and $X_{k'}^g$.

- *Decentralized subgroups:* Unlike centralized machine learning, individual subgroups X_k^g and $X_{k'}^g$ are spread across clients in FL, making it difficult to assess and impose the Lipschitz condition without breaking FL privacy.

4.2 Our Solution: `LipFed` Framework

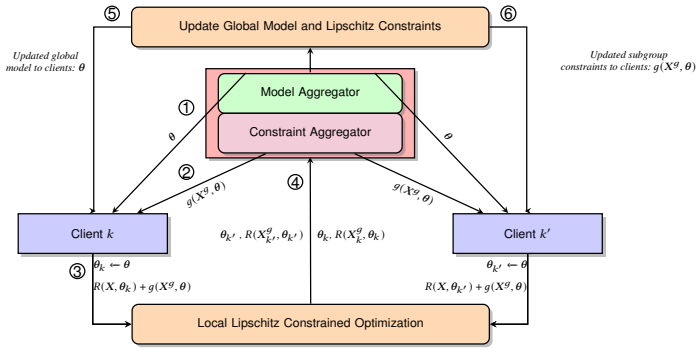


Fig. 3. Schematic of our proposed individual subgroup bias mitigation approach `LipFed`. X_k^g is the subgroup data and θ_k are the local model parameters for client k. $R(X_k^g, \theta_k)$ and $R(X, \theta_k)$ measure the subgroup and overall data risks, respectively. The numbered circle indicates sequential FL steps.

4.2.1 Subgroup Similarity Based Distance Metric $d(\cdot)$. In this section, we explain the foundation for subgroup fairness through an interpretable and meaningful distance metric. To address the challenges associated with defining and implementing a distance metric $d(\cdot)$ for assessing the similarity between individual subgroups in FL, we refine the concept of individual subgroup similarity. Subgroups denoted as X_k^g comprise samples from intersectional groups that belong to specific groups $g \in G$. The similarity metric between these subgroups across clients is designed to capture the essence of intersectional characteristics that they share. The proposed similarity metric is based on client subgroups' most commonalities. For example, in the case of *black and white women*, the shared characteristic of being *women* is utilized as the primary basis for defining similarity. This approach hypothesizes that such commonalities are the best available approximation for assessing subgroup similarity in a decentralized data environment. The distance metric is designed and approximated such that the distance between the loss outcomes of individual subgroups is minimized, ideally not exceeding a small threshold ($\|h_\theta(X_k^g) - h_\theta(X_{k'}^g)\| \approx d(X_k^g, X_{k'}^g) \leq \epsilon$). This minimization reflects the foundational assumption that subgroups sharing key attributes should yield similar performance outputs, thereby ensuring that the model's fairness extends across the network while respecting the inherent diversity and privacy of the FL setup.

4.2.2 Privacy-preserving Decentralized Lipschitz Constraints. Next, we demonstrate the process of how `LipFed` addresses privacy concerns while enabling individual subgroup fairness constraints in FL. In FL, a central server coordinates training across multiple clients, each possessing a unique local subgroup dataset X_k^g. The FL process follows the steps in Figure 3: ① The *Model Aggregator* initializes a global model, θ, and broadcasts it to all participating clients $k, k' \in K$. ③ Each client trains θ on its local subgroup dataset, optimizing local losses $R(X_k, \theta)$ and $R(X_{k'}, \theta)$. ④ After local training, each client computes and sends its updated local model parameters θ_k to the server for aggregation into a global model. ⑤ The updated aggregated model is iteratively sent back to the clients, enabling further local training and refinement until convergence or a specified stopping criterion is met. While this procedure enables collaborative learning, individual subgroup fairness is not guaranteed (Section 3). To address this challenge, `LipFed` integrates fairness

constraints into the standard FL workflow (② and (⑥)). Specifically, after each local training step (③), the *Constraint Aggregator* updates individual subgroup fairness constraints as discussed below:

Individual Subgroup Fairness Constraint $\mathscr{C}_f$. Computing individual subgroup constraints across subgroup outcomes $D(h_\theta(X_k^g), h_\theta(X_{k'}^g)) = \|h_\theta(X_k^g) - h_\theta(X_{k'}^g)\|$ at client k hosting the local subgroup X_k^g poses privacy issues, as there's a lack of global information about decentralized subgroups $X_{k'}^g$ residing on other clients k'. To solve this issue, the *Constraint Aggregator* receives subgroup losses in a privacy-preserving manner from individual clients (④) and computes the individual subgroup constraint for client k as the difference $\mathscr{C}_f = \|h_\theta(X_k^g) - h_\theta(X_{k'}^g)\|$ between client k's subgroup loss $h_\theta(X_k^g)$ and the weighted aggregation of decentralized subgroup losses $h_\theta(X_{k'}^g)$ from other clients' k' as:

$$\mathscr{C}_f = D(h_\theta(X_k^g;\theta), h_\theta(X_k^g;\theta) = \|R(X_k^g;\theta) - R(X_{k'}^g;\theta)\| \approx \|R(X_k^g;\theta) - \sum_{k'} w_{g,k'} R(X_{k'}^g;\theta)\| \tag{4}$$

where $w_{g,k'}$ denotes the relative importance subgroup loss weight for client k' in the aggregation The expression $\|\cdot\|$ quantifies the total discrepancy or distance between the loss performances of the global model on client k's subgroup and the weighted subgroup losses across other clients k'. A small discrepancy value indicates that the model's individual subgroup performance aligns well with all clients' collective performance without bias. At the beginning of the next round, the server distributes $\mathscr{C}_f$ (②) for each client k to carry out the following local constrained optimization:

$$\min_\theta \frac{1}{K}\sum_{k=1}^{K} R(X_k,\theta) \quad \text{s.t} \quad \forall X_k^g, X_{k'}^g \in G : \mathscr{C}_f(\theta) = \sum_{g=1}^{G} \|R(X_k^g;\theta) - \sum_{k'} w_{g,k'} R(X_{k'}^g;\theta)\| \leq \epsilon \tag{5}$$

$\mathscr{C}_f$ in Equation 5 ensures that the difference between the loss of a subgroup on client k and the aggregated losses of subgroup with overlaps across other clients k' is small.

Decentralized Aggregation with Importance Weights $w_{g,k'}$. In `LipFed`, individual subgroup importance weights are computed as the ratio of the number of samples in a subgroup on a client to the total number of samples across all clients, reflecting the subgroup's proportional contribution to the global model. This design is motivated by prior work in FL [40], which demonstrates that sample-proportional weighting ensures the aggregated model reflects the global data distribution effectively. Formally, $w_{g,k'} = \frac{n_{g,k'}}{N}$, where $n_{g,k'}$ is the number of samples in a subgroup on client k', and N is the total sample size across all clients. This approach ensures that subgroups with larger representation influence the global model proportionally, while smaller subgroups are not overshadowed, promoting a balanced contribution.

Differentially Private Individual Subgroup Fairness Constraints. Inspired by the work in [1], which safeguards the privacy of sensitive metadata exchanged between the server and clients during training, we protect the sensitive loss outcomes of individual subgroups $R(Xk'^g;\theta)$. These are shared by each client with the server to compute components of individual subgroup fairness constraints $\sum_{g=1}^{G}\|\sum_{k'} w_{g,k'} R(X_{k'}^g;\theta)$. To achieve the privacy goal, we use differential privacy (DP) [1] by introducing varying levels of Laplace noise [50] to the local subgroup losses. The implementation of DP via Laplace noise guarantees the confidentiality of individual subgroup data while still enabling the assessment and enforcement of fairness constraints across diverse subgroups.

4.2.3 Reformulation to an Unconstrained Problem. Now to simplify `LipFed` optimization to make fairness constraints practically enforceable in FL, we reformulate Equation 5 using a linear penalty approach, yielding an unconstrained problem formulation:

$$\min_\theta \left\{(1-\lambda)R(X_k,\theta) + \lambda \max\left(0, g(X_g;\theta)\right)\right\} \tag{6}$$

$g(\boldsymbol{X}_g; \boldsymbol{\theta}) = \mathscr{C}_f(\boldsymbol{\theta}) - \epsilon$ at ③. This formulation aims to minimize the empirical risk while enforcing the constraint that the performance across individual subgroups should not degrade below a threshold defined by ϵ. This approach balances the primary objective with the need to meet the fairness constraint, thereby mitigating individual subgroup performance discrepancies in non-IID data settings without significantly compromising overall group fairness.

4.2.4 Balancing Fairness and Utility with λ. The optimal value of λ, which balances fairness and utility in a model, is context-dependent and stems from specific priorities of the application and the data. Higher values of λ prioritize fairness, enhancing equity across individual subgroups, which is essential in high-stakes domains like healthcare or finance. Conversely, lower λ values emphasize the model's utility, focusing on performance metrics such as accuracy, suitable for less sensitive applications. Determining the best λ involves empirical testing, such as cross-validation, to observe how changes affect both fairness metrics (e.g., error rate disparities) and utility metrics (e.g., overall accuracy). Sensitivity analysis can help assess the impact on subgroups, ensuring the model does not disproportionately benefit or harm individual subgroups. Engaging with stakeholders and considering ethical implications also play critical roles in setting λ, ensuring that the model aligns with societal norms and regulatory expectations. Ultimately, choosing λ is an iterative process that may require dynamic adjustments as new data emerge and societal expectations evolve.

5 THEORETICAL ANALYSIS

This section presents a comprehensive theoretical analysis of individual subgroup and group fairness in FL models. We explore the empirical loss upper bounds for `LipFed` optimization and delve into the trade-offs between Lipschitz continuity, empirical risk, and fairness constraints. These analyses provide crucial insights into how model properties relate to empirical risk outcomes, which is essential for quantifying fairness (for assumptions and more details, see §D).

5.1 Empirical Loss Bounds for `LipFed`

Here, we establish the theoretical upper bounds of empirical loss for both individual subgroups and broader groups within the context of FL, which are fundamental to understanding the performance capabilities and limitations of `LipFed`.

5.1.1 Subgroup Empirical Loss Upper Bound.

Theorem 5.1.1. ***Subgroup empirical loss upper bound.*** *Under* ***Assumption 1-4*** *in §D on the global empirical risk function $R(\boldsymbol{\theta})$ as per recent FL works [35, 36], we have:*

$$\mathbb{E}[R(\boldsymbol{\theta}_T) - R^*] \leq \frac{L}{2}\frac{v}{\gamma + T} + \frac{\kappa}{\gamma + T}\left(\frac{2 \cdot \sum_{k=1}^{K} p_k^2 \sigma_k^2 + 6L\Gamma_1 + 8(E-1)^2 G^2}{\mu} + \frac{\mu(\gamma+1)}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|^2\right), \tag{7}$$

where Γ_1 quantifies the degree of individual subgroup data heterogeneity across clients; if the data for subgroups with similarities across clients is non-iid, then Γ_1 is nonzero, and its magnitude reflects the heterogeneity of the individual subgroup data distribution [36]. p_k is the weight of the k-th device such that p_k is proportional to the device's local data size and $p_k \geq 0$, E is the number of local training rounds/epochs for each device k, and $\gamma = \max\{8\kappa, E\}$.

5.1.2 Group Empirical Loss Upper Bound.

Theorem 5.1.2. ***Group empirical loss upper bound.*** *Under* ***Assumption 1-4*** *in §D on the global empirical risk function $R(\boldsymbol{\theta})$ as per recent FL works [35, 36], we have:*

$$\mathbb{E}[R(\boldsymbol{\theta}_T) - R^*] \leq \frac{L}{2}\frac{v}{\gamma + T} + \frac{\kappa}{\gamma + T}\left(\frac{2 \cdot \sum_{k=1}^{K} p_k^2 \sigma_k^2 + 6L\Gamma_2 + 8(E-1)^2 G^2}{\mu} + \frac{\mu(\gamma+1)}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|^2\right), \tag{8}$$

where Γ_2 quantifies the degree of local group data heterogeneity across distinct groups; if the data for distinct groups is non-iid, then Γ_2 is nonzero, and its magnitude reflects the heterogeneity of the global group data distribution [36]. p_k is the weight of the k-th device such that p_k is proportional to the device's local data size and $p_k \geq 0$, E is the number of local training rounds/epochs for each device k, and $\gamma = \max\{8\kappa, E\}$.

5.2 Trade-Off Analysis Between Individual Subgroup and Group Fairness

This subsection examines how the previously established empirical loss bounds impact the model's ability to achieve fairness across both individual subgroups and broader groups, focusing on the roles of Lipschitz continuity L and data heterogeneity terms Γ_1 and Γ_2.

5.2.1 Impact of the Lipschitz Constant and Data Heterogeneity on Fairness. The Lipschitz constant L acts as a regulatory mechanism that controls how much one can change the model's parameters in response to changes in input data, affecting how tightly the model fits to individual subgroup characteristics. A lower L value enhances individual subgroup fairness by reducing their model performance variance, thereby minimizing the impact of Γ_1, the measure of individual subgroup heterogeneity. This is advantageous for achieving uniformity and fairness across individual subgroups despite being located across different clients. However, the benefits of a lower L for subgroup fairness do not guarantee local group fairness, particularly when distinct local group heterogeneity Γ_2 is significant. This suggests that fairness to individual subgroups may compromise equity across more diverse groups due to data heterogeneity of group data.

5.2.2 Optimizing the Lipschitz Constant. Optimizing L involves finding a balance such that both $L\Gamma_1$ and $L\Gamma_2$ are minimized. The ideal L should be flexible enough to allow for adjustments to individual subgroup characteristics without causing significant disparities within broader groups. This might involve dynamically adjusting L during different phases of model training or for different segments of data to cater both to subgroup nuances and group-level diversity using controlled regularization as in §6.

5.2.3 Formalizing the Trade-Off. The theoretical upper bounds derived from our theorems quantify individual subgroup and group performance trade-off dynamics. For instance, the individual subgroup performance bound indicates how lowering L decreases $L\Gamma_1$, reducing subgroup errors. Conversely, the group fairness bound reflects how the same reduction in L can not have a major impact in reducing $L\Gamma_2$ if there are significant complexities across unique groups, indicating potential disparities in group-level errors.

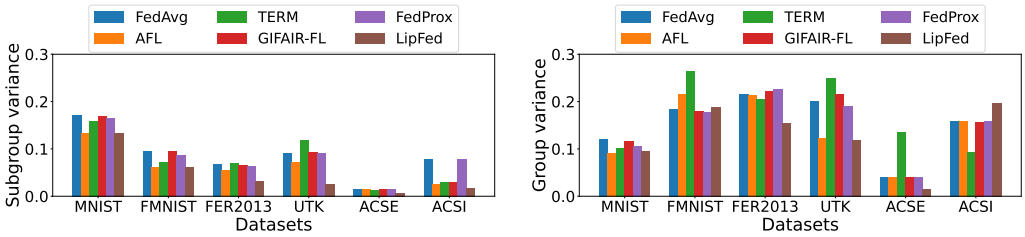
Takeaway: *This analysis shows that fine-tuning L can improve subgroup fairness but may impact group fairness in diverse datasets. Future work should explore adaptive strategies to balance these factors, potentially through machine learning techniques that adjust L in response to real-time assessments of individual subgroups and group heterogeneity.*

6 EVALUATION

In this section, we evaluate `LipFed`'s effectiveness in achieving individual subgroup fairness for subgroups with intersecting attributes that are distributed across different clients while adhering to three key constraints: (1) maintaining group fairness, (2) preserving model utility, and (3) data privacy.

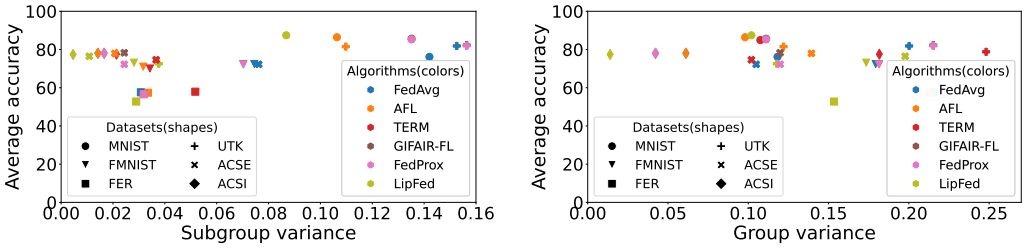
6.1 Experimental Setup

Models and datasets. Our study assesses `LipFed`'s efficacy using the setup in §3.2. We compare `LipFed` with SOTA baselines on benchmark datasets and evaluate its real-world applicability using the UTK dataset and ACS fairness datasets, examining bias mitigation across different client partitions in FL.



(a) TPR variations in subgroups. (b) TPR variations in groups.

Fig. 4. Demonstrating individual subgroup bias in model performance for different datasets and baselines.



(a) TPR variations in subgroups. (b) TPR variations in groups.

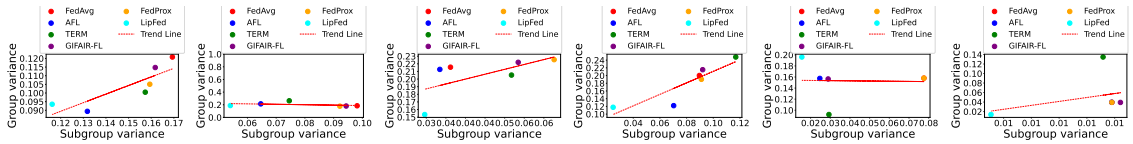
Fig. 5. Demonstrating model utility vs. discrepancy for different datasets and baselines.

Baselines. We evaluate `LipFed` across three key categories of state-of-the-art FL techniques: 1) The *FL baseline category*, represented by FedAvg and FedProx [34], serve as the standard learning schemes in FL. 2) The *FL client and group bias-reduction category* includes AFL[41], TERM[33], and GIFAIR-FL [62], which use empirical risk reweighting strategies to mitigate bias and adapt the global model to diverse local data distributions.*Note: We use client and group bias baselines, as to the best of our knowledge, no existing techniques are specifically designed to address subgroup bias. We provide an additional evaluation of FL robustness techniques that are not specifically focused on fairness in §I.1.*

6.2 Comparative Evaluation of `LipFed` on Benchmark and Real-World Datasets

We use six datasets to compare `LipFed` with bias mitigation baselines in achieving individual subgroup fairness. In the MNIST and Fashion-MNIST datasets, `LipFed` significantly outperforms baselines in reducing individual subgroup bias, as illustrated in Figure 4a. This improvement is largely due to `LipFed`'s use of Lipschitz continuity constraints, which directly address discrepancies in individual subgroup performance. In contrast, existing fairness techniques focus primarily on group fairness, which does not inherently guarantee individual subgroup fairness. However, `LipFed` occasionally exhibits higher group performance variance group (Figure 4b), indicating that improving individual subgroup fairness does not always translate into improved group fairness, a point further explored in the theoretical analysis §D. Nevertheless, `LipFed` maintains competitive model utility compared to baseline methods not only at the individual subgroup level (Figure 5a) but also at the group level (Figure 5b). The trends are consistent in real-world datasets (FER2013, UTK, ACSI, and ACSE) with those observed in the benchmark datasets, validating `LipFed`'s ability to balance individual subgroup fairness and utility in practical, non-IID FL settings.

Takeaway: *`LipFed` mitigates individual subgroup bias for non-IID subgroups across clients and maintains competitive utility compared to baselines without compromising performance on all six datasets.*



(a) MNIST (b) FMNIST (c) FER2013 (d) UTK (e) ACSI (f) ACSE

Fig. 7. Group Fairness vs. Individual subgroup fairness on different baselines and datasets.

6.3 Impact of `LipFed` Integration with Traditional FL Methods on Subgroup Fairness

We evaluate the impact of combining `LipFed` with other FL algorithms, such as AFL and TERM, to reduce individual subgroup bias. Our goal is to *investigate whether `LipFed` can address individual subgroup fairness beyond the FedAvg technique, particularly in scenarios with feature heterogeneity*. By integrating `LipFed` with AFL and TERM, resulting in AFL+`LipFed` and TERM+`LipFed`, we aim to ensure consistent model performance across clients. Using the same datasets and metrics, we find that both AFL+`LipFed` and TERM+`LipFed` consistently demonstrate lower individual subgroup variance discrepancies compared to AFL and TERM alone (Figure 6). This improvement is driven by `LipFed`'s enforcement of Lipschitz continuity constraints, which specifically target and penalize individual subgroup performance discrepancies. In contrast, most fairness techniques focus primarily on group fairness, which is insufficient to fully address individual subgroup fairness challenges.

Takeaway. *`LipFed` enhances the effectiveness of other group fairness methods in FL in reducing subgroup bias.*

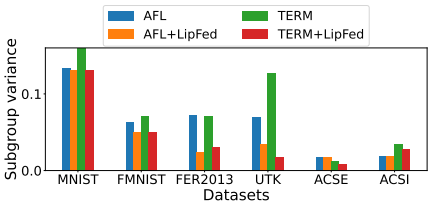


Fig. 6. Enhancing fairness across other FL algorithms: `LipFed` improves subgroup bias mitigation in FL algorithms across datasets.

6.4 Trade-off Between Individual Subgroup and Group Fairness

Figure 7 illustrates the empirical trade-off between individual subgroup and group fairness, complementing the theoretical analysis discussed earlier. The red lines indicate trends in various algorithms' ability to mitigate individual subgroup and group bias. A negative slope highlights the trade-off, where improving one type of fairness often compromises the other. `LipFed`, shown at the leftmost marker, effectively enhances individual subgroup fairness but slightly compromises group fairness due to the challenge of balancing these trade-offs during optimization. The mixed trends observed can be attributed to *Dataset characteristics and feature distribution* as they influence this trade-off. For instance, MNIST's uniform feature distribution helps align individual subgroup and group fairness, whereas FMNIST's variability in textures and styles causes a divergence between the two. Our results show that bias mitigation techniques exhibit varying trends depending on data heterogeneity and training parameters (λ). Careful parameter tuning is key to balancing subgroup and group fairness, with dataset complexity playing a major role in their alignment or divergence across clients.

Takeaway: *Balancing individual subgroup and group fairness requires trade-offs and careful parameter tuning.*

6.5 Privacy Preservation and Its Impact on Fairness and Utility

To evaluate that sensitive client metadata remains protected while allowing for calculating fairness constraints for `LipFed`, we assess the impact of differential privacy on individual subgroup fairness and model performance. We add varying levels of Laplace noise ($\epsilon \in 0.8, 1.0, 1.4$) to local subgroup losses exchanged between clients and the server to compute individual subgroup fairness constraints. The ϵ values range aligns with standard privacy-preserving practices in FL [1].

Table 1. Impact of differential privacy levels on individual subgroup fairness and model utility.

ϵ	MNIST		FMNIST		FER2013		UTK		ACSI		ACSE	
	Sub. Var.	Avg. Var.	Sub. Var.	Avg. Var.	Sub. Var.	Avg. Var.	Sub. Var.	Avg. Var.	Sub. Var.	Avg. Var.	Sub. Var.	Avg. Var.
0.8	0.13	87.00%	0.06	73.49%	0.03	48.43%	0.03	74.89%	0.01	77.53%	0.02	71.50%
1.0	0.13	87.11%	0.06	73.49%	0.03	48.83%	0.03	74.89%	0.01	77.53%	0.02	71.50%
1.4	0.13	87.11%	0.06	72.59%	0.03	48.83%	0.02	74.89%	0.01	77.53%	0.02	71.50%
no-DP	0.12	87.18%	0.06	73.40%	0.03	51.07%	0.02	73.00%	0.01	77.53%	0.02	71.50%

We evaluate the impact of different privacy levels on individual subgroup variance and model utility for benchmark datasets as shown in Table 1. Differential privacy has minimal effect on individual subgroup fairness and utility. For instance, at $\epsilon = 0.8$, MNIST shows a variance of 0.13 and 87% accuracy, while FMNIST shows a 0.06 discrepancy and 73.49% accuracy. These results remain consistent across varying privacy levels and without privacy (no-DP), indicating that privacy does not significantly degrade fairness or performance. `LipFed` 's inherent Lipschitz continuity and subgroup similarity provide natural privacy protection by reducing sensitivity to individual data points without needing explicit noise addition. The mathematical framework in §E can be used to argue that our technique naturally satisfies differential privacy criteria, meaning the technique limits information leakage about individual data points in the dataset to the extent that no single data point significantly alters the statistical characteristics of the output, thereby offering privacy protection as an inherent feature. A detailed theoretical privacy analysis is presented in §E.

Takeaway. *`LipFed` effectively preserves sensitive client information through differential privacy while having only a negligible impact (0.01%) on model accuracy and maintaining stable individual subgroup fairness.*

6.6 Effect of the Fairness Regularization Parameter

In this experiment, we investigate the effect of the fairness regularization parameter λ in Equation 6 on the classifier's utility and individual subgroup fairness. The parameter λ controls the trade-off between the utility of the classifier and its individual subgroup fairness, and tuning this parameter is usually dependent on the network or dataset. To that end, our investigation uses the experimental setup in §3.2 with three different values for $\lambda = \{0.003, 0.1, 0.4\}$. The results are summarized in Figure 8. The results show that the individual subgroup fairness satisfaction can increase without a significant drop in accuracy.

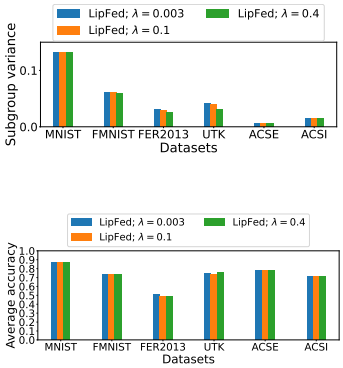


Fig. 8. Impact of varying the fairness regularization parameter λ on the classifier's utility and individual subgroup fairness. Increasing λ improves subgroup fairness with minimal impact on utility.

Takeaway: *Adjusting the λ improves individual subgroup fairness with minimal impact on classifier utility, demonstrating that `LipFed` can balance fairness and utility effectively.*

7 CONCLUSION AND FUTURE WORK

The heterogeneity of statistical features in local data across clients in FL models leads to subgroup bias. To address this, we introduce `LipFed`, a framework leveraging the Lipschitz fairness constraint `LipFed` ensures that similar subgroups have performance outcomes with a statistical distance within their similarity measure, improving subgroup fairness without significantly sacrificing utility, as delineated by our theoretical analysis, which shows a trade-off in group fairness. Our extensive experiments validate `LipFed`'s efficacy in subgroup bias mitigation, demonstrating its superiority over six state-of-the-art bias mitigation techniques and enhancing the fairness of traditional FL methods.

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

[3] Canyu Chen, Yueqing Liang, Xiongxiao Xu, Shangyu Xie, Yuan Hong, and Kai Shu. 2022. On Fair Classification with Mostly Private Sensitive Attributes. *arXiv preprint arXiv:2207.08336* (2022).

[4] R Courtland. 2018. Bias detectives: The researchers striving to make algorithms fair. Na-ture. Retrieved March 15, 2022.

[5] 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.

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

[7] Anamika Dhillon and Gyanendra K Verma. 2020. Convolutional neural network: a review of models, methodologies and applications to object detection. *Progress in Artificial Intelligence* 9, 2 (2020), 85–112.

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

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[10] 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.

[11] 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.

[12] 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.

[13] Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

[14] 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.

[15] Panagiotis Giannopoulos, Isidoros Perikos, and Ioannis Hatzilygeroudis. 2018. Deep learning approaches for facial emotion recognition: A case study on FER-2013. *Advances in hybridization of intelligent methods: Models, systems and applications* (2018), 1–16.

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

[17] 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.

[18] Samuel Horvath, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Lane. 2021. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems* 34 (2021), 12876–12889.

[19] David W Hosmer, Trina Hosmer, Saskia Le Cessie, and Stanley Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine* 16, 9 (1997), 965–980.

[20] 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.

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

[22] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. 2022. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering* 9, 4 (2022), 2039–2051.

[23] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[24] 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.

[25] 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.

[26] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*. PMLR, 2564–2572.

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

[28] Anoop Krishnan, Ali Almadan, and Ajita Rattani. 2020. Understanding fairness of gender classification algorithms across gender-race groups. In *2020 19th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 1028–1035.

[29] 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.

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

[31] Gihun Lee, Minchan Jeong, Yongjin Shin, Sangmin Bae, and Se-Young Yun. 2021. Preservation of the Global Knowledge by Not-True Distillation in Federated Learning. *arXiv preprint arXiv:2106.03097* (2021).

[32] 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.

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

[34] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems* 2 (2020), 429–450.

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

[36] 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).

[37] 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.

[38] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830* (2015).

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

[40] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

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

[42] 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.

[43] 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.

[44] Judea Pearl. 2022. Comment: understanding Simpson's paradox. In *Probabilistic and causal inference: The works of judea Pearl*. 399–412.

[45] Chaim Perelman. 1963. The idea of justice and the problem of argument. (1963).

[46] Vyacheslav Polonski. 2018. AI is convicting criminals and determining jail time, but is it fair. In *World Economic Forum*, Vol. 19.

[47] 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.

[48] pytorch 2019. PyTorch Documentation. https://pytorch.org/.

[49] 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.

[50] Rathindra Sarathy and Krishnamurty Muralidhar. 2011. Evaluating Laplace noise addition to satisfy differential privacy for numeric data. *Trans. Data Priv.* 4, 1 (2011), 1–17.

[51] Andrey V Savchenko. 2021. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 119–124.

[52] Khotso Selialia, Yasra Chandio, and Fatima M Anwar. 2023. Mitigating Group Bias in Federated Learning for Heterogeneous Devices. *arXiv preprint arXiv:2309.07085* (2023).

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

[54] 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.

[55] 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).

[56] Huan Tian, Bo Liu, Tianqing Zhu, Wanlei Zhou, and S Yu Philip. 2024. MultiFair: Model Fairness With Multiple Sensitive Attributes. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[57] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems* 33 (2020), 7611–7623.

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[59] Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. 2021. Federated Learning with Fair Averaging. *CoRR* abs/2104.14937 (2021). https://doi.org/10.24963/IJCAI.2021/223

[60] 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).

[61] Xubo Yue, Maher Nouiehed, and Raed Al Kontar. 2023. Gifair-fl: A framework for group and individual fairness in federated learning. *INFORMS Journal on Data Science* 2, 1 (2023), 10–23.

[62] 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).

[63] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*. PMLR, 962–970.

[64] Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710* (2015).

Appendix

We provide additional information for our paper, `LipFed`: *Mitigating Subgroup Bias in Federated Learning with Lipschitz Constraints*, in the following order:

- Limitations and Future Work (Appendix A)
- Subgroup Fairness (Appendix B)
- Novelty of Lipschitz Constraints (Appendix C)
- Detailed Theoretical Analysis (Appendix E)
- Privacy Analysis (Appendix E)
- More Related Work (Appendix F)
- Experimental Setup (Appendix G)
- Metrics (Appendix H)
- Additional Results (Appendix I)

A LIMITATIONS AND FUTURE WORK

A.1 Limitations

Despite the effectiveness of the `LipFed` framework in mitigating subgroup bias, several limitations remain. Firstly, the reliance on the Lipschitz property to ensure subgroup fairness introduces constraints that may not universally apply across all types of models or datasets. There is a possibility that different models exhibit varying degrees of sensitivity to Lipschitz constraints, which could lead to inconsistent results when applied to non-IID data distributions. Second, the effectiveness of our method is influenced by the proper selection of the hyperparameter ϵ that governs the Lipschitz constraint. Finding the optimal balance between subgroup and group fairness may require extensive tuning and could differ based on the specific characteristics of the datasets being used.

Furthermore, while our approach shows improvements over existing methods, the trade-off between subgroup and group fairness necessitates careful calibration, which may not be straightforward. As subgroup variance decreases, the potential for bias to still emerge in certain groups remains a challenge. Lastly, the additional computational overhead of enforcing Lipschitz constraints during the optimization process may not be feasible for all practical applications, especially in resource-constrained environments.

A.2 Future Work

Further empirical studies are needed to evaluate `LipFed`'s performance in diverse real-world scenarios, including applications beyond image classification, such as text and audio data. It would also be beneficial to investigate our method's scalability in federated learning environments with a large number of clients and significantly diverse data distributions. Moreover, it would be valuable to explore dynamic tuning mechanisms for the hyperparameter ϵ, potentially through adaptive methods that can adjust to the evolving characteristics of the data during the training process. This would facilitate achieving a more nuanced balance between subgroup and group fairness.

B SUBGROUP FAIRNESS

B.1 Federated Learning Subgroup Fairness vs. Centralized Learning Subgroup Fairness

Subgroup fairness in FL differs significantly from centralized learning. In centralized learning, all data is aggregated in one location, making it easier to apply fairness constraints uniformly across subgroups. However, FL operates on decentralized data distributed across multiple clients, where non-IID data distributions pose significant challenges. Achieving subgroup fairness in FL requires ensuring that each client contributes equitably to the global model despite these variations. This decentralized setup demands sophisticated model aggregation techniques to maintain subgroup fairness, as direct access to all client data is not possible.

B.2 Subgroup Fairness vs. Fairness across multiple sensitive attributes

Fairness across multiple sensitive attributes, discussed in MultiFair[56], ensures that fairness constraints are satisfied *for each sensitive attribute individually* (regardless of their number) without necessarily focusing on their intersections. Consider a loan approval algorithm that aims to ensure fairness. The algorithm might be designed to approve loans at the same rate for men and women (gender fairness) and at the same rate for people of different ages (age fairness). Each attribute (gender, age) is treated separately to ensure fairness, but the algorithm might not specifically check if it's fair to, for instance, young women or older men. Subgroup fairness (intersectional attributes focus) and multiple sensitive attributes (individual attribute focus) have some overlap, but they are not closely related. The distinction between these approaches is well-recognized in the literature [26]. In centralized learning, there is a clear separation between ensuring fairness for individual attributes and addressing fairness at the intersection of multiple attributes (subgroup fairness). As noted in the paper [26], the need to ensure fairness across intersectional subgroups is paramount to avoid fairness gerrymandering, where a model appears fair across individual attributes but fails at the intersection of these attributes.

B.3 Additional Causes for Subgroup Fairness

Several factors contribute to subgroup unfairness, one of the most prominent being differences in group sizes. This issue is commonly referred to as *label distribution skew*, where imbalances in the distribution of labels across groups lead to biased outcomes. This challenge has been extensively studied in recent federated learning fairness research [26, 61].

In contrast, our work `LipFed` deliberately focuses on a less explored yet equally important issue: the *same label, different features* phenomenon. This refers to instances where subgroups that share the same label exhibit significantly different feature distributions, leading to unfair treatment across those subgroups. By addressing this underexamined factor, our work provides new insights into the complexities of achieving subgroup fairness in FL.

B.4 Average Variance of Image Pixel Weighting Scheme

Pixel-level variance reflects differences in texture, lighting, and other visual features that affect image data similarity and heterogeneity [64]. By computing subgroup importance weights based on the average variance of image pixels, subgroups with higher pixel variance, indicating less robustness, are prioritized during training to improve model performance [58]. In [27], the authors present a mathematical framework showing how feature variance, such as image pixel variance, influences fairness by affecting loss discrepancy. Here are the relevant equations and their implications in scenarios of binary groups (0 and 1, say):

$$Disc \propto \left|(\Lambda\beta)^\top \Delta\Sigma_z(\Lambda\beta) - (P[g=1] - P[g=0])((\Lambda\beta)^\top \Delta\mu_z)^2\right| \tag{9}$$

where $\Lambda = (\Sigma_z + \Sigma_u)^{-1}\Sigma_u$ is a matrix that balances the variance of the latent features (Σ_z) with the variance of noise in those features (Σ_u), ensuring that features with lower noise are weighted more heavily.

The terms $\Delta\Sigma_z = Var[z \mid g = 1] - Var[z \mid g = 0]$ and $\Delta\mu_z = E[z \mid g = 1] - E[z \mid g = 0]$ represent the difference in the variance and the mean of the latent features between the two groups, $g = 1$ and $g = 0$, respectively. Larger differences in these values signify a greater potential for bias, as one group's feature distribution deviates significantly from the other's. The proportions $P[g = 1]$ and $P[g = 0]$ reflect the relative sizes of the two groups, which influence how much weight the second term in the equation has on the overall discrepancy. The model's learned parameters, β, determine the importance of each latent feature in the prediction process. The interaction between the feature variances and the model parameters, captured by the term $(\Lambda\beta)^\top \Delta\Sigma_z(\Lambda\beta)$, increases as feature variance (Σ_z) increases, indicating that higher variance in features leads to a larger loss discrepancy between groups.

Building on previous studies, we assign higher importance to subgroups with higher variance, which indicates potential model bias. This method aligns with other techniques that prioritize training samples based on characteristics like gradient norm, assessing robustness through feature heterogeneity. This loss discrepancy directly contributes to model bias, suggesting unequal treatment of different groups. Our weighting scheme aims to mitigate this bias by assigning higher importance to subgroups with greater variance. We compare our fairness weighting scheme with GIFAIR-FL, a framework for fairness in FL[61]. GIFAIR-FL uses regularization to penalize variations in client group losses, adapting to statistical differences at each communication round. This approach aligns with our fairness definitions by ensuring equitable performance across data groups.

C NOVELTY OF LIPSCHITZ CONSTRAINTS

While Lipschitz continuity itself is not a novel concept, our work introduces one of the first adaptations of Lipschitz constraints in FL to specifically address subgroup fairness. `LipFed` leverages these constraints to calculate the importance of each subgroup on a client, enabling the model to assign different weights to subgroups based on the variability in their data. This approach helps mitigate the effects of non-IID data by prioritizing subgroups that experience greater bias.

What sets `LipFed` apart is its ability to enforce Lipschitz constraints without requiring access to clients' raw data, preserving privacy—a crucial aspect in federated settings. By focusing on the balance between subgroup fairness and data privacy, `LipFed` offers an innovative solution to address fairness in FL systems without compromising privacy.

D DETAILED THEORETICAL ANALYSIS

D.1 Assumptions

We make the following assumptions on the functions $R_1, \ldots, R_N$ as [36].

Assumption 1. $F_1, \ldots, F_N$ are all L-smooth: for all $\mathbf{v}$ and θ,

$$R_k(\mathbf{v}) \leq R_k(\theta) + \nabla R_k(\theta)^T(\mathbf{v} - \theta) + \frac{L}{2}\|\mathbf{v} - \theta\|^2.$$

Assumption 2. $R_1, \ldots, R_N$ are all μ-strongly convex: for all $\mathbf{v}$ and θ,

$$R_k(\mathbf{v}) \geq R_k(\theta) + \nabla R_k(\theta)^T(\mathbf{v} - \theta) + \frac{\mu}{2}\|\mathbf{v} - \theta\|^2.$$

Assumption 3. Let ξ_k^t be sampled from the k-th device's local data uniformly at random. The variance of stochastic gradients in each device is bounded:

$$\mathbb{E}[\|\nabla R_k(\theta_k^t, \xi_k^t) - \nabla R_k(\theta_k^t)\|^2] \leq \sigma_k^2.$$

Assumption 4. The expected squared norm of stochastic gradients is uniformly bounded, i.e.,

$$\mathbb{E}[\|\nabla R_k(\theta_k^t, \xi_k^t)\|^2] \leq G^2$$

for all $k = 1, \ldots, N$ and $t = 1, \ldots, T-1$. This section presents a theoretical analysis of subgroup and group fairness in ML models. Theorems here establish upper bounds for `LipFed` optimization and explore trade-offs between Lipschitz continuity, empirical risk, and fairness constraints. These theorems provide insights into the relationships between model properties, fairness constraints, and empirical risk outcomes.

Theorem D.1.1. ***Subgroup fairness upper bound.*** *Under* ***Assumption 1-4*** *on the global empirical risk function $R(\boldsymbol{\theta})$ as per recent FL works [35, 36], we have:*

$$\mathbb{E}[R(\boldsymbol{\theta}_T) - R^*] \leq \frac{L}{2}\frac{v}{\gamma + T} + \frac{\kappa}{\gamma + T}\left(\frac{2 \cdot \sum_{k=1}^{K} p_k^2 \sigma_k^2 + 6L\Gamma_1 + 8(E-1)^2 G^2}{\mu} + \frac{\mu(\gamma+1)}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|^2\right), \tag{10}$$

where Γ_1 quantifies the degree of data heterogeneity; if the data are non-iid, then Γ is nonzero and its magnitude reflects the heterogeneity of the data distribution [36]. p_k is the weight of the k-th device such that p_k is proportional the device's local data size and $p_k \geq 0$, E is the number of local training rounds/epochs for each device k, and $\gamma = \max\{8\kappa, E\}$.

Theorem D.1.2. ***Group fairness upper bound.*** *Under* ***Assumption 1-4*** *on the global empirical risk function $R(\boldsymbol{\theta})$ as per recent FL works [35, 36], we have:*

$$\mathbb{E}[R(\boldsymbol{\theta}_T) - R^*] \leq \frac{L}{2}\frac{v}{\gamma + T} + \frac{\kappa}{\gamma + T}\left(\frac{2 \cdot \sum_{k=1}^{K} p_k^2 \sigma_k^2 + 6L\Gamma_2 + 8(E-1)^2 G^2}{\mu} + \frac{\mu(\gamma+1)}{2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}^*\|^2\right), \tag{11}$$

where Γ_2 quantifies the degree of data heterogeneity; if the data are non-iid, then Γ is nonzero and its magnitude reflects the heterogeneity of the data distribution [36]. p_k is the weight of the k-th device such that p_k is proportional the device's local data size and $p_k \geq 0$, E is the number of local training rounds/epochs for each device k, and $\gamma = \max\{8\kappa, E\}$.

D.2 Proof of Theorems D.1.1 and D.1.2

We analyze FedAvg in the setting of full device participation in this section.

D.2.1 Additional Notation. Let $\boldsymbol{\theta_k^t}$ be the model parameter maintained in the k-th device at the t-th step. Let T_E be the set of global synchronization steps, i.e., $T_E = \{nE | n = 1, 2, \ldots\}$. If $t+1 \notin T_E$, i.e., the time step to communication, FedAvg activates all devices. Then the update of FedAvg with partial devices active can be described as:

$$\boldsymbol{v_k^{t+1}} = \boldsymbol{\theta_k^t} - \eta_t \nabla R_k(\boldsymbol{\theta_k^t}, \xi_k^t), \tag{12}$$

$$\boldsymbol{\theta_k^{t+1}} = \begin{cases} \boldsymbol{v_k^{t+1}} & \text{if } t+1 \notin T_E, \\ \frac{\sum_{k=1}^{N} p_k \boldsymbol{v_k^{t+1}}}{\sum_{k=1}^{N} p_k} & \text{if } t+1 \in T_E. \end{cases} \tag{13}$$

Here, an additional variable $\boldsymbol{v_k^{t+1}}$ is introduced to represent the immediate result of one step SGD update from $\boldsymbol{\theta_k^t}$. We interpret $\boldsymbol{\theta_k^{t+1}}$ as the parameter obtained after communication steps (if possible).

D.2.2 Key Lemmas. To convey the proof clearly, it would be necessary to prove certain useful lemmas. We refer the reader to [36] for detailed proofs.

Lemma 1 (Results of one step SGD): Assume Assumption 1 and 2. If $\eta_t \leq \frac{1}{4L}$, we have

$$\mathbb{E}[\|\boldsymbol{v}_k^{t+1} - \boldsymbol{\theta}^*\|^2] \leq (1-\eta_t\mu)\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2] + \eta_t^2\mathbb{E}[\|g_t - \bar{g}_t\|^2] + 6L\eta_t^2\Gamma + 2\eta_t^2\sum_{k=1}^{N} p_k\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_k^t\|^2],$$

where $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^* \geq 0$.

Lemma 2 (Bounding the variance): Assume Assumption 3 holds. It follows that

$$\mathbb{E}[\|g_t - \bar{g}_t\|^2] \leq \sum_{k=1}^{N} p_k^2\sigma_k^2.$$

Lemma 3 (Bounding the divergence of $\{w_k^t\}$): Assume Assumption 4, that η_t is non-increasing and $\eta_t \leq 2\eta_{t+E}$ for all $t \geq 0$. It follows that

$$\mathbb{E}\left[\sum_{k=1}^{N} p_k\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_k^t\|^2\right] \leq 4\eta_t^2(E-1)^2G^2.$$

Proof. It is clear that no matter whether $t+1 \in \mathscr{E}$ or $t+1 \notin \mathscr{E}$, we always have $\bar{\boldsymbol{\theta}_{t+1}} = \bar{\boldsymbol{v}_{t+1}}$. Let $\Delta_t = \mathbb{E}[\|\bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|^2]$. From Lemma 1, Lemma 2 and Lemma 3, it follows that:

$$\Delta_{t+1} \leq (1-\eta_t\mu)\Delta_t + \eta_t^2 B$$

where

$$B = \sum_{k=1}^{N} p_k^2\sigma_k^2 + 6L\Gamma + 8(E-1)^2G^2.$$

For a diminishing stepsize, $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{\mu}$ and $\gamma > 0$ such that $\eta_1 \leq \min\left(\frac{1}{\mu}, \frac{1}{4L}\right) = \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+E}$. We will prove $\Delta_t \leq \frac{v}{t+\gamma}$ where $v = \max\left(\frac{\beta^2 B}{(\beta\mu-1)}, (\gamma+1)\Delta_1\right)$.

We prove it by induction. Firstly, the definition of v ensures that it holds for $t = 1$. Assume the conclusion holds for some t, it follows that:

$$\begin{aligned}\Delta_{t+1} &\leq (1-\eta_t\mu)\Delta_t + \eta_t^2 B \leq \left(1 - \frac{\beta\mu}{t+\gamma}\right)\frac{v}{t+\gamma} + \frac{\beta^2 B}{(t+\gamma)^2} \\ &= \frac{v(t+\gamma-1) + \beta^2 B}{(t+\gamma)^2} \leq \frac{v(t+\gamma) + \beta B - v}{(t+\gamma)^2} = \frac{v}{t+\gamma+1}.\end{aligned}$$

Then by the L-smoothness of $F(\cdot)$,

$$\mathbb{E}[F(W_t)] - F^* \leq \frac{L}{2}\Delta_t \leq \frac{L}{2}\frac{v}{t+\gamma}$$

Specifically, if we choose $\beta = \frac{2}{\mu}$, $\gamma = \max(8L, E) - 1$ and denote $k = \frac{L}{\mu}$, then $\eta_t = \frac{2}{\mu(t+\gamma)}$. One can verify that the choice of η_t satisfies $\eta_t \leq 2\eta_{t+E}$ for $t \geq 1$. Then, we have

$$v = \max\left(\frac{\beta^2 B}{(\beta\mu-1)}, (\gamma+1)\Delta_1\right) \leq \frac{4B}{\mu^2}(\gamma+1)\Delta_1,$$

and

$$\mathbb{E}[F(\bar{\boldsymbol{\theta}}_t)] - F^* \leq \frac{L}{2}\frac{v}{\gamma+t} \leq \frac{L}{2}\frac{v}{\gamma+t} = \frac{\kappa}{\gamma+t}\left(\frac{2B}{\mu} + \frac{\mu(\gamma+1)}{2}\|\bar{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}^*\|^2\right).$$

E PRIVACY ANALYSIS

In this section, we present a detailed mathematical analysis of how differential privacy (DP) is applied in `LipFed` to protect subgroup losses and fairness constraints while maintaining model utility. The goal is to ensure that sensitive data remains private without compromising the ability to mitigate subgroup bias.

E.1 Differential Privacy in `LipFed`

Differential privacy ensures that the inclusion or exclusion of a single data point (or client) does not significantly affect the outcome of the computation, thereby protecting sensitive data. `LipFed` integrates DP by adding *Laplace noise* to the local subgroup losses, ensuring privacy in the exchange of fairness-related metrics between clients and the server.

Definition of Differential Privacy. A randomized algorithm A satisfies ϵ-differential privacy if, for any two adjacent datasets D and D' (differing by only one data point), and for any set S of possible outputs:

$$P(A(D) \in S) \leq e^{\epsilon} \cdot P(A(D') \in S)$$

where ϵ is the *privacy budget*, controlling the amount of noise added and the trade-off between privacy and accuracy.

E.2 Applying Differential Privacy to Subgroup Losses

In `LipFed`, we introduce Laplace noise to the local subgroup losses to maintain privacy. The randomized mechanism for applying DP to subgroup losses is defined as:

$$A(D) = \hat{R}(X_g; \theta) + \text{Laplace}\left(\frac{\Delta R}{\epsilon}\right) \tag{14}$$

where $\hat{R}(X_g; \theta)$ is the true risk or loss function for subgroup X_g; ΔR is the sensitivity of the loss function, measuring the maximum change in output by modifying a single client's data; ϵ is the privacy budget controlling the amount of noise added.

E.3 Sensitivity of Subgroup Losses

The *sensitivity* ΔR of the loss function is the maximum possible difference in the loss function due to the change in one client's data. If $R(X_g; \theta)$ represents the loss for subgroup X_g, then:

$$\Delta R = \max_{D,D'} |R(D; \theta) - R(D'; \theta)| \tag{15}$$

where D and D' are neighboring datasets differing by only one data point.

E.4 Noise Addition and Privacy Guarantee

For each subgroup, we add Laplace noise $\text{Laplace}\left(\frac{\Delta R}{\epsilon}\right)$ to ensure that the differences in the subgroup losses remain indistinguishable. The magnitude of the noise is proportional to the sensitivity ΔR and inversely proportional to ϵ, where larger ϵ implies less noise and weaker privacy guarantees. This ensures that the exchange of sensitive subgroup performance information between the server and clients is protected by differential privacy.

E.5 Impact on Fairness and Utility

The introduction of DP in `LipFed` does not significantly degrade fairness or model utility, as seen in the experimental results. For instance, different privacy budgets $\epsilon \in \{0.8, 1.0, 1.4\}$ only minimally affect subgroup fairness and accuracy.

Theoretical Privacy Bound. `LipFed` ensures that the discrepancy between the loss values of similar subgroups is bounded by ϵ-differential privacy. Given the Lipschitz continuity constraint $D(h_\theta(X), h_\theta(X')) \leq \epsilon \cdot d(X, X')$, we enforce that:

$$|R(X_g; \theta) - R(X'_g; \theta)| \leq \epsilon^2 \cdot \Gamma \tag{16}$$

where Γ measures the heterogeneity in data distribution across clients. This bound ensures that subgroup discrepancies remain within the privacy budget while preserving fairness.

F MORE RELATED WORK

FL algorithms aimed at achieving a globally fair model are typically classified into three distinct categories, including *client-fairness*[6, 18, 22, 33, 35, 41], *group-fairness*[5, 43, 52, 62], and *robustness techniques* [25, 32].

F.1 Client fairness in Federated Learning

Ensuring fairness among clients in FL is vital to counteract biases from non-IID data distributions across devices. Techniques like the Federated Fair Averaging (FedFV)[59] adjust gradient directions and magnitudes to balance model *average performance* based on each client's conflict level and contribution[43]. GIFair-FL [62] dynamically adjusts model updates using a fairness-aware aggregator to reduce *average loss* across clients, while FjORD [18] employs ordered dropout to tailor model sizes to clients' device capacities, enhancing fairness and accuracy.

Additional approaches that build upon these fairness-enhancing techniques include Agnostic Federated Learning (AFL)[41], which optimizes the global model against any potential target distribution by accommodating unknown distribution mixes among clients. q-FFL[35] addresses data heterogeneity by reweighting losses to prioritize devices with poorer performance, promoting uniform model accuracy across devices. Tilted empirical risk minimization (TERM) [33] adjusts the influence of outliers and balances class representation through a flexible tilt hyperparameter. These methods enhance *average performance* in FL systems operating in heterogeneous environments.

F.2 Group Fairness in Federated Learning

Recent advancements in FL have highlighted the importance of addressing fairness concerns, particularly group fairness, where biases against protected demographic groups are mitigated. [11] introduced FairFed, a strategy that ensures fair model training by employing a fairness-aware aggregation method. In FairFed, each client performs local debiasing using their own dataset to maintain decentralization and privacy. Clients evaluate the global model's fairness in each FL round, and aggregation weights are adjusted in collaboration with the server based on the mismatch between global and local fairness metrics. This method, supported by secure aggregation protocols, enhances group fairness under heterogeneous data conditions and allows for client-specific debiasing techniques, showing significant improvement over traditional fairness approaches in FL settings. FairFed's empirical validation confirms its effectiveness in achieving group fairness, with plans for future enhancements to accommodate various application scenarios and integrate broader fairness concepts, such as collaborative and client-based fairness.

In a parallel effort, [43] explore group fairness in FL through their FedMinMax algorithm, which is crafted to establish minimax fairness across demographic groups, an approach that differs from traditional methods aimed at equalizing performance across clients. FedMinMax strategically employs alternating optimization techniques—projected gradient ascent for optimizing weights and stochastic gradient descent for the model—tailoring the learning process to balance fairness among demographic groups effectively. This method has demonstrated competitive or superior

performance against established benchmarks in various FL setups, showcasing its capability to uphold group fairness robustly. Simultaneously, [5] propose the FCFL framework, which addresses both algorithmic fairness and performance consistency across distributed data sources in FL. Derived from a constrained multi-objective optimization perspective, FCFL aims to maximize the utility of the least advantaged client while meeting fairness constraints, achieving Pareto optimality via gradient-based methods. Theoretical and empirical validations of FCFL underscore its ability to outperform existing models in ensuring fairness and consistent performance across clients, making it a viable solution for real-world applications where these attributes are crucial. These developments collectively signal a shift towards more ethical and equitable federated learning environments, emphasizing the necessity for continuous innovation in fairness-oriented methodologies within the field.

Table 2. Partitioning of datasets with added Gaussian noise

Client	Samples	Noise STD	Test Data	Samples	Noise STD	Test Data	Samples	Noise STD	Test Data
	MNIST			**FashionMNIST**			**FER2013**		
1	6,000	0.4	Original + 0.4	6,000	0.4	Original + 0.4	2870	0.0	Original + 0.0
2	6,000	0.5	Original + 0.5	6,000	0.5	Original + 0.5	2870	0.09	Original + 0.09
3	6,000	0.7	Original + 0.7	6,000	0.7	Original + 0.7	2870	0.18	Original + 0.18
4	6,000	1.0	Original + 1.0	6,000	1.0	Original + 1.0	2870	0.27	Original + 0.27
5	6,000	1.5	Original + 1.5	6,000	1.5	Original + 1.5	2870	0.36	Original + 0.36
6	6,000	0.4	Original + 0.4	6,000	0.4	Original + 0.4	2870	0.0	Original + 0.0
7	6,000	0.5	Original + 0.5	6,000	0.5	Original + 0.5	2870	0.09	Original + 0.09
8	6,000	0.7	Original + 0.7	6,000	0.7	Original + 0.7	2870	0.18	Original + 0.18
9	6,000	1.0	Original + 1.0	6,000	1.0	Original + 1.0	2870	0.27	Original + 0.27
10	6,000	1.5	Original + 1.5	6,000	1.5	Original + 1.5	2870	0.36	Original + 0.36
	UTK			**ACS Income (ASCI)**			**ACS Employment (ASCE)**		
1	1920	0.0	Original + 0.0	26621	-	State test	6656	-	State test
2	1920	0.1	Original + 0.1	11143	-	State test	2768	-	State test
3	1920	0.3	Original + 0.3	156532	-	State test	39133	-	State test
4	1920	0.5	Original + 0.5	32091	-	State test	8023	-	State test
5	1920	0.7	Original + 0.7	41653	-	State test	10414	-	State test
6	1920	0.0	Original + 0.0	108739	-	State test	27185	-	State test
7	1920	0.1	Original + 0.1	13069	-	State test	3268	-	State test
8	1920	0.3	Original + 0.3	12645	-	State test	3162	-	State test
9	1920	0.5	Original + 0.5	17604	-	State test	4402	-	State test
10	1920	0.7	Original + 0.7	17814	-	State test	4454	-	State test

G EXPERIMENTAL SETUP

G.1 Dataset Details

Choice of Datasets. In our experiments, we evaluated the `LipFed` framework using four small datasets, including MNIST, Fashion-MNIST, FER2013, and UTK, and two large scale dataset, including ASCI and ASCE, with a 10 clients. These datasets were chosen to represent a diverse set of applications, thereby providing a comprehensive evaluation

of the feasibility and initial effectiveness of the proposed subgroup fairness technique. Each dataset presents unique characteristics and challenges related to bias studies.

The MNIST dataset consists of handwritten digit images. This dataset is often used as a benchmark for image classification tasks and serves as a starting point for evaluating model performance on simple, grayscale images. It helps in understanding basic biases that might arise from digit shapes and writing styles. Fashion-MNIST is a dataset of grayscale images of clothing items. This dataset is used to test model performance on more complex visual patterns compared to MNIST. It introduces variability in clothing styles, textures, and shapes, which can help identify biases related to visual feature extraction and classification. The FER2013 dataset contains grayscale images of facial expressions. This dataset is crucial for studying biases related to facial recognition and emotion detection. It includes images with diverse facial expressions and varying degrees of emotion intensity, which can reveal biases in recognizing and classifying emotional states, especially across different demographic groups. The UTKFace dataset includes images of faces with annotations for age, gender, and ethnicity. This dataset is particularly valuable for studying intersectional biases involving age, gender, and ethnicity. It allows for an in-depth analysis of how different demographic attributes can impact model performance and fairness, revealing potential biases in facial recognition systems across diverse population groups. Despite the aforementioned datasets, we recognize the importance of assessing the model's scalability and robustness on larger datasets, we perform further evaluations on large real-world datasets used in fairness studies (ACSI and ACSE).

Data Partitions. As it is customary to partition benchmark datasets across clients in FL research [21, 57], we adopt this strategy and distribute samples of an individual group equally across clients according to the Dirichlet distribution [21]. This distribution is demonstrated in Table 2, where the third column shows that distributing samples of an individual group equally across clients leads to clients with the equal number of samples in their local data $\mathscr{D}_k$. The uniform data partitioning strategy is motivated by the desire to demonstrate that even in FL settings with balanced groups across clients, feature noise heterogeneity still leads to subgroup bias across clients.

Heterogeneous Feature Distributions. We introduce feature noise across data partitions to simulate real-world scenarios where images are non-IID, deviating from the feature distribution of pristine training images [14, 49, 53]. The noise is added to an image by adding a random value sampled from a Gaussian distribution to each pixel of the image. Mathematically, this is represented as:

$$\tilde{I}(x, y) = I(x, y) + \epsilon \tag{17}$$

where $\epsilon \sim \mathscr{N}(0, \sigma^2)$, with $\tilde{I}(x, y)$ and $I(x, y)$ denoting noisy and original pixel values at (x, y), respectively. The parameter σ controls the amount of noise added to the image. The larger the value of σ, the more intense the noise. Specifically, Gaussian noise with σ of 0.03 or higher is incorporated, reflecting conditions observed in real-world deployments [39]. The noise addition to each client's local training dataset $\mathscr{D}_k$ is demonstrated in Table 2, where the fourth column shows all local datasets across different clients have different feature noise distributions. The difference in feature noise across clients is motivated by the desire to understand how the nonIID-ness in subgroup data of an individual group affects the global model's bias across subgroups.

Local Test Data. Each client utilizes a replicated version of the original benchmark test set, aligning similar noise feature distributions between the training and test data for individual clients. For example, as depicted in Table 2, client 1 employs the original FMNIST test dataset with noise levels consistent with those of the training partition. This approach is motivated by the assumption that the local and training data for each client share similar feature distributions, which may differ from those of other clients.

G.2 Training Parameters

Table 3 outlines the primary training parameters used across all models and datasets in this work. We implemented the system using PyTorch [48] on Ubuntu 22.04 (8GB NVIDIA Quadro P2200 GPU). [1]

Table 3. Model Training Parameters.

Algorithm	Dataset	Train time per round (minutes)	Model	Minibatch size	Momentum	Weight decay	Learning rate	# Local epochs	# Rounds	Loss function
FedAvg	MNIST	2.25	LeNet	256	0.9	0.0001	0.01	5	65	Cross entropy
	Fashion-MNIST	2.23	VGGNet	256	0.9	0.0005	0.01	5	65	Cross entropy
	FER2013	8.93	ResNet-18	128	0.9	0.0005	0.01	5	30	Cross entropy
	UTK	4.98	ResNet-18	64	0.9	0.0005	0.01	5	75	Cross entropy
	ACSIncome	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
	ACSEmpoyment	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
AFL	MNIST	2.22	LeNet	256	0.9	0.0001	0.01	5	65	Cross entropy
	Fashion-MNIST	2.25	VGGNet	256	0.9	0.0005	0.01	5	65	Cross entropy
	FER2013	8.76	ResNet-18	128	0.9	0.0005	0.01	5	30	Cross entropy
	UTK	4.94	ResNet-18	64	0.9	0.0005	0.01	5	75	Cross entropy
	ACSIncome	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
	ACSEmpoyment	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
TERM	MNIST	2.27	LeNet	256	0.9	0.0001	0.01	5	65	Cross entropy
	Fashion-MNIST	2.28	VGGNet	256	0.9	0.0005	0.01	5	65	Cross entropy
	FER2013	9.15	ResNet-18	128	0.9	0.0005	0.01	5	30	Cross entropy
	UTK	4.99	ResNet-18	64	0.9	0.0005	0.01	5	75	Cross entropy
	ACSIncome	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
	ACSEmpoyment	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
GIFAIR-FL	MNIST	2.05	LeNet	256	0.9	0.0001	0.01	5	65	Cross entropy
	Fashion-MNIST	1.98	VGGNet	256	0.9	0.0005	0.01	5	65	Cross entropy
	FER2013	8.27	ResNet-18	128	0.9	0.0005	0.01	5	30	Cross entropy
	UTK	4.51	ResNet-18	64	0.9	0.0005	0.01	5	75	Cross entropy
	ACSIncome	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
	ACSEmpoyment	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
FedNTD	MNIST	2.53	LeNet	256	0.9	0.0001	0.01	5	65	Cross entropy
	Fashion-MNIST	2.57	VGGNet	256	0.9	0.0005	0.01	5	65	Cross entropy
	FER2013	10.23	ResNet-18	128	0.9	0.0005	0.01	5	30	Cross entropy
	UTK	5.62	ResNet-18	64	0.9	0.0005	0.01	5	75	Cross entropy
	ACSIncome	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
	ACSEmpoyment	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
Scaffold	MNIST	0.71	LeNet	256	0.9	0.0001	0.01	5	65	Cross entropy
	Fashion-MNIST	0.82	VGGNet	256	0.9	0.0005	0.01	5	65	Cross entropy
	FER2013	2.29	ResNet-18	128	0.9	0.0005	0.01	5	30	Cross entropy
	UTK	1.63	ResNet-18	64	0.9	0.0005	0.01	5	75	Cross entropy
	ACSIncome	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
	ACSEmpoyment	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
`LipFed`	MNIST	2.61	LeNet	256	0.9	0.0001	0.01	5	65	Cross entropy
	Fashion-MNIST	5.14	VGGNet	256	0.9	0.0005	0.01	5	65	Cross entropy
	FER2013	9.94	ResNet-18	128	0.9	0.0005	0.01	5	30	Cross entropy
	UTK	5.14	ResNet-18	64	0.9	0.0005	0.01	5	75	Cross entropy
	ACSIncome	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy
	ACSEmpoyment	-	Logistic R.	128	-	-	0.001	5	10	Binary Cross entropy

[1]The 'readme.txt' file at the root of the project folder consists of the steps required to run the code: Download Zipped Folder

G.3 Adaptation to tabular datasets

Our approach of using the average variance of image pixels is directly applicable to tabular data. We first present a detailed methodology for adapting `LipFed` for two tabular datasets from fair ML Retiring Adult datasets [8], ACSIncome and ACSEmployment:

The steps to compute subgroup weights/moments (e.g., variance) for subgroups are as follows:

(1) Data Separation: Divide data into subgroups based on intersecting attributes (income>50K, demographics).
(2) Variance Calculation: Calculate variance (subgroup weight) for each subgroup: $\sigma_g^2 = \frac{1}{N_g} \sum_{i=1}^{N_g} (x_i - \mu_g)^2$.

Here, N_g is the number of samples in subgroup g, x_i are the feature values, μ_g is the mean of the feature for subgroup g, and σ_g is the standard deviation of the feature for subgroup g. We use the ACS PUMS [8] as the basis for both prediction tasks income and employment:

Example: ACSIncome Prediction. We use ACS PUMS data to gather income-related features, race, and state information, ensuring each data point includes the state it belongs to. Data is distributed across clients based on the state attribute (randomly selected USA states), with each client representing data from a specific state. We define two groups:

(1) Income True: Individuals with income above a certain threshold (e.g., $50,000).
(2) Income False: Individuals with income below this threshold.

The state serves as an *implicit sensitive attribute* due to its correlation with demographic distribution, forming subgroups by income level and demographic region (Income True and California).

Example: ACSEmployment Prediction. For ACSEmployment, the task is to predict whether an individual is employed after filtering ACS PUMS data to include individuals between the ages of 16 and 90 [8]; We define following groups:

(1) Employed: Individuals who are currently employed.
(2) Unemployed: Individuals who are not employed.

The steps to compute subgroup weights for this dataset are similar: Divide data into subgroups based on employment status and demographic attributes (e.g., employed and from California). Compute variance for each subgroup as described earlier, allowing us to weigh the subgroups' importance and enforce subgroup fairness in `LipFed` optimization .

This methodology illustrates the adaptability of the `LipFed` framework to diverse data types, emphasizing its utility in addressing fairness across multiple domains.

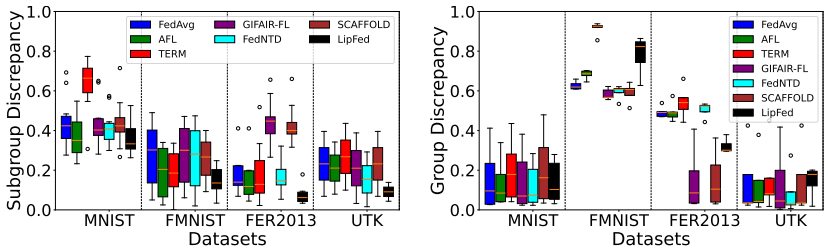
H METRICS

H.1 True Positive Rate (TPR)

The True Positive Rate (TPR) is a critical metric for assessing model performance, as it measures the proportion of actual positives correctly predicted by the model. Variations in TPR across subgroups indicate discrepancies in the model's generalization across different subpopulations. In FL, TPR variation is often a result of non-IID data across clients. Subgroups with diverse characteristics—such as demographic differences, sensor quality, or geographical factors—lead to varied feature distributions, causing differential model performance. Mathematically, TPR is defined as:

$$TPR_g = \frac{TP_g}{TP_g + FN_g} \tag{18}$$

where TP_g and FN_g represent the true positives and false negatives for subgroup g, respectively.



(a) TPR variations among subgroups. (b) TPR variations among groups.

Fig. 9. Demonstrating subgroup bias in model performance for different datasets and baselines.

H.1.1 Variation in TPR Across Subgroups. The variation in TPR can quantify the performance discrepancies between subgroups. Let the TPR for each subgroup g be denoted as TPR_g. The difference between the highest can measure the discrepancy in performance among subgroups- and lowest-performing subgroups:

$$Disc(TPR) = \max_g(TPR_g) - \min_g(TPR_g) \tag{19}$$

A large discrepancy suggests that some subgroups benefit more from the model than others, highlighting the presence of subgroup bias. In non-IID FL settings, subgroup g on one client may have very different feature distributions compared to the same subgroup on another client, leading to inconsistent TPRs across subgroups.

In our `LipFed` framework, which applies Lipschitz constraints to reduce subgroup bias, the goal is to minimize the performance discrepancy across subgroups. The performance difference is constrained by a Lipschitz continuity condition that controls how much the TPR can vary based on subgroup similarity. This condition ensures that:

$$D(h_\theta(x), h_\theta(x')) \leq \epsilon \cdot d(x, x') \tag{20}$$

where $D(h_\theta(x), h_\theta(x'))$ represents the Euclidean distance between the model's outputs for two subgroup instances x and x', and $d(x, x')$ is a distance metric quantifying the similarity between the subgroups. Thus, variations in TPR among subgroups are restricted by the parameter ϵ, which limits the magnitude of subgroup performance differences:

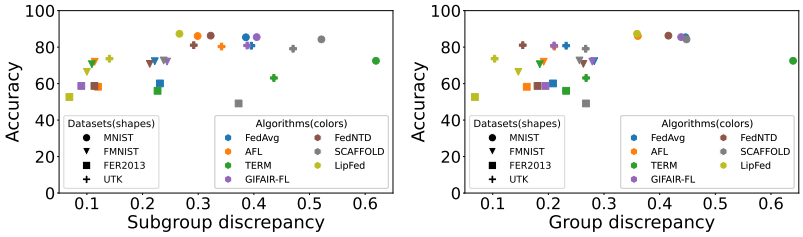
$$Disc(TPR) \leq \epsilon \tag{21}$$

By enforcing these Lipschitz constraints, `LipFed` reduces the subgroup performance disparity, resulting in more equitable TPRs across clients.

I ADDITIONAL RESULTS

I.1 LipFed's Performance Against Consistency and Robustness Benchmarks

In addition to fairness benchmarks (AFL, TERM, GIFAIR), we compare LipFed, our fairness-focused technique, against FL algorithms such as Scaffold [25] and FedNDT [32], which prioritize robustness and consistency over fairness. In the *FL heterogeneity category*, FedNTD addresses performance loss due to data heterogeneity by managing global model memory. In the *FL robustness category*, SCAFFOLD [25] focuses on enhancing resilience against outliers and noisy data, mitigating the impact of irregularities in local datasets. Scaffold addresses client drift, while FedNDT targets model discrepancies due to non-IID data. This evaluation measures LipFed's performance in reducing subgroup bias and maintaining model utility compared to these non-fairness benchmarks.



(a) Utility vs. Subgroup fairness. (b) Utility vs. Group fairness.

Fig. 10. Demonstrating model utility vs. discrepancy for different datasets and baselines.

The results indicate that while Scaffold and FedNDT exhibit strong robustness and consistency across various datasets, LipFed outperforms both in terms of reducing subgroup bias, as shown in Figure 9. For example, in the MNIST dataset, LipFed achieves a 20% lower subgroup bias than Scaffold, demonstrating its effectiveness in mitigating bias without sacrificing much performance. Importantly, although LipFed is designed to mitigate subgroup bias, it also improves group fairness, showing reductions in group discrepancy similar to those seen in subgroup fairness. This demonstrates that LipFed's benefits extend beyond subgroup bias mitigation.

Additionally, in Figure 10, we show that LipFed maintains competitive performance across all datasets, with trends in utility closely mirroring those of the robustness-focused methods. While Scaffold and FedNDT slightly outperform LipFed in raw performance metrics, the trade-off is minimal, showcasing that LipFed effectively balances both fairness and performance across diverse data conditions. This comparison highlights that while methods like Scaffold and FedNDT excel in providing robustness, LipFed offers a balanced solution by significantly reducing subgroup bias while still maintaining strong performance across diverse datasets.

I.2 Convergence Analysis

We evaluate the convergence behavior of LipFed in comparison to baseline techniques such as AFL, TERM, and GIFAIR. The goal is to assess how quickly the training process reduces subgroup discrepancies across multiple datasets, including MNIST, Fashion-MNIST, FER2013, and UTK. Convergence here refers to the stability and speed at which the subgroup discrepancy is minimized during the training process.

As shown in Figure 11, LipFed consistently demonstrates faster convergence and lower subgroup discrepancy across all datasets. This rapid reduction in subgroup bias is primarily due to the Lipschitz continuity constraints imposed by LipFed, which ensure that performance differences between subgroups are bounded early in the training process. In contrast, the baseline techniques either converge more slowly or stabilize at higher subgroup discrepancy values, highlighting their inability to efficiently address subgroup fairness in non-IID settings. For example, on the FER2013 dataset, LipFed achieves a 30% reduction in subgroup discrepancy within the first 50 iterations compared to AFL, which converges much slower. Similarly, on the UTK dataset, LipFed stabilizes subgroup fairness more effectively than other methods, reaching a lower discrepancy in fewer iterations. This consistent performance across datasets illustrates LipFed's efficiency in addressing fairness concerns in federated learning environments with heterogeneous client data. In summary, LipFed's convergence behavior demonstrates its ability to quickly and efficiently reduce subgroup discrepancies, outperforming other fairness-focused techniques in both speed and effectiveness.

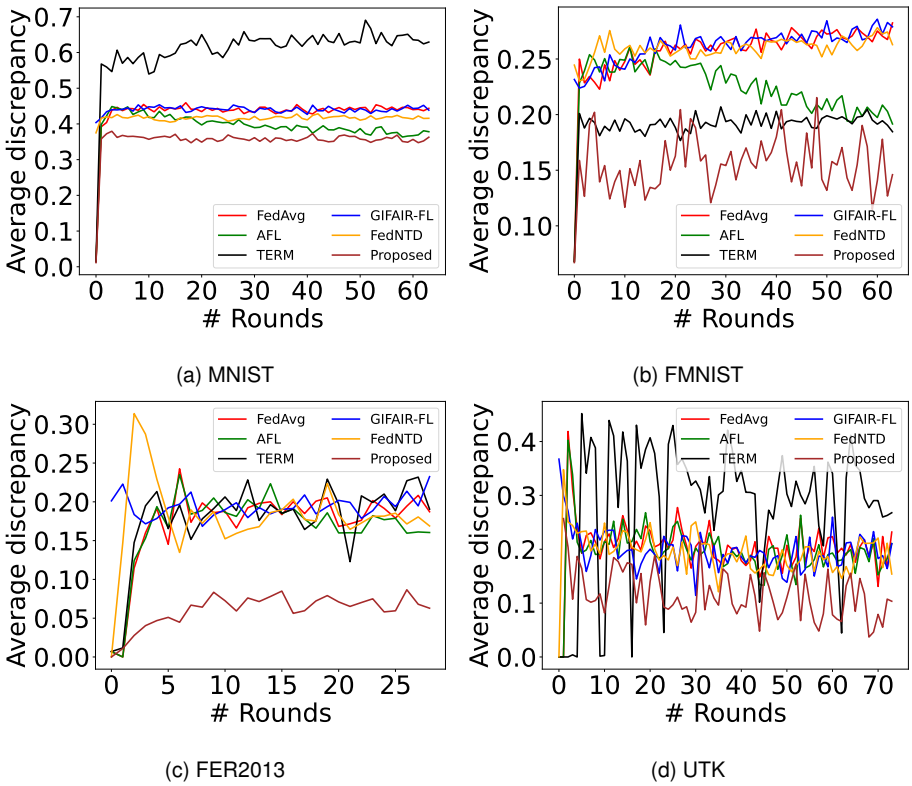


Fig. 11. Convergence of the training subgroup discrepancy of LipFed and other baseline techniques across multiple datasets. LipFed consistently exhibits lower subgroup discrepancy across all iterations.