From Optimization to Generalization: Fair Federated Learning against Quality Shift via Inter-Client Sharpness Matching

Nannan Wu , **Zhuo Kuang** , **Zengqiang Yan**[*] and **Li Yu**

School of Electronic Information and Communications, Huazhong University of Science and Technology

{wnn2000, kuangzhuo, z_yan, hustlyu}@hust.edu.cn

Abstract

Due to escalating privacy concerns, federated learning has been recognized as a vital approach for training deep neural networks with decentralized medical data. In practice, it is challenging to ensure consistent imaging quality across various institutions, often attributed to equipment malfunctions affecting a minority of clients. This imbalance in image quality can cause the federated model to develop an inherent bias towards higher-quality images, thus posing a severe fairness issue. In this study, we pioneer the identification and formulation of this new fairness challenge within the context of the imaging quality shift. Traditional methods for promoting fairness in federated learning predominantly focus on balancing empirical risks across diverse client distributions. This strategy primarily facilitates fair optimization across different training data distributions, yet neglects the crucial aspect of generalization. To address this, we introduce a solution termed Federated learning with Inter-client Sharpness Matching (FedISM). FedISM enhances both local training and global aggregation by incorporating sharpness-awareness, aiming to harmonize the sharpness levels across clients for fair generalization. Our empirical evaluations, conducted using the widely-used ICH and ISIC 2019 datasets, establish FedISM's superiority over current state-of-the-art federated learning methods in promoting fairness. Code is available at https://github.com/wnn2000/FFL4MIA.

1 Introduction

In light of escalating concerns regarding data privacy, federated learning (FL) [McMahan *et al.*, 2017] has emerged as a promising approach for training deep neural networks in the realm of medical image analysis [Dou *et al.*, 2021]. A significant challenge within FL is the inherent data heterogeneity [Ye *et al.*, 2023; Huang *et al.*, 2022; Huang *et al.*, 2023a] observed across various medical institutions, primarily attributed to their independent data collection processes.
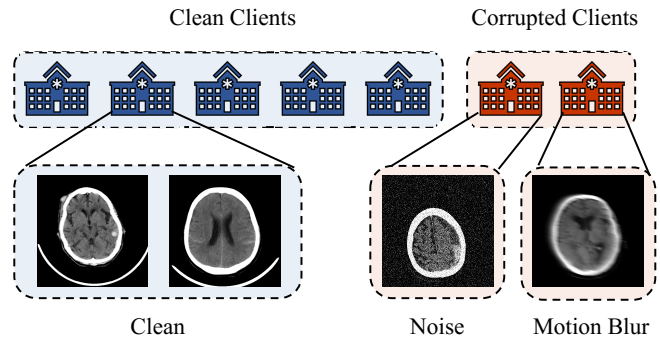
---

[*]Corresponding author

Figure 1: Imaging quality shift across clients. Most clients possess clean images, while others have corrupted images (*e.g.*, exhibiting noise or blur).

This heterogeneity has been examined from several perspectives in existing research, *e.g.*, domain shift [Li *et al.*, 2021; Liu *et al.*, 2021; Jiang *et al.*, 2023b], label skew [Zhang *et al.*, 2022; Wu *et al.*, 2023b], and label quality variation [Wu *et al.*, 2023a; Chen *et al.*, 2023; Wu *et al.*, 2024]. Nevertheless, the prevalent issue of quality heterogeneity [Fang *et al.*, 2023] in medical imaging, a factor that could potentially raise new challenges for FL, remains under-explored.

Despite relatively rigorous protocols in medical imaging environments, uniformity in image quality across various institutions cannot be strictly assured. As illustrated in Fig. 1, images from some clients may exhibit noise due to equipment malfunctions or the necessity for low-dose imaging. Moreover, random movements from either patients or cameras in certain cases can result in motion blur. Typically, the proportion of low-quality (*i.e.*, corrupted) images is smaller compared to high-quality (*i.e.*, clean) images [Huang *et al.*, 2023d]. In such scenarios, characterized by quality shifts across clients, FL tends to exhibit a bias towards the more-prevalent clean images, thereby compromising the performance on the less-frequent corrupted images. Such a bias can be a critical concern when applying federated models to clients with corrupted images. Therefore, it is imperative to address the issue of biased performance under quality shifts.

In this paper, we pioneer the identification and formulation of this significant challenge in FL within real-world medical contexts. In FL, especially under the mantle of privacy
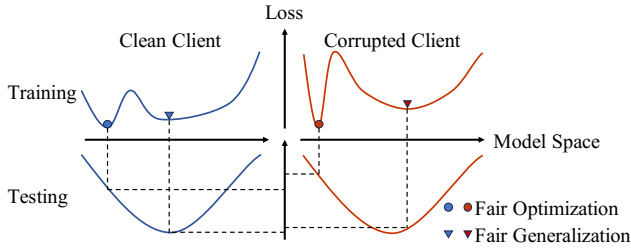
Figure 2: Motivations behind previous fair optimization and our fair generalization. Fair optimization aims for uniform and low loss values across clients, often resulting in convergence at sharp minima and poor testing performance. Comparatively, we focus on achieving uniform sharpness and converging at flat minima, thus enhancing fair generalization and testing performance.

protection, there is an absence of prior knowledge about the imaging quality of each client, suggesting the necessity for a data-agnostic solution to this issue. We theoretically demonstrate improving performance on the poorest-quality images is equivalent to achieving client-level fairness in FL. Hence, we remodel this challenge as a problem of client-level fairness, *i.e.*, ***how to ensure equitable FL performance among clients with clean and corrupted image distributions?*** This investigation marks the first effort to promote fair FL across clients with image quality shifts, a departure from previous research focused on fairness under domain shifts [Jiang *et al.*, 2023a] or class distribution shifts [Li *et al.*, 2020].

In this work, we observe that existing methods to foster client-level performance fairness in FL typically modify the importance weights in global aggregation to harmonize certain training metrics (*e.g.*, loss) across clients [Mohri *et al.*, 2019; Li *et al.*, 2020; Jiang *et al.*, 2023a]. However, as illustrated in Fig. 2, while these methods facilitate fairness regarding optimization during training, they do not necessarily ensure fairness regarding generalization during testing. This gap stems from the issue that equalizing training losses for all clients may result in a simplistic convergence towards a sharp minimum, particularly for the minority distributions (*i.e.*, corrupted image distributions), thereby impairing generalization on testing sets. To counter this, ***our focus extends beyond fair optimization to fair generalization***. Recent advancements in sharpness-aware minimization reveal an inverse relationship between the generalization capability and the sharpness of the loss surface [Foret *et al.*, 2021; Zhuang *et al.*, 2022]. Inspired by this finding, we introduce a novel **Fed**erated learning framework with **I**nter-client **S**harpness **M**atching (**FedISM**), which aims to equalize the sharpness levels across clients for fair generalization. In this framework, local optimization in FL is made sharpness-aware, where each client's local update aims to minimize sharpness on its local data. Subsequently, weights for global aggregation are determined based on each client's sharpness level, with higher weights assigned to clients exhibiting greater sharpness. In this way, the global model update is more effective in minimizing sharpness in higher-level clients, leading to more uniform sharpness and fair generalization across both clean and corrupted image distributions as

shown in Fig. 2. It is worth mentioning that FedISM involves only simple modifications to local optimization and global aggregation compared to FedAvg [McMahan *et al.*, 2017], thus eschewing complex loss functions and additional regularization and offering ease of implementation.

The contributions are summarized into three folds:

- **New Fairness Challenge in FL - Imaging Quality Shift:** We identify and articulate a new challenge in FL, specifically concentrating on achieving performance fairness across clients with diverse imaging qualities.

- **Innovative Solution - FedISM:** To address this challenge, our focus extends beyond previous fair optimization to fair generalization. Our proposed solution, FedISM, integrates sharpness-aware local updates with sharpness-dependent global aggregation, promoting uniform sharpness across clients and achieving more equitable generalization.

- **Extensive Validation:** We validate the effectiveness of FedISM through a series of experiments on two real-world medical image classification datasets, *i.e.*, RSNA ICH and ISIC 2019. FedISM demonstrates superior performance, outperforming several state-of-the-art methods in promoting fairness in FL.

2 Related Work

2.1 Fair Federated Learning

Fairness [Huang *et al.*, 2023c] has become a crucial topic in FL, primarily concentrating on collaborative fairness [Lyu *et al.*, 2020; Xu *et al.*, 2021] and performance fairness [Li *et al.*, 2020]. Collaborative fairness aims for the reward of each participant to be proportional to its contribution to the federation. In contrast, performance fairness advocates for unbiased performance across different devices/attributes/distributions. Existing solutions mainly achieve this goal by balancing weights of different objects to optimize fairly. For instance, q-FedAvg [Li *et al.*, 2020] addresses this by prioritizing clients that are harder to optimize; FedCE [Jiang *et al.*, 2023a] tackles this issue by considering task-specific performance. Nevertheless, fairness respecting generalization has not been considered thoroughly till now, leading to sub-optimal outcomes.

2.2 Sharpness of Loss Surface

Bridging the gap between training optimization and testing generalization remains a pivotal challenge in machine learning. Recent developments in sharpness-aware minimization suggest that models tend to generalize better on flat minima than on sharp minima [Foret *et al.*, 2021; Zhuang *et al.*, 2022]. Motivated by this insight, various studies have incorporated the concept of loss surface sharpness to address poor generalization. SharpDRO [Huang *et al.*, 2023d] merges sharpness with GroupDRO [Sagawa *et al.*, 2020] to achieve robust generalization. ImbSAM [Zhou *et al.*, 2023] focuses on enhancing the performance for tail classes in long-tailed recognition by minimizing sharpness in these classes. In addition to typical machine learning, it has recently been applied in FL [Qu *et al.*, 2022; Caldarola *et al.*, 2022; Sun *et al.*, 2023]. However, the relationship between sharpness and fairness within FL has not yet been investigated.

3 Methodology

3.1 Preliminaries

For a typical image classification task with C classes, we consider a cross-silo FL scenario with K participants. Each k-th participant possesses a private dataset $D_k = \{(\boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y})\}_{i=1}^{N_k}, k \in [K]$, where $\mathcal{X}$ and $\mathcal{Y} = [C]$ represent the input image and label spaces, respectively. An image-label pair $(\boldsymbol{x}_i, y_i)$ is drawn from the client-specific distribution $\mathbb{P}_k(\boldsymbol{x}, y \mid a_k)$, with $a_k \in [A]$ indicating an attribute only influencing image quality in client k. This paper highlights the uneven distribution of clients across various quality attributes. Define $f(\cdot; \boldsymbol{\theta}) : \mathcal{X} \rightarrow \Delta^{C-1}$ as a deep learning model parameterized by $\boldsymbol{\theta}$ and $\ell : \Delta^{C-1} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ as the loss function, where Δ denotes the probability simplex. In this paper, the goal of FL is to optimize $\boldsymbol{\theta}$ for performance enhancement on images with the worst-performing quality:

$$\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta}} \{\max_{a \in [A]} \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathbb{P}(\boldsymbol{x}, y|a)}[\ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y)]\}. \quad (1)$$

When treating each quality as a group, Eq. 1 achieves group fairness which is similar to the aim in distributionally robust optimization [Sagawa *et al.*, 2020]. However, in FL, there is no prior knowledge of group information (*i.e.*, the imaging quality in each client), making group-wise design impractical. Therefore, we achieve Eq. 1 via client fairness, as stated in the following theorem following [Papadaki *et al.*, 2022]:

Theorem 1 (Equivalence). *Assuming class distributions of the testing set and all clients' training sets are identical, we have:*

$$\boldsymbol{\theta}^*, \boldsymbol{\lambda}^* = \arg \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \in \Delta^{K-1}} \sum_{k=1}^{K} \lambda_k \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathbb{P}_k(\boldsymbol{x}, y|a_k)}[\ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y)], \quad (2)$$

and

$$\boldsymbol{\theta}^*, \boldsymbol{\mu}^* = \arg \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\mu} \in \Delta^{A-1}} \sum_{u=1}^{A} \mu_u \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathbb{P}(\boldsymbol{x}, y|u)}[\ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y)], \quad (3)$$

where $\mu_u^* = \sum_{k=1}^{K} \mathbb{1}_{a_k=u} \lambda_k^*$.

The proof is detailed in the Appendix. Notably, even when label distribution shifts occur, they can be theoretically addressed by logit adjustment [Menon *et al.*, 2021; Zhang *et al.*, 2022], thereby not considered to be addressed in this paper. Theorem 1 demonstrates that achieving client fairness (Eq. 2) inherently ensures group fairness (Eq. 3). Therefore, this paper subsequently concentrates on enhancing fairness across clients with various imaging qualities.

3.2 Previous Solution: Fair Optimization

Client-level unfairness in FL arises from a tendency to overlook certain clients, particularly those with limited data or outlier distributions. These clients are often overlooked because optimizing them does not significantly contribute to the overall optimization objective. To address this challenge, various approaches have been proposed. AFL [Mohri *et al.*, 2019] concentrates efforts on the worst-performing client; q-FedAvg [Li *et al.*, 2020] and FairFed [Ezzeldin *et al.*, 2023] give greater weights to those clients with higher training losses; and FedCE [Jiang *et al.*, 2023a] focuses on clients

with lower task-specific metrics. These methods collectively aim to balance optimization across the empirical distribution of each client, striving for more uniform risks among clients, as represented by the following optimization problem:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \in \Delta^{K-1}} \sum_{k=1}^{K} \frac{\lambda_k}{N_k} \sum_{(\boldsymbol{x}, y) \in D_k} \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y). \quad (4)$$

Essentially, these solutions seek to achieve a balance in optimization by prioritizing clients that are performing worse. However, this strategy may result in clients with poorer performance rapidly converging to sharp minima, primarily because the rate of loss minimization is most pronounced when moving towards them. Due to the discrepancy between empirical and expected risks, this strategy to optimize fairly does not necessarily lead to client fairness in a strict sense, as denoted by Eq. 2. A conceptual illustration of this issue is presented in Fig. 2. Thus, it becomes crucial to extend the exploration of client fairness to include considerations of generalization, beyond mere optimization.

3.3 Measurement of Generalization: Sharpness

Since it is sub-optimal to measure generalization capacity by a single value of empirical risk (*i.e.*, training loss), promoting fairness by merely seeking uniformity in such a metric is inappropriate. To tackle this problem, we should identify a new indicator more closely correlated with generalization ability.

A key limitation of using single training loss as a metric is its insensitivity to the geometric properties of the loss landscape, treating sharp and flat minima indiscriminately. To overcome this, we propose focusing on a range of loss values, specifically the sharpness of the loss surface [Foret *et al.*, 2021; Zhuang *et al.*, 2022]. It is defined as the largest loss change in the vicinity of the initial model parameters:

$$\mathcal{S} := \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \{\ell(f(\boldsymbol{x}; \boldsymbol{\theta} + \boldsymbol{\epsilon}), y) - \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y)\}, \quad (5)$$

where ρ is a positive step parameter controlling the search radius. Calculating sharpness as per Eq. 5 poses a challenge due to the continuous and infinite nature of the perturbation. To simplify, the difference can be approximated linearly through the Taylor series when ρ is sufficiently small:

$$\ell(f(\boldsymbol{x}; \boldsymbol{\theta} + \boldsymbol{\epsilon}), y) - \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y) \approx \boldsymbol{\epsilon}^\top \nabla \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y). \quad (6)$$

This approximation enables us to identify the optimal perturbation by maximizing the right-hand side of the equation:

$$\boldsymbol{\epsilon}^* = \arg \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \boldsymbol{\epsilon}^\top \nabla \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y) = \rho \frac{\nabla \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y)}{\|\nabla \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y)\|_2}. \quad (7)$$

Therefore, sharpness can be computed more feasibly as:

$$\mathcal{S} \approx \ell(f(\boldsymbol{x}; \boldsymbol{\theta} + \boldsymbol{\epsilon}^*), y) - \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y). \quad (8)$$

Unlike a single training loss, sharpness reflects the rate of change in training loss across the loss surface. Prior research has shown that this rate of change is closely related to generalization capacity. To be specific, models generally perform better on a flat minimum (*i.e.*, with smaller sharpness) than on a sharp minimum (*i.e.*, with larger sharpness) [Foret *et al.*, 2021; Zhuang *et al.*, 2022]. Therefore, we select this indicator to measure generalization capacity.
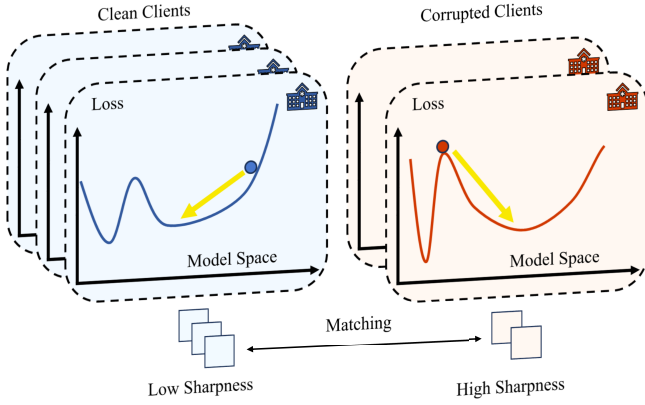
Figure 3: Illustration of FedISM. Our insight is to ensure the uniformity of sharpness across clients, leading to fair generalization.

3.4 Sharpness Matching for Fair Generalization

As the sharpness of the loss surface is indicative of a model's generalization ability, we propose an FL method with inter-client sharpness matching (FedISM) to establish a uniform sharpness distribution across clients, thereby achieving fairer generalization. The overview of FedISM is depicted in Fig. 3, and details are as follows.

Our proposed method is structured around the following objective:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \in \Delta^{K-1}} \sum_{k=1}^{K} \frac{\lambda_k}{N_k} \sum_{(\boldsymbol{x},y) \in D_k} \mathcal{S}(f(\boldsymbol{x};\boldsymbol{\theta}),y), \qquad (9)$$

where $\mathcal{S}(f(\boldsymbol{x};\boldsymbol{\theta}),y)$ represents the sharpness for a given sample and model, as defined in Eq. 8. Unlike previous solutions which focus on a uniformly-low loss (Eq. 4), our key objective is to achieve a uniformly-low sharpness across clients, thereby avoiding convergence to sharp minima while aligning empirical risks. Considering that lower sharpness typically reflects better generalization, this strategy emphasizes fairness regarding generalization more.

In conventional local training in FL, *e.g.*, FedAvg [McMahan *et al.*, 2017], the objective is to minimize the loss on local data via gradient descent:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y), \qquad (10)$$

with η as the learning rate. However, it tends to find an isolated point of low loss, without aiming for a flat minimum as desired in Eq. 9. To address this, local training should incorporate sharpness-awareness. Inspired by sharpness-aware minimization [Foret *et al.*, 2021], the update rule is adapted to:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla \ell(f(\boldsymbol{x};\boldsymbol{\theta}+\boldsymbol{\epsilon}^*),y), \qquad (11)$$

where $\boldsymbol{\epsilon}^*$ represents the optimal perturbation that maximizes the change in the training loss as Eq. 7. This modification directs the optimization towards a point where the surrounding area exhibits the lowest loss, effectively minimizing the sharpness of the loss surface.

After each round of local training, gradients from all clients are transmitted to the server for global aggregation. Notably,

Algorithm 1 Pseudocode of `FedAvg` and our `FedISM`

Input: number of clients K, local datasets $\{D_1, \ldots, D_K\}$, local dataset size, total communication rounds T, learning rate of local training η.
Output: final global model $\boldsymbol{\theta}_{T+1}$
 1: Initialize the global model $\boldsymbol{\theta}_1$
 2: **for** $t = 1, 2, \ldots, T$ **do**
 3:     **for** Client $k = 1, 2, \ldots, K$ in parallel **do**
 4:         $\boldsymbol{\theta}_{(t,k)} \leftarrow \boldsymbol{\theta}_t$ ▷ download the global model
 5:         **for** $(\boldsymbol{x}_i, y_i) \in D_k$ **do**
 6:             Update $\boldsymbol{\theta}_{(t,k)}$ with $(\boldsymbol{x}_i, y_i)$ by Eq. 10
 7:             Update $\boldsymbol{\theta}_{(t,k)}$ with $(\boldsymbol{x}_i, y_i)$ by Eq. 11
 8:         **end for**
 9:     **end for**
10:     $\boldsymbol{\theta}_{t+1} \leftarrow$ Aggregate $\{\boldsymbol{\theta}_{(t,k)}\}_{k=1}^{K}$ with $\boldsymbol{w}_{\text{Avg}}$ (Eq. 12)
11:     $\boldsymbol{\theta}_{t+1} \leftarrow$ Aggregate $\{\boldsymbol{\theta}_{(t,k)}\}_{k=1}^{K}$ with $\boldsymbol{w}$ (Eq. 14)
12: **end for**
13: **return** $\boldsymbol{\theta}_{T+1}$

---

under sharpness-aware minimization (Eq. 11), each client's gradient is representative of the direction to minimize sharpness with respect to local data. The crux of achieving uniform sharpness across clients lies in how these gradients are fairly aggregated. Traditional FedAvg [McMahan *et al.*, 2017] assigns aggregation weights based on the quantity of data each client contributes, as defined by:

$$\boldsymbol{w}_{\text{Avg}} = \frac{1}{\sum_{k=1}^{K} N_k} [N_1, N_2, \cdots, N_K]^\top. \qquad (12)$$

However, such a weighting policy may not facilitate uniform sharpness. Given that clients with corrupted images typically have less data, the global update tends to be dominated by clients with clean images. Consequently, this might only reduce the sharpness for clean clients, leading to a disparity in sharpness similar to the loss disparity addressed by fair optimization [Mohri *et al.*, 2019; Li *et al.*, 2020; Ezzeldin *et al.*, 2023; Jiang *et al.*, 2023a]. To solve this, we propose sharpness-aware aggregation:

$$\widetilde{\boldsymbol{w}}_t = \frac{1}{\sum_{k=1}^{K} \mathcal{S}_{k,t}^q} [\mathcal{S}_{1,t}^q, \mathcal{S}_{2,t}^q, \cdots, \mathcal{S}_{K,t}^q]^\top, \qquad (13)$$

where $\mathcal{S}_{k,t}$ denotes the sharpness calculated on the entire local dataset D_k at any round t by Eq. 8, and q is a predetermined positive parameter. This strategy emphasizes clients with higher sharpness levels during aggregation. As q increases, the aggregation process focuses more on these high-sharpness clients. In extreme cases, aggregation will exclusively focus on the client exhibiting the highest sharpness given $q \to +\infty$. To conclude, this method ensures that the sharpness levels of clients with initially higher sharpness are primarily reduced, thus promoting a uniform sharpness distribution across all clients. In this way, it accomplishes the objective defined in Eq. 9. To maintain stability in federated training, a moving average is further employed for rounds

$t > 1$, formulated as:

$$\boldsymbol{w}_t = \beta \widetilde{\boldsymbol{w}}_t + (1 - \beta) \boldsymbol{w}_{t-1}. \qquad (14)$$

In particular, we set $\boldsymbol{w}_1 = \widetilde{\boldsymbol{w}}_1$ for the first round.

It is important to note that FedISM only requires clients to share their sharpness values instead of any information regarding data distributions (*e.g.*, imaging quality of each client), which is privacy-preserving.

To facilitate a clearer understanding, the processes of both FedISM and the conventional FedAvg [McMahan *et al.*, 2017] are summarized in Algorithm 1. Compared to FedAvg, FedISM modifies only the local optimizer and global aggregation weights, without the need to introduce complex loss functions or additional regularization techniques. Such a streamlined approach enhances the ease of implementation.

4 Experiments

We conduct a series of experiments to assess the effectiveness of FedISM. More results and discussion can be found in Appendix.

4.1 Experimental Setup

Datasets

Two medical datasets are used for evaluation, in line with prior FL research [Jiang *et al.*, 2022; Wu *et al.*, 2023a]:

- RSNA ICH [Flanders *et al.*, 2020]: The task is to classify each CT slice into five intracranial hemorrhage (ICH) subtypes. Following [Jiang *et al.*, 2022], we randomly select 25000 images for experiments.

- ISIC 2019 [Tschandl *et al.*, 2018; Codella and others, 2018; Combalia *et al.*, 2019]: This dataset contains 25331 images for developing models to classify eight skin diseases.

Both datasets are split into training and test sets in an 8:2 ratio and resized to 224×224 pixels following the standard operation [Jiang *et al.*, 2022]. For data partitioning, training sets are partitioned into 20 clients using a Dirichlet distribution (*i.e.*, $Dir(1.0)$), simulating the prevalent label distribution shifts though this paper does not focus mainly on it. Imaging quality shifts are created via Gaussian noise added on a subset of clients following [Hendrycks and Dietterich, 2019].

Model

For standard evaluation, pretrained ResNet-18 [He *et al.*, 2016] is used as the base model for all experiments.

Implement Details

To mitigate label distribution shifts between local training and testing sets, we incorporate logit adjustment [Menon *et al.*, 2021] in the training of local models. We train batches of 32 images using the Adam optimizer, with a constant learning rate of 0.0003, beta values of (0.9, 0.999), and a weight decay of 0.0005. For FL specifics, we set a maximum of 300 communication rounds and the local epoch as 1. These parameters are consistent across all experiments to facilitate fair comparison. In FedISM, we implement GSAM [Zhuang *et al.*, 2022] for sharpness-aware minimization, with default settings of $q = 2.0$ and $\beta = 0.5$.

Evaluation Strategy

The federated model's performance is evaluated on both an unaltered clean testing set and a generated corrupted testing set (with the identically distributed Gaussian noise in corrupted clients). We use class-balanced accuracy (ACC) and the area under the receiver operating characteristic curve (AUC) as our evaluation metrics. To ensure the robustness of results, three independent experiments are conducted, and performance is averaged over the last five communication rounds, in line with [Huang *et al.*, 2023b].

4.2 Comparison to State-of-the-Arts

To validate the superiority, we compare FedISM with several leading methods, including the basic FL approach FedAvg [McMahan *et al.*, 2017], and five advanced fair FL methods: Agnostic-FL [Mohri *et al.*, 2019], q-FedAvg [Li *et al.*, 2020], FairFed [Ezzeldin *et al.*, 2023], FedCE [Jiang *et al.*, 2023a], and FedGA [Zhang *et al.*, 2023]. Comprehensive details on these methods and implementation strategies are available in the Appendix. In our experiments, 4 out of 20 clients are equipped with corrupted images, constituting a corrupted client ratio of 20%.

Tab. 1 summarizes the mean and standard deviation of ACC and AUC for both clean and corrupted images, as well as their average. Although most fair FL methods improve the performance on corrupted images, excluding the less stable Agnostic-FL, they focus primarily on fair optimization rather than generalization, which can be suboptimal. Comparatively, our FedISM, with its emphasis on sharpness minimization, achieves better generalization on corrupted images. For instance, on the ICH dataset, FedISM outperforms FedAvg and the second-best method (FedGA) by 10.85% and 3.78%, respectively. Notably, previous methods often boost the performance on corrupted images at the expense of performance degradation on clean images in this setting. For instance, FedGA's ACC on clean ICH images falls by 4.84% compared to FedAvg. This trade-off can be problematic in medical scenarios, potentially discouraging high-quality institutions from participation in FL. In contrast, FedISM not only enhances the performance on corrupted images but also maintains top results on clean images. This balanced improvement is crucial in medical scenarios, ensuring accurate diagnostics and encouraging broader participation in FL.

4.3 Ablation Study

FedISM is comprised of two key components: Sharpness-Aware Local Training (SALT, Eq. 11) and Sharpness-Aware Global Aggregation (SAGA, Eqs. 13 and 14). To evaluate the effectiveness of each component, we conduct ablation studies by integrating them individually with FedAvg, with performance summaries provided in Table 2. SALT, by directing the model towards flat minima, enhances FedAvg's performance on both clean and corrupted images. Nonetheless, this approach is not fully optimized for fairness, as it fails to address the disparity in sharpness between clean and corrupted image distributions. This is where SAGA comes into play; it helps the model prioritize distributions with poorer generalization capacity, typically those of corrupted images. As a result, the combination of both SALT and SAGA yields

Category	Method	Dataset (Corrupted Clients Ratio: 20%)											
		ICH						ISIC 2019					
		Clean		Corrupted		Average		Clean		Corrupted		Average	
		ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑
Naive FL	FedAvg (AISTATS'17)	76.77 (0.68)	94.58 (0.20)	53.66 (1.28)	84.37 (0.73)	65.21 (0.75)	89.48 (0.34)	64.43 (1.01)	91.91 (0.40)	38.26 (1.57)	78.03 (1.32)	51.35 (0.70)	84.97 (0.64)
Fair FL	Agnostic-FL (ICML'19)	55.13 (7.28)	83.20 (4.86)	46.69 (7.54)	78.13 (4.79)	50.91 (2.56)	80.67 (1.65)	39.66 (12.94)	78.64 (9.08)	33.55 (8.88)	75.62 (6.48)	36.60 (3.30)	77.13 (2.52)
	q-FedAvg (ICLR'20)	75.94 (1.05)	94.36 (0.28)	59.46 (1.90)	86.81 (0.95)	67.70 (0.69)	90.58 (0.40)	65.20 (1.26)	91.59 (0.67)	44.54 (0.80)	82.88 (0.61)	54.87 (0.58)	87.24 (0.47)
	FairFed (AAAI'23)	74.13 (1.15)	93.55 (0.30)	60.27 (1.08)	86.74 (0.63)	67.20 (0.79)	90.15 (0.38)	60.36 (1.59)	90.29 (0.54)	49.04 (1.80)	84.86 (0.64)	54.70 (1.39)	87.58 (0.52)
	FedCE (CVPR'23)	75.82 (0.45)	94.31 (0.10)	58.77 (1.92)	86.93 (0.70)	67.29 (0.94)	90.62 (0.37)	62.21 (1.27)	90.37 (0.56)	45.25 (2.65)	83.37 (1.72)	53.73 (1.11)	86.87 (0.91)
	FedGA (CVPR'23)	71.93 (1.43)	92.88 (0.36)	60.73 (1.24)	87.45 (0.47)	66.33 (0.99)	90.16 (0.33)	59.56 (1.55)	89.53 (0.55)	48.16 (1.38)	84.64 (0.47)	53.86 (1.03)	87.08 (0.47)
Ours	FedISM	**77.52** (0.58)	**95.02** (0.16)	**64.51** (0.64)	**89.56** (0.22)	**71.01** (0.36)	**92.29** (0.13)	**66.94** (0.84)	**93.21** (0.32)	**51.62** (1.39)	**86.89** (0.49)	**59.28** (0.52)	**90.05** (0.18)

Table 1: Comparison of the mean (%) and standard deviation (%) for ACC and AUC against state-of-the-art methods. Values without parentheses represent the mean, while those within parentheses indicate the standard deviation.

the best performance on corrupted images, demonstrating the comprehensive effectiveness of FedISM's designs.

4.4 SALT Helps Fair Optimization

A key motivation for developing Sharpness-Aware Local Training (SALT) is to address the issue of fair optimization methods converging to sharp minima, often resulting in sub-optimal performance (see Fig. 2). To validate this premise and demonstrate SALT's effectiveness, we integrate SALT with existing fair optimization methods and analyze the impact on performance enhancement. The results, presented in Tab. 3, reveal that, aside from Agnostic-FL [Mohri *et al.*, 2019] which shows large standard deviations, other methods exhibit performance improvements when combined with SALT. It not only confirms that SALT is an effective and adaptable component for fair FL, but also validates our initial motivation for its design.

4.5 Robustness to Different Imbalanced Ratios

Our experiments also demonstrate the robustness of FedISM, showcasing stable performance enhancements across varying ratios of clean and corrupted clients. Quantitative results under various ratios of clients with corrupted images through $\{0.1, 0.2, 0.3\}$ are illustrated in Fig. 4. Across all ratios and on both image distributions, FedISM consistently outperforms other methods in the two evaluation metrics. Such findings further highlight FedISM's strong adaptability and effectiveness irrespective of the corruption ratio.

4.6 Discussion on Parameters

The parameter q is closely related to how strongly FedISM concentrates on clients with higher sharpness levels. We examine this impact by varying q through $\{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$ and assess the performance on the ICH dataset with 20% corrupted clients, as shown in Fig. 5. For comparison, we also report the performance of FedAvg and the second-best fair FL method. As q increases, FedISM increasingly focuses on clients with greater sharpness (typically those with corrupted images), which results in a decrease in performance
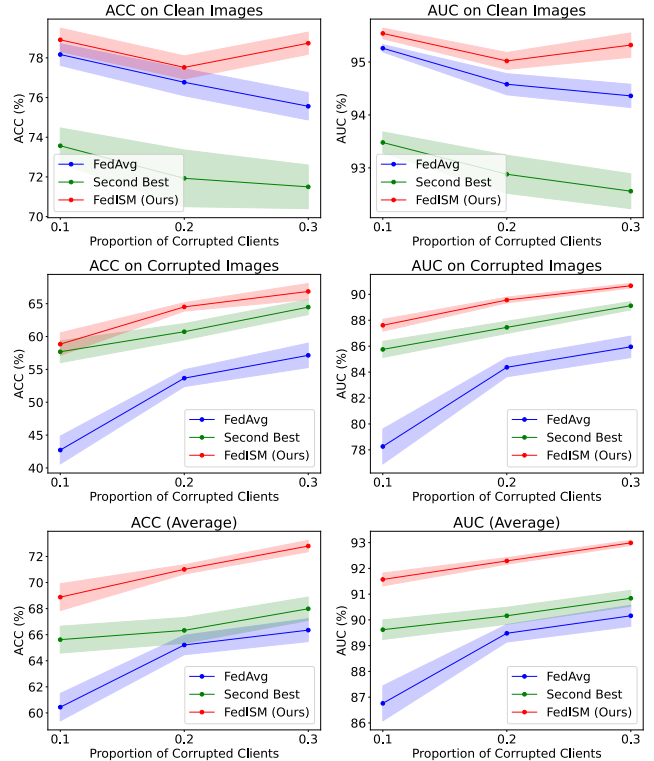

Figure 4: Evaluation across diverse ratios of clean and corrupted clients. Solid lines denote the mean values, while the transparent areas depict the standard deviations. Second best refers to the second best fair FL methods from Section 4.2, specifically in terms of their performance on corrupted images.

on clean images and an improvement on corrupted images. It is important to note that using very high values of q, such as 5 and 10, would slightly degrade the performance on corrupted images, likely due to training instability. Overall, FedISM demonstrates improved performance across a wide range of q values, reducing the burden of parameter tuning in practice.

| Method | ICH (Corrupted Clients Ratio: 20%) | | | | | | | | | | | |
| | Clean | | | | Crorrupted | | | | Average | | | |
	ACC ↑	ΔACC ↑	AUC ↑	ΔAUC ↑	ACC ↑	ΔACC ↑	AUC ↑	ΔAUC ↑	ACC ↑	ΔACC ↑	AUC ↑	ΔAUC ↑
FedAvg (AISTATS'17)	76.77 (0.68)	-	94.58 (0.20)	-	53.66 (1.28)	-	84.37 (0.73)	-	65.21 (0.75)	-	89.48 (0.34)	-
FedAvg +SALT	**79.21** **(0.21)**	**+2.44**	**95.66** **(0.10)**	**+1.08**	55.29 (2.05)	+1.63	86.61 (0.70)	+2.24	67.25 (1.08)	+2.04	91.14 (0.32)	+1.66
FedAvg +SAGA	75.45 (1.20)	−1.32	93.91 (0.49)	−0.67	61.24 (1.81)	+7.58	87.46 (0.95)	+3.09	68.34 (0.60)	+3.13	90.68 (0.26)	+1.20
FedISM (Ours)	77.52 (0.58)	+0.75	95.02 (0.16)	+0.44	**64.51** **(0.64)**	**+10.85**	**89.56** **(0.22)**	**+5.19**	**71.01** **(0.36)**	**+5.80**	**92.29** **(0.13)**	**+2.81**

Table 2: Component-wise ablation study in mean (%) and standard deviation (%) in ACC and AUC. Values without parentheses represent the mean, while those within parentheses indicate the standard deviation. ΔACC and ΔAUC represent the difference to ACC and AUC of FedAvg, respectively.

| Method | ICH (Corrupted Clients Ratio: 20%) | | | |
| | Corrupted | | | |
	ACC	ΔACC ↑	AUC	ΔAUC ↑
FedAvg	53.66 (1.28)	-	84.37 (0.73)	-
+ SALT	55.29 (2.05)	+1.63	86.61 (0.70)	+2.24
Agnostic-FL	46.69 (7.54)	-	78.13 (4.79)	-
+ SALT	45.91 (10.89)	−0.78	78.23 (6.90)	+0.10
q-FedAvg	59.46 (1.90)	-	86.81 (0.95)	-
+ SALT	61.30 (1.57)	+1.84	88.66 (0.51)	+1.85
FairFed	60.27 (1.08)	-	86.74 (0.63)	-
+ SALT	62.86 (1.08)	+2.59	89.14 (0.39)	+2.40
FedCE	58.77 (1.92)	-	86.93 (0.70)	-
+ SALT	62.34 (0.94)	+3.57	88.81 ((0.41)	+1.88
FedGA	60.73 (1.24)	-	87.45 (0.47)	-
+ SALT	61.76 (0.78)	+1.03	88.56 (0.15)	+1.11

Table 3: Performance enhancement in mean (%) and standard deviation (%) on corrupted images by combining SALT with existing FL for fair optimization.

Figure 5: Evaluation on different q. Solid lines denote the mean values, while the transparent areas depict the standard deviations. Second best refers to the second best fair FL methods from Section 4.2, specifically in terms of their performance on corrupted images.

5 Conclusion

In this paper, we pioneer the identification and formulation of a new fairness challenge in FL, specifically concerning imaging quality shifts across clients. To address this problem, existing FL approaches have primarily concentrated on balancing the empirical risk among distinct client distributions. Despite their effectiveness for fair optimization, they often neglect the crucial aspect of fair generalization. To address this overlooked area, we introduce Federated Learning with Inter-client Sharpness Matching (FedISM). FedISM innovatively refine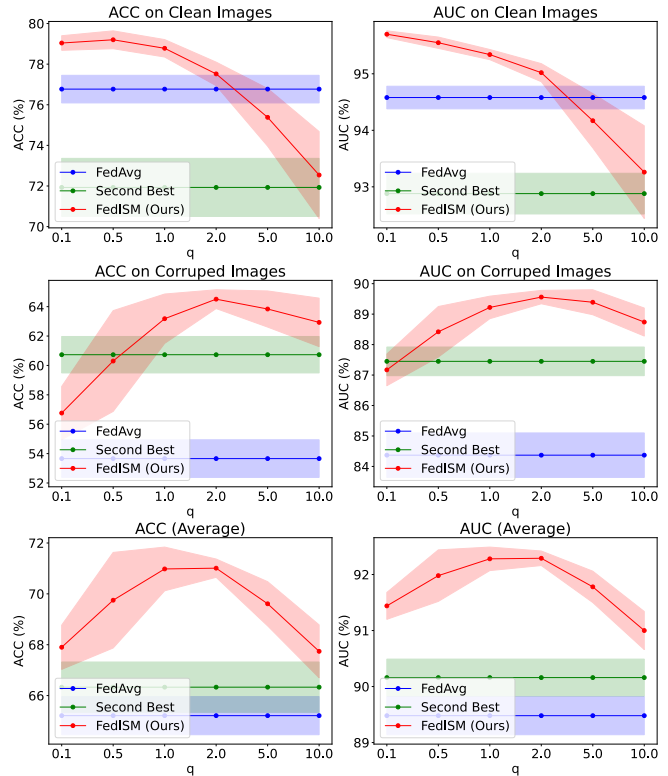s both local training and global aggrega-tion by integrating sharpness-awareness, effectively harmonizing sharpness levels across clients to achieve fair generalization in FL. Extensive empirical evaluations, on the well-recognized RSNA ICH and ISIC 2019 datasets, clearly demonstrate FedISM's superiority over current state-of-the-art FL methods in terms of promoting fairness. It highlights the effectiveness of FedISM in resolving fairness issues stemming from imaging quality shifts across medical datasets. We are confident that our proposed challenge and solution will pave the way for more equitable and effective FL system designs in medical applications and beyond.

Acknowledgments

References

[Caldarola *et al.*, 2022] Debora Caldarola, Barbara Caputo, and Marco Ciccone. Improving generalization in federated learning by seeking flat minima. In *ECCV*, 2022.

[Chen *et al.*, 2023] Zhen Chen, Wuyang Li, Xiaohan Xing, and Yixuan Yuan. Medical federated learning with joint graph purification for noisy label learning. *Medical Image Anal.*, 90:102976, 2023.

[Codella and others, 2018] Noel CF Codella et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *ISBI*, pages 168–172, 2018.

[Combalia *et al.*, 2019] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. BCN20000: Dermoscopic lesions in the wild. *arXiv:1908.02288*, 2019.

[Dou *et al.*, 2021] Qi Dou, Tiffany Y So, Meirui Jiang, Quande Liu, Varut Vardhanabhuti, Georgios Kaissis, Zeju Li, Weixin Si, Heather HC Lee, Kevin Yu, et al. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *npj Digit. Medicine*, 4(1):60, 2021.

[Ezzeldin *et al.*, 2023] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *AAAI*, 2023.

[Fang *et al.*, 2023] Xiuwen Fang, Mang Ye, and Xiyuan Yang. Robust heterogeneous federated learning under data corruption. In *ICCV*, 2023.

[Flanders *et al.*, 2020] Adam E Flanders, Luciano M Prevedello, George Shih, Safwan S Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T Mongan, Anouk Stein, Felipe C Kitamura, Matthew P Lungren, et al. Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3), 2020.

[Foret *et al.*, 2021] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hendrycks and Dietterich, 2019] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

[Huang *et al.*, 2022] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, 2022.

[Huang *et al.*, 2023a] Wenke Huang, Mang Ye, Zekun Shi, and Bo Du. Generalizable heterogeneous federated cross-correlation and instance similarity learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[Huang *et al.*, 2023b] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, 2023.

[Huang *et al.*, 2023c] Wenke Huang, Mang Ye, Zekun Shi, Guancheng Wan, He Li, Bo Du, and Qiang Yang. Federated learning for generalization, robustness, fairness: A survey and benchmark. *arXiv:2311.06750*, 2023.

[Huang *et al.*, 2023d] Zhuo Huang, Miaoxi Zhu, Xiaobo Xia, Li Shen, Jun Yu, Chen Gong, Bo Han, Bo Du, and Tongliang Liu. Robust generalization against photon-limited corruptions via worst-case sharpness minimization. In *CVPR*, 2023.

[Jiang *et al.*, 2022] Meirui Jiang, Hongzheng Yang, Xiaoxiao Li, Quande Liu, Pheng-Ann Heng, and Qi Dou. Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance. In *MICCAI*, pages 196–206, 2022.

[Jiang *et al.*, 2023a] Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *CVPR*, pages 16302–16311, 2023.

[Jiang *et al.*, 2023b] Meirui Jiang, Hongzheng Yang, Chen Cheng, and Qi Dou. IOP-FL: Inside-outside personalization for federated medical image segmentation. *IEEE Trans. Medical Imaging*, 2023.

[Li *et al.*, 2020] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *ICLR*, 2020.

[Li *et al.*, 2021] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-iid features via local batch normalization. In *ICLR*, 2021.

[Liu *et al.*, 2021] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021.

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

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks

from decentralized data. In *AISTATS*, pages 1273–1282, 2017.

[Menon *et al.*, 2021] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021.

[Mohri *et al.*, 2019] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *ICML*, pages 4615–4625, 2019.

[Papadaki *et al.*, 2022] Afroditi Papadaki, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and Miguel Rodrigues. Minimax demographic group fairness in federated learning. In *ACM FAccT*, pages 142–159, 2022.

[Qu *et al.*, 2022] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *ICML*, 2022.

[Sagawa *et al.*, 2020] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

[Sun *et al.*, 2023] Yan Sun, Li Shen, Shixiang Chen, Liang Ding, and Dacheng Tao. Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape. In *ICML*, 2023.

[Tschandl *et al.*, 2018] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

[Wu *et al.*, 2023a] Nannan Wu, Li Yu, Xuefeng Jiang, Kwang-Ting Cheng, and Zengqiang Yan. Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. In *IJCAI*, 2023.

[Wu *et al.*, 2023b] Nannan Wu, Li Yu, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. Fediic: Towards robust federated learning for class-imbalanced medical image classification. In *MICCAI*, 2023.

[Wu *et al.*, 2024] Nannan Wu, Zhaobin Sun, Zengqiang Yan, and Li Yu. Feda3i: Annotation quality-aware aggregation for federated medical image segmentation against heterogeneous annotation noise. In *AAAI*, 2024.

[Xu *et al.*, 2021] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. 2021.

[Ye *et al.*, 2023] Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3):1–44, 2023.

[Zhang *et al.*, 2022] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *ICML*, 2022.

[Zhang *et al.*, 2023] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *CVPR*, 2023.

[Zhou *et al.*, 2023] Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *ICCV*, pages 11345–11355, 2023.

[Zhuang *et al.*, 2022] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *ICLR*, 2022.