Fair Federated Medical Image Classification Against Quality Shift via Inter-Client Progressive State Matching

Nannan Wu, Zhuo Kuang, Zengqiang Yan, Ping Wang, *Fellow, IEEE,* and Li Yu, *Senior Member, IEEE,*

Abstract—Despite the potential of federated learning in medical applications, inconsistent imaging quality across institutions—stemming from lower-quality data from a minority of clients—biases federated models toward more common high-quality images. This raises significant fairness concerns. Existing fair federated learning methods have demonstrated some effectiveness in solving this problem by aligning a single 0th- or 1st-order state of convergence (*e.g.*, training loss or sharpness). However, we argue in this work that fairness based on such a single state is still not an adequate surrogate for fairness during testing, as these single metrics fail to fully capture the convergence characteristics, making them suboptimal for guiding fair learning. To address this limitation, we develop a generalized framework. Specifically, we propose assessing convergence using multiple states, defined as sharpness or perturbed loss computed at varying search distances. Building on this comprehensive assessment, we propose promoting fairness for these states across clients to achieve our ultimate fairness objective. This is accomplished through the proposed method, FedISM+. In FedISM+, the search distance evolves over time, progressively focusing on different states. We then incorporate two components in local training and global aggregation to ensure cross-client fairness for each state. This gradually makes convergence equitable for all states, thereby improving fairness during testing. Our empirical evaluations, performed on the well-known RSNA ICH and ISIC 2019 datasets, demonstrate the superiority of FedISM+ over existing state-of-the-art methods for fair federated learning. The code is available at https://github.com/wnn2000/FFL4MIA.

Index Terms—Federated Learning, Fairness, Medical Image Classification, State Matching

I. Introduction

IN response to growing concerns over data privacy, federated learning (FL) [1] has been recognized as a promising paradigm for safeguarding sensitive information through decentralized training of deep neural networks, particularly in medical domains [2]. Despite its potential, FL faces a significant challenge of data heterogeneity across medical institutions [3], [4], [5], resulting from independent data acquisition processes led by different institutions. Existing research has examined this heterogeneity from various perspectives, including domain shift [6], [7], [8], label distribution skew [9], [10], and label quality variation [11], [12], [13], [14].

Nannan Wu, Zhuo Kuang, Zengqiang Yan, and Li Yu are with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China (e-mail: wnn2000@hust.edu.cn; kuangzhuo@hust.edu.cn; z_yan@hust.edu.cn; hustlyu@hust.edu.cn)

Ping Wang is with the Department of Electrical Engineering and Computer Science, Lassonde School of Engineering, York University, Toronto, Canada (e-mail: ping.wang@lassonde.yorku.ca).
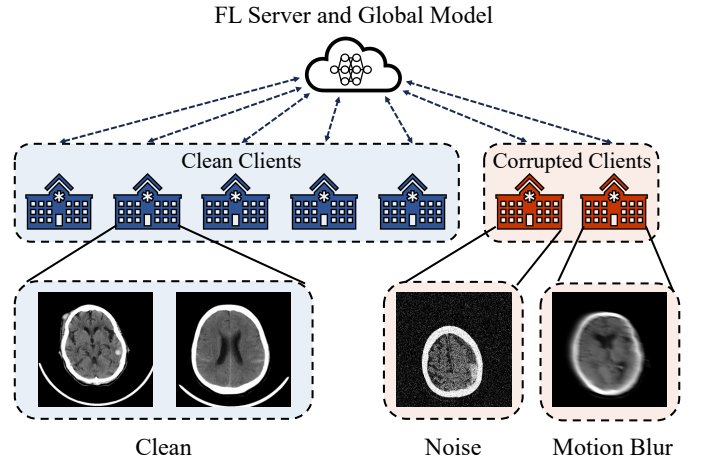
This is the preprint version.

Fig. 1: Imaging quality shifts across clients, where most clients have high-quality (*i.e.*, clean) images, while others experience corruption, such as noise or motion blur.

Nevertheless, the prevalent issue of quality heterogeneity [15] in medical imaging, which may pose new challenges for FL, has not been fully explored.

Although strict protocols are followed in medical imaging environments, image quality uniformity across institutions is not guaranteed. As depicted in Fig. 1, variations in equipment and imaging conditions lead to discrepancies in image quality, resulting in corruption in some images—such as noise from equipment malfunctions or motion blur from involuntary movements—while others remain unaffected. Typically, only a small proportion of images are low-quality (*i.e.*, corrupted) [16], creating a special quality shift. This shift causes standard FL methods, *e.g.*, FedAvg [1], to prioritize the more common clean images, leading to degraded performance on the rarer corrupted data and limiting the applicability of federated models. In this work, we pioneer the identification of this real-world challenge as the significant performance disparity across data quality and demonstrate theoretically that addressing this issue is equivalent to ensuring Rawlsian Max-Min fairness [17] for client performance [18] (see Sec. III-A). This issue is then framed as a client-level fairness problem, namely: ***how can we enhance performance lower bound across clients with varying image quality distributions, including both clean and corrupted images?*** Our study marks the inaugural effort to address fairness in FL across clients experiencing imaging quality shifts, which differs from existing studies that focus on
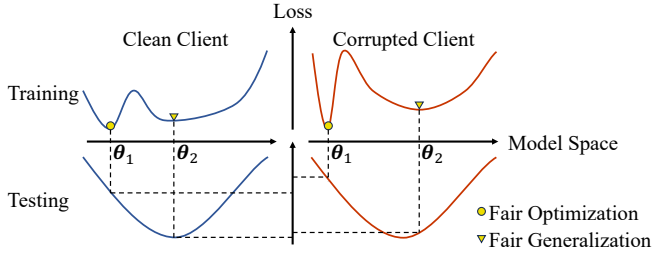
Fig. 2: Comparison of fair optimization and fair generalization: Fair optimization typically leads to convergence at sharp minima (with model parameters θ_1), resulting in uniformly low loss values but poor testing performance. In contrast, fair generalization emphasizes uniformly low sharpness (with model parameters θ_2), which improves absolute performance and fairness when testing.

fairness under domain [19] and label distribution shifts [20].

Previous fair FL solutions can be categorized into two approaches. The first, summarized as *fair optimization* (see Sec. III-B1), focuses on improving fairness during the optimization process across different clients/distributions [19], [20], [21], [22], [23]. Specifically, they monitor a client-specific indicator (*e.g.*, training loss), which directly reflects the degree of the optimization process carried out on each distribution. In response to the selected indicator, the methods dynamically adjust the weights assigned to clients in FL, prioritizing those with poorer performance. This strategy can ultimately reduce the upper bound of the loss across all clients, backed by theoretical guarantees [20], thereby promoting fairness.

However, this approach is sub-optimal as it only considers the 0th-order state of convergence[1], which captures optimization on training data but fails to fully assess generalization to testing data. As shown in Fig. 2, strict adherence to this metric may lead to superficial fairness, such as convergence to a sharp minimum for all clients that severely impairs generalization [24], [25]. Minimizing such a gap points to the second approach, *i.e.*, *fair generalization* (see Sec. III-B2). Motivated by recent advances in sharpness-aware minimization, which highlight the inverse correlation between generalization performance and the 1st-order state of convergence[2] (*e.g.*, sharpness) [24], [25], a preliminary work proposes to minimize and uniformize sharpness across clients [18], as shown in Fig. 2. This is accomplished through a method, federated learning via inter-client sharpness matching (FedISM) [18], which computes sharpness empirically within a predetermined search distance.

Though such a 1st-order state (*i.e.*, sharpness) is better than the 0th-order state for guiding fair learning, it still fails to fully capture the comprehensive characteristics of convergence on the loss surface. For instance, a convergence with uniformly low sharpness (at a fixed search distance) on different distributions can still exhibit varying high loss and high sharpness (at other search distances). This phenomenon leads to sub-optimal generalization, because the optimal sharpness, which

needs to be minimized for better performance, is defined at different search distances across distributions [16]. This indicates that a single state, whether 0th- or 1st-order, is not a reliable surrogate for assessing convergence and guiding the process of promoting fairness. We therefore focus on the following question: *What is the precise surrogate for the generalization ability of the convergence?* To address this, we propose a comprehensive framework by considering multiple states (see Sec. III-C). Specifically, we compute sharpness or perturbed loss at varying search distances, from zero to a maximum. This enables us to pursue the final fairness target by achieving fairness for all these states of convergence across clients. We note that this new framework generalizes previous approaches—fair optimization and fair generalization—where the search distance is fixed at 0 or a positive constant, respectively. Thus, our framework provides a better assessment for convergence.

Based on such comprehensive assessment of convergence, the key to the final fairness target is *how to achieve fairness for these states across clients.* To address this, we propose an enhanced FL method called inter-client progressive states matching (**FedISM+**) (see Sec. III-D). We first introduce a scheme that gradually adjusts the search distance during training, allowing us to focus on different states at different stages of FL. For each state, FedISM+ incorporates sharpness-aware local optimization, with each client minimizing sharpness. During global aggregation, weights are adjusted based on each client's sharpness or perturbed loss level. This ensures that the global update effectively reduces sharpness/perturbed loss, particularly for clients with higher initial levels, resulting in a more uniform state distribution across clients. Throughout the entire FL process, FedISM+ progressively focuses on different states, considering fairness for all states.

We note that FedISM+ is not only easy to implement but also highly effective. Compared to FedAvg [1], it requires only three lines of code modifications. Its effectiveness has been validated through extensive empirical evaluations on two widely used medical image datasets, demonstrating superior performance over state-of-the-art methods (see Sec. IV). We also discuss its communication and privacy properties (see Sec. IV-D6).

This paper is a **substantial extension of our preliminary work** [18] reported in IJCAI-2024 from the following main aspects:

- New Insight: Using a single state to assess convergence and guide the process of promoting fairness, as done in [18], is inadequate for either 1st-order or 0th-order states.
- Comprehensive Convergence Assessment: We introduce a new framework that generalizes previous methods relying on a single state for convergence assessment and fair learning, by considering multiple states.
- An Enhanced FL Method: Based on this comprehensive assessment, we propose achieving the final fairness goal by ensuring cross-client fairness for all these states of convergence. This is done by progressively aligning sharpness/perturbed loss at different search distances, from zero to a maximum. By considering multiple states, this approach effectively combines fair optimization and

---

[1]The 0th-order state of convergence refers to a specific metric (*e.g.*, loss or accuracy) evaluated at convergence itself.

[2]The 1st-order state of convergence refers to the rate of change or trend of a specific metric (*e.g.*, loss or accuracy) near convergence

fair generalization.
- Extensive Empirical Evaluation: We assess the performance of FedISM+ by conducting extensive experiments on two real-world medical image classification datasets, *i.e.*, RSNA ICH and ISIC 2019, demonstrating its superiority over current FL methods in enhancing fairness despite variations in imaging quality.

II. RELATED WORK

A. Fair Federated Learning

Fairness [26] has emerged as a vital topic in FL, focusing mainly on collaborative fairness [27], [28] and performance fairness [20]. Collaborative fairness ensures that each participant's reward is proportional to their contribution to the federation, which is crucial for maintaining motivation and engagement among clients. In contrast, this paper focuses on performance fairness, which typically strives for uniformly high performance (*i.e.*, a low upper bound of loss as Eq. 2) across various devices/features/distributions, thereby ensuring equal benefits from the federation. Current solutions primarily aim to achieve this goal by aligning the single state of convergence across clients. A series of works focus on the 0th-order state [19], [20], [21], [22], [23], while our preliminary work in [18] focuses on the 1st-order state [18]. However, these approaches fail to assess convergence comprehensively, leading to a significant gap between the fairness measured by the chosen metric and the final fairness target.

B. Sharpness of Loss Surface

A major challenge in machine learning is to reconcile training optimization with testing generalization. Recent research on sharpness-aware minimization suggests that models generalize more effectively when they converge to flat minima, as opposed to sharp ones [24], [25]. Building on this understanding, several studies have integrated sharpness metrics to address issues of poor generalization. For instance, SharpDRO [16] combines sharpness with GroupDRO [29] to enhance robust generalization. ImbSAM [30] targets enhancing recognition of tail classes in long-tailed image datasets by reducing the sharpness within these classes. While these approaches have been extensively studied in traditional machine learning contexts, their application in FL has gained attention [31], [32], [33]. Yet, the specific relationship between sharpness and fairness within FL remains largely unexplored.

III. METHODOLOGY

A. Preliminaries

For a standard image classification task with C classes, we consider a cross-silo FL setting involving K participants. Each participant k holds a private dataset $D_k = \{(\boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y})\}_{i=1}^{N_k}, k \in [K] = \{1, 2, \cdots, K\}$, where $\mathcal{X}$ and $\mathcal{Y} = [C] = \{1, 2, \cdots, C\}$ represent the input image and label spaces, respectively. Each 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 indicating an attribute influencing the image quality in client k, *i.e.*, $\mathbb{P}_k(\boldsymbol{x}, y \mid a_k) = \mathbb{P}_k(\boldsymbol{x} \mid y, a_k)\mathbb{P}_k(y)$. This paper highlights

that this quality attribute is non-identical across clients, *i.e*, $\{a_1, a_2, \cdots, a_K\} = [A]$ and $A > 1$. Let $f(\cdot; \boldsymbol{\theta}) : \mathcal{X} \to \Delta^{C-1}$ represent a deep neural network parameterized by $\boldsymbol{\theta}$, and $\ell : \Delta^{C-1} \times \mathcal{Y} \to \{0\} \cup \mathbb{R}+$ denote the loss function, where Δ is the probability simplex. The goal of this paper is to train a model that achieves Rawlsian Max-Min fairness [17] with respect to imaging quality, formulated as:

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

Fair learning that prioritizes the worst group is required to achieve Eq. 1. In traditional centralized learning with all data aggregated, this can be effectively solved through distributionally robust optimization [29], leveraging attribute information (*i.e.*, the imaging quality of each image) obtained in advance. However, such prior knowledge is unavailable due to privacy constraints in FL, making attribute-aware design impractical. Therefore, we approach the problem in a data-agnostic manner inspired by [34], as outlined in the following theorem:

Theorem 1 (Equivalence). *Assuming label distributions of all clients 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 $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ denote client- and attribute-wise weights, respectively, and $\mu_u^* = \sum_{k=1}^{K} \mathbb{1}_{a_k=u}\lambda_k^*$.

As shown above, prioritizing low-performance clients with high expected risk (Eq. 2), *i.e.*, promoting Rawlsian Max-Min fairness [17] for client performance, inherently satisfies the primary objective (Eq. 3). It is important to note that even in the presence of heterogeneity in label distributions, this conclusion remains unchanged, as such heterogeneity can be effectively addressed using logit adjustment techniques [35], [9]. In summary, Theorem 1 provides a solution to the problem that bypasses the quality attribute. We therefore shift the focus of this paper to the newly constructed objective in Eq. 2, *i.e.*, how to achieve performance-fair FL across clients with varying image quality.

B. Previous Solutions: Single State Alignment

Client performance unfairness in FL often arises from the neglect of certain clients with scarce data or atypical distributions (*i.e.*, clients with corrupted images in this work), as improving their performance has negligible impact on the overall objective. To address this issue, previous solutions propose achieving Rawlsian Max-Min fairness [17] in specific metrics, which can be categorized into two main approaches:

1) Fair Optimization: The training loss reflects how well the model fits the training data, making fairness in this metric a potential surrogate for performance fairness. Consequently,

the following objective is constructed:

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

which requires the model to be fairly optimized on data from different clients. This approach is thus referred to as fair optimization. Several methods have been applied to achieve this, including AFL [21], which directs attention to clients with the largest training loss, and q-FedAvg [20], FairFed [22], and FedGA [23], which assign higher importance to clients with larger training losses. FedCE [19] takes an implicit method by focusing on clients with worse task-specific performance metrics.

Despite some effectiveness, it remains challenging for these methods to fully achieve the objective in Eq. 2. As conceptually illustrated in Fig. 2, fair optimization methods may cause the model to quickly converge to sharp minima, as the most significant loss reduction occurs near these points, which Eq. 4 favors. However, this rapid convergence does not necessarily guarantee the achievement of Eq. 2. The reason for this is the gap between training loss and expected loss, especially for smaller-scale data (*e.g.*, corrupted images) [36]. Therefore, it is crucial to broaden the investigation of performance fairness to include generalization, rather than focusing solely on optimization.

2) Fair Generalization: Considering this limitation of fair optimization, our preliminary work proposes shifting the focus to another metric [18]. This starts by analyzing the shortcomings of relying solely on training loss. A major issue is its lack of sensitivity to the underlying geometry of the convergence process, treating sharp and flat minima in the same way. This happens because training loss only reflects the 0th-order state of convergence on the loss surface. To address this, this work proposes focusing on the 1st-order state of convergence, which also considers the training loss in its vicinity. The chosen metric is the sharpness of the loss surface [24], [25], defined as the greatest change in loss when perturbing the initial model parameters:

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

where ρ is a predefined positive parameter that controls the perturbation's search distance. Building on this, the modified objective is formulated as:

$$\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\lambda} \in \Delta^{K-1}} \sum_{k=1}^{K} \frac{\lambda_k}{N_k} \sum_{(\boldsymbol{x},y) \in D_k} \mathcal{S}(f(\boldsymbol{x};\boldsymbol{\theta}),y). \qquad (6)$$

This problem is addressed by the previous method FedISM [18]. In contrast to fair optimization methods that focus on lower bounds of losses (Eq. 4), this method aims to achieve Max-Min fairness in terms of sharpness, promoting convergence to flat minima across all clients' data, as illustrated in Fig. 2. Since reduced sharpness often correlates with improved generalization [24], [25], this method places greater emphasis on fairness in terms of generalization rather than optimization alone.

Summary: The two categories of methods discussed share a common goal: achieving fairness for a single state of convergence on the loss surface (*i.e.*, aligning the state across clients), as illustrated in Fig. 2. The underlying motivation is to treat the fairness of the single metric as a surrogate for the final fairness objective. Specifically, the first category focuses on the 0th-order state of the convergence, while the second emphasizes the 1st-order state by exploring the region within a fixed search distance.

C. New Assessment with Comprehensive States

Although the aforementioned two approaches show some effectiveness in addressing fairness problem, we argue that both are suboptimal. This is because a single state does not fully capture the convergence characteristics, making fairness based on this metric (Eq. 4 and Eq. 6) an imperfect surrogate for the final fairness objective (Eq. 2). In fair optimization, a convergence with uniformly low loss can exhibit varying levels of sharpness, which can be suboptimal for generalization on the test set [18]. For fair generalization, a convergence with uniformly low sharpness (at a fixed search distance) across different distributions can still exhibit high loss and sharpness when considered at other search distances. This leads to suboptimal fairness because the optimal sharpness, which should be minimized for better performance, is defined at different search distances across distributions [16]. Considering their limitations, we ask: how can we better assess a convergence than by using a single state (*i.e.*, loss and sharpness), so that we can leverage this assessment to guide fair learning?

To address this issue, we extend the focus to multiple states. Specifically, we use a series of sharpness (or its variants) defined at different search distances $\rho \in [0, \rho_{max}]$, where ρ_{max} is a predefined maximum.

We begin by defining sharpness as a function of ρ:

$$\mathcal{S}(\rho) := \max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \{\ell(f(\boldsymbol{x};\boldsymbol{\theta}+\boldsymbol{\epsilon}),y) - \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y)\}. \qquad (7)$$

This term is difficult to compute directly due to the continuous and unbounded nature of the perturbation. To make this more manageable, we approximate the difference linearly using the Taylor series expansion, assuming ρ is always small enough:

$$\ell(f(\boldsymbol{x};\boldsymbol{\theta}+\boldsymbol{\epsilon}),y) - \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y) \approx \boldsymbol{\epsilon}^\top \nabla \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y). \qquad (8)$$

Using this approximation, we can determine the optimal perturbation by maximizing the right-hand side expression:

$$\boldsymbol{\epsilon}^* = \arg\max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \boldsymbol{\epsilon}^\top \nabla \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y) = \rho \frac{\nabla \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y)}{\|\nabla \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y)\|_2}. \qquad (9)$$

It is clear that $\boldsymbol{\epsilon}^*$ is a function of ρ, and thus it is denoted as $\boldsymbol{\epsilon}^*(\rho)$. In this way, sharpness can be calculated more feasibly.

$$\mathcal{S}(\rho) \approx \ell(f(\boldsymbol{x};\boldsymbol{\theta}+\boldsymbol{\epsilon}^*(\rho)),y) - \ell(f(\boldsymbol{x};\boldsymbol{\theta}),y). \qquad (10)$$

In addition to sharpness, we explore its variant, the perturbed loss, which also reflects multiple states:

$$\mathcal{L}(\rho) \approx \ell(f(\boldsymbol{x};\boldsymbol{\theta}+\boldsymbol{\epsilon}^*(\rho)),y). \qquad (11)$$

Please note that these two quantities, *i.e.*, $\mathcal{S}$ and $\mathcal{L}$, vary as ρ changes. Therefore, we can assess the convergence by a series
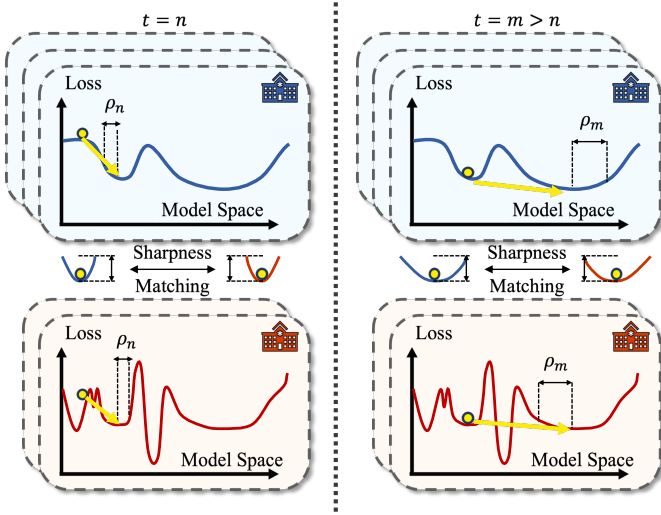
Fig. 3: Illustration of FedISM+. The key insight is to maintain uniformly low sharpness across clients, defined at progressively increasing search distances as training progresses.

of sharpness or perturbed loss, *i.e.*, $\{\mathcal{S}(\rho) \mid \rho \in [0, \rho_{max}]\}$ or $\{\mathcal{L}(\rho) \mid \rho \in [0, \rho_{max}]\}$. This assessment is, in fact, a generalized version of those using the single state discussed in Sec. III-B, as it not only includes the 0th-order state of convergence when $\rho = 0$, but also considers the 1st-order state defined at varying distances $\rho > 0$.

D. FedISM+: Extending State Matching Further

Building on the comprehensive assessment discussed above, we extend the previous single-state matching methods from Sec. III-B. Our insight is to achieve Max-Min fairness for these comprehensive states across clients, rather than relying on the single state as in Eq. 4 and Eq. 6, which better promotes fairness on testing sets.

To achieve this, we propose an enhanced method, called inter-client progressive state matching (**FedISM+**). The overview is depicted in Fig. 3. Building on the previous method, FedISM [18], the key feature of FedISM+ is its progressive scheme that focuses on fairness across multiple states of convergence. This scheme is implemented by defining a search distance for computing sharpness (Eq. 10) and perturbed loss (Eq. 11), as follows:

$$\rho(t) = \rho_{max} \left(\frac{t}{T} \right)^{\tau}, \qquad (12)$$

where ρ_{max} and $\tau > 0$ are predefined parameters, t denotes the current round, and T represents the maximum communication round in FL. This design ensures that the search distance increases from 0 to ρ_{max} during training, controlled by τ. Initially, this scheme effectively considers the 0th-order state since $\rho(t) \approx 0$, and progressively incorporates the 1st-order state as training progresses.

Next, we describe how fairness is achieved for these states. In traditional local training within FL (*e.g.*, FedAvg [1]), the objective is to minimize the local loss via gradient descent:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla \ell(f(\boldsymbol{x}; \boldsymbol{\theta}), y), \qquad (13)$$

Algorithm 1 `FedAvg` and `FedISM+`

Input: Number of clients K, local datasets $\{D_1, \dots, D_K\}$, local dataset size, total communication rounds T, learning rate of local training η, maximum search distance ρ_{max}, parameter τ.

Output: Global model $\boldsymbol{\theta}_{T+1}$

1: Initialize the global model $\boldsymbol{\theta}_1$
2: **for** $t = 1, 2, \dots, T$ **do**
3:      $\rho \leftarrow \rho_{max} \left(\frac{t}{T} \right)^{\tau}$ ▷ set ρ for `FedISM+`
4:      **for** Client $k = 1, 2, \dots, K$ in parallel **do**
5:          $\boldsymbol{\theta}_{(t,k)} \leftarrow \boldsymbol{\theta}_t$ ▷ download the global model
6:          **for** $(\boldsymbol{x}_i, y_i) \in D_k$ **do**
7:              Update $\boldsymbol{\theta}_{(t,k)}$ with $(\boldsymbol{x}_i, y_i)$ by Eq. 13
8:              Update $\boldsymbol{\theta}_{(t,k)}$ with $(\boldsymbol{x}_i, y_i)$ by Eq. 14
9:          **end for**
10:     **end for**
11:     $\boldsymbol{\theta}_{t+1} \leftarrow$ Aggregate $\{\boldsymbol{\theta}_{(t,k)}\}_{k=1}^{K}$ with $\boldsymbol{w}_{\texttt{Avg}}$ (Eq. 15)
12:     $\boldsymbol{\theta}_{t+1} \leftarrow$ Aggregate $\{\boldsymbol{\theta}_{(t,k)}\}_{k=1}^{K}$ with $\boldsymbol{w}_t$ (Eq. 18)
13: **end for**
14: **return** $\boldsymbol{\theta}_{T+1}$

---

where η is the learning rate. However, this method only considers the 0th-order state and does not account for other states, *i.e.* $\mathcal{S}(\rho(t))$ and $\mathcal{L}(\rho(t))$. To overcome this limitation, we introduce sharpness-awareness into local training. Drawing from sharpness-aware minimization [24], we modify the update rule to:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla \ell(f(\boldsymbol{x}; \boldsymbol{\theta} + \boldsymbol{\epsilon}^*(\rho(t))), y), \qquad (14)$$

where $\boldsymbol{\epsilon}^*(\rho(t))$ refers to the optimal perturbation that maximizes the change in training loss, as outlined in Eq. 9. This adjustment guides the optimization process to a destination where the neighboring region has a small loss. The perturbation is a function of $\rho(t)$, which allows for progressively exploring the neighboring region in a comprehensive manner. For simplicity, we refer to this as sharpness-aware local training (SALT).

After each local training round, the gradients from all participating clients are sent to the server for aggregation. Achieving fairness in the states of convergence across clients depends on whether these gradients are aggregated fairly. In vanilla FL [1], aggregation weights are determined by the size of the dataset each client holds, as expressed by:

$$\boldsymbol{w}_{\texttt{Avg}} = \frac{1}{\sum_{k=1}^{K} N_k} [N_1, N_2, \cdots, N_K]^{\top}. \qquad (15)$$

However, this weighting scheme may not achieve our fairness goal. As discussed in Fig. 1, clients with corrupted data typically have fewer data points, leading to the global model update being predominantly driven by clients with clean data. As a result, sharpness and perturbed loss are minimized mainly for clients with clean data, causing a disparity in these metrics that mirrors the loss disparity in fair optimization [21], [20], [22], [19]. To address this issue, we introduce sharpness-aware

TABLE I: Component-wise ablation study. Numbers outside the parentheses correspond to the mean (%), while those within the parentheses are the standard deviation (%). ΔACC (%) and ΔAUC (%) denote the difference to ACC and AUC of FedAvg. The best results are in bold.

Type	Method Name	Component			RSNA ICH											
					Clean				Corrupted				Average			
		SALT	SAGA (S)	SAGA (L)	ACC ↑	ΔACC ↑	AUC ↑	ΔAUC ↑	ACC ↑	ΔACC ↑	AUC ↑	ΔAUC ↑	ACC ↑	ΔACC ↑	AUC ↑	ΔAUC ↑
-	FedAvg [1]				76.77 (0.68)	-	94.58 (0.20)	-	53.66 (1.28)	-	84.37 (0.73)	-	65.21 (0.75)	-	89.48 (0.34)	-
Ablation of FedISM+	-	✓			**78.84 (0.72)**	**+2.07**	**95.64 (0.08)**	**+1.06**	53.53 (2.79)	−0.13	86.50 (0.94)	+2.13	66.19 (1.11)	+0.98	91.07 (0.44)	+1.59
	-		✓		74.48 (1.73)	−2.29	93.73 (0.54)	−0.85	59.16 (2.42)	+5.50	86.96 (0.94)	+2.59	66.82 (1.06)	+1.61	90.34 (0.34)	+0.86
	-			✓	73.44 (1.59)	−3.33	93.38 (0.43)	−1.20	61.39 (1.50)	+7.73	87.68 (0.36)	+3.31	67.42 (0.81)	+2.21	90.53 (0.22)	+1.05
	FedISM+ (S)	✓	✓		77.19 (0.67)	+0.42	95.01 (0.10)	+0.43	**64.55 (0.78)**	**+10.89**	**89.93 (0.22)**	**+5.56**	70.87 (0.42)	+5.66	92.47 (0.09)	+2.99
	FedISM+ (L)	✓		✓	78.09 (0.78)	+1.32	95.31 (0.14)	+0.73	63.83 (1.02)	+10.17	89.72 (0.28)	+5.35	**70.96 (0.70)**	**+5.75**	**92.51 (0.17)**	**+3.03**

global aggregation (SAGA), where the aggregation weights are determined by sharpness, as follows:

$$\widetilde{w}_t = \frac{1}{\sum_{k=1}^{K} \mathcal{S}_{k,t}^{(q)}} [\mathcal{S}_{1,t}^{(q)}, \mathcal{S}_{2,t}^{(q)}, \cdots, \mathcal{S}_{K,t}^{(q)}]^{\top}, \qquad (16)$$

Alternatively, we can use the perturbed loss as the weighting criterion:

$$\widetilde{w}_t = \frac{1}{\sum_{k=1}^{K} \mathcal{L}_{k,t}^{(q)}} [\mathcal{L}_{1,t}^{(q)}, \mathcal{L}_{2,t}^{(q)}, \cdots, \mathcal{L}_{K,t}^{(q)}]^{\top}, \qquad (17)$$

where $\mathcal{S}_{k,t}$ and $\mathcal{L}_{k,t}$ represent the sharpness and perturbed loss computed for the complete local dataset D_k at round t using Eq. 10 and Eq. 11, respectively, with $q > 0$ as the exponent. Methods using these two aggregation weights (denoted as SAGA (S) and SAGA (L)) are referred to as FedISM+ (S) and FedISM+ (L), respectively. This strategy assigns greater weight to clients with higher sharpness or perturbed loss during aggregation. Importantly, when using SALT (Eq. 14), each client's gradient inherently points towards reducing the sharpness specific to its local data. Therefore, the combination of SALT and SAGA prioritizes minimizing the sharpness/perturbed loss for clients with initially higher sharpness/perturbed loss, effectively promoting fairness in these progressive states. To ensure the stability of FL, a simple moving average is further applied for rounds $t > 1$, as follows:

$$w_t = \beta \widetilde{w}_t + (1 - \beta) w_{t-1}. \qquad (18)$$

Particularly, we set $w_1 = \widetilde{w}_1$ for the first round.

It is important to note that FedISM+ only requires clients to transmit their sharpness or perturbed loss, not any information about their data distributions. This preserves privacy, as discussed in Sec. IV-D6.

For clarity, the procedure for FedISM+ is summarized in Alg. 1, where we also include FedAvg [1] for better understanding. It shows that FedISM+ introduces modifications only to the local optimizer and the global aggregation weights compared to FedAvg, along with a scheme to determine the search distance, avoiding the introduction of complex loss functions or regularization methods. This pseudocode illustrates the ease of implementation.

IV. EXPERIMENTS

We conduct a range of experiments to assess the effectiveness of our proposed solution.

A. Experimental Setup

1) Datasets: Two widely-used medical image classification datasets, commonly employed in FL research [37], [11], are used for evaluation:

- RSNA ICH [38]: The task involves classifying CT slices into five intracranial hemorrhage subtypes. Following [37], 25,000 images are randomly selected for experiments.
- ISIC 2019 [39], [40], [41]: This dataset contains 25331 images for training models to classify eight skin diseases.

Both datasets are divided into training and test sets in an 8:2 ratio and resized to 224×224 pixels, in line with the standard preprocessing steps [37].

2) Client Training Data Preparation: Following [42], the training sets are distributed among 20 clients using a Dirichlet distribution (*i.e.*, Dir(1.0)), simulating prevalent label distribution shifts. To create imaging quality shifts, Gaussian noise or motion blur is added to images for a subset of the clients, as depicted in Fig. 1, following [43].

3) Model: Pretrained ResNet-18 [44] is selected as the base model for all experiments for standard evaluation.

4) Implement Details: To mitigate label distribution shifts, we incorporate logit adjustment [35] in local training. Following previous settings [18], we train the local models using batches of 32 images with the Adam optimizer, applying a constant learning rate of 0.0003, beta values of (0.9, 0.999), and a weight decay of 0.0005. For FL setup, we configure a maximum of 300 communication rounds and set the local epoch to 1. These hyperparameters are kept consistent across all experiments to ensure a fair comparison. For FedISM+, we adopt GSAM [25] for sharpness-aware minimization and set $\rho_{max} = 0.1$, $\tau = 0.5$, $q = 2.0$ and $\beta = 0.5$ defaultly.

5) Evaluation Strategy: We follow the previous work [18] to design the evaluation strategy. To maximize the utilization of limited testing images, the testing set is not directly divided among clients. Instead, following [43], a new corrupted testing set is generated from the initial clean one by adding the same type of corruption that is applied to the training set of corrupted clients. The performance of FL is then evaluated on both the original clean and the newly generated corrupted test sets to demonstrate how Eq. 1 is achieved. We use the area under the receiver operating characteristic curve (AUC) and classification accuracy (ACC) as our evaluation metrics. To minimize the impact of randomness, all experiments are

TABLE II: Performance comparison on RSNA ICH. Numbers outside the parentheses correspond to the mean (%), while those within the parentheses are the standard deviation (%). The best results are in bold.

Category	Method	RSNA ICH											
		Gaussian Noise						Motion Blur					
		Clean		Corrupted		Average		Clean		Corrupted		Average	
		ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑
Vanilla FL	FedAvg [1] (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)	75.97 (1.11)	94.74 (0.22)	68.69 (0.74)	90.98 (0.41)	72.33 (0.81)	92.86 (0.29)
Fair Optimization (0th-Order State)	Agnostic-FL [21] (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)	63.57 (3.48)	89.45 (1.32)	57.27 (7.12)	86.35 (3.32)	60.42 (3.27)	87.90 (1.78)
	q-FedAvg [20] (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)	75.68 (1.59)	94.77 (0.34)	70.53 (1.18)	91.87 (0.53)	73.10 (0.54)	93.32 (0.17)
	FairFed [22] (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)	74.59 (1.12)	94.34 (0.21)	69.66 (1.08)	91.75 (0.30)	72.12 (0.66)	93.05 (0.18)
	FedCE [19] (CVPR'23)	75.82 (0.45)	94.31 (0.10)	58.77 (1.92)	86.93 (0.94)	67.29 (0.94)	90.62 (0.37)	76.83 (0.83)	95.00 (0.15)	70.38 (0.72)	92.04 (0.28)	73.60 (0.73)	93.52 (0.17)
	FedGA [23] (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)	72.97 (0.78)	93.75 (0.16)	69.64 (0.60)	91.78 (0.27)	71.31 (0.54)	92.76 (0.16)
Fair Generalization (1st-Order State)	FedISM [18] (IJCAI'24)	77.52 (0.58)	95.02 (0.16)	64.51 (0.64)	89.56 (0.22)	**71.01** (**0.36**)	92.29 (0.13)	75.43 (0.97)	95.06 (0.11)	72.54 (0.79)	93.46 (0.15)	73.99 (0.54)	94.26 (0.12)
Ours (Multiple States)	FedISM+ (S)	77.19 (0.67)	95.01 (0.10)	**64.55** (**0.78**)	**89.93** (**0.22**)	70.87 (0.42)	92.47 (0.09)	76.50 (0.70)	95.17 (0.07)	**72.97** (**0.60**)	93.75 (0.10)	74.73 (0.34)	94.46 (0.05)
	FedISM+ (L)	**78.09** (**0.78**)	**95.31** (**0.14**)	63.83 (1.02)	89.72 (0.28)	70.96 (0.70)	**92.51** (**0.17**)	**77.60** (**0.58**)	**95.50** (**0.07**)	72.88 (0.64)	**93.89** (**0.11**)	**75.24** (**0.42**)	**94.70** (**0.06**)

independently conducted three times. We present the mean and standard deviation calculated from the last five communication rounds, following the strategy in [45].

B. Ablation Study

FedISM+ consists of two primary components: SALT and SAGA. To assess the impact of each component, we perform ablation studies on the RSNA ICH dataset, using 20% of the clients with Gaussian noise corruption, and integrate each component individually with FedAvg [1]. The results of AUC and ACC for both clean and corrupted images, as well as their average, are summarized in Tab. I. SALT, which encourages the model to converge towards flatter minima, typically improves the performance of FedAvg. However, it does not fully lead to fairness, as it overlooks the cross-client difference. This problem is addressed by SAGA. The results demonstrate that both SAGA (S) and SAGA (L) effectively prioritize worst-performing distributions, often those corresponding to corrupted images. Therefore, combining both SALT and SAGA achieves the best results on corrupted images, showcasing the comprehensive benefits of the FedISM+.

C. Comparison to State-of-the-Arts

To demonstrate the superiority, we compare FedISM+ against several state-of-the-art methods, including:

- FedAvg [1]: The vanilla FL approach.
- Agnostic-FL [21]: A pioneering work in fair FL that focuses updates on the poorest-performing client.
- q-FedAvg [20]: Integrates training loss into the global aggregation stage, with an adjustment including a multiplicative constant for improved convergence. Its parameter q is tuned across the values $\{0.5, 1.0, 2.0, 5.0\}$.
- FairFed [22]: Seeks to achieve group fairness by optimizing across clients in a balanced manner, with the parameter β varied within the range $\{0.1, 0.5, 1.0\}$.
- FedCE [19]: Aims to foster fairness in FL, specifically for medical image segmentation. Adapted for classification tasks in this study.

- FedGA [23]: Focuses on improving domain generalization by ensuring fairness across diverse clients.
- FedISM [18]: Aims for fairness in sharpness that is defined at a fixed search distance. This is the previous version of this work. We set $q = 2.0$ and $\rho = 0.05$ as the origin paper.

Among these methods, Agnostic-FL [21], q-FedAvg [20], FairFed [22], FedCE [19], and FedGA [23] focus on the fairness of the 0th-order state of convergence, corresponding to fair optimization, while FedISM [18] focuses on the fairness of the 1st-order state of convergence, corresponding to fair generalization.

In this section, experiments are conducted on two datasets in a setting where 4 out of 20 clients (corruption ratio: 20%) possess images corrupted by Gaussian noise or motion blur.

1) Classification Performance Comparison: Quantitative comparison results measured by the mean and standard deviation (stemming from multiple runs) are summarized in Tabs. II and III. Under the given quality shifts, FedAvg demonstrates a clear performance bias toward clean images, resulting in a diminished lower performance bound across distributions. To address this, previous fair optimization methods focus primarily on aligning the 0th-order state of convergence. In this way, most methods improve the performance on corrupted images, with the exception of the less stable Agnostic-FL. As discussed before, they are sub-optimal due to their neglect of the 1st-order state. In contrast, FedISM, which focuses on the sharpness of the loss surface, achieves better generalization on corrupted images. For instance, on the Gaussian noise-corrupted distribution of the RSNA ICH dataset, FedISM outperforms FedAvg and the best fair optimization method (*i.e.*, FedGA) by 10.85% and 3.78% in ACC, respectively. However, FedISM's limited perception of the loss surface restricts its generalization capacity estimation, thus impairing performance. This limitation is addressed by FedISM+, which progressively increases the search distance from zero, considering both 0th- and multiple 1st-order states. This approach better bridges the gap between training and testing

TABLE III: Performance comparison on ISIC 2019. Numbers outside the parentheses correspond to the mean (%), while those within the parentheses are the standard deviation (%). The best results are in bold.

Category	Method	ISIC 2019											
		Gaussian Noise						Motion Blur					
		Clean		Corrupted		Average		Clean		Corrupted		Average	
		ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑
Vanilla FL	FedAvg [1] (AISTATS'17)	64.43 (1.01)	91.91 (0.40)	38.26 (1.57)	78.03 (1.32)	51.35 (0.70)	84.97 (0.64)	67.16 (1.15)	92.60 (0.37)	54.98 (1.72)	88.16 (0.54)	61.07 (1.31)	90.38 (0.41)
Fair Optimization (0th-Order State)	Agnostic-FL [21] (ICML'19)	39.66 (2.94)	78.64 (9.08)	33.55 (8.88)	75.62 (6.48)	36.60 (3.30)	77.13 (2.52)	51.24 (5.49)	87.90 (1.91)	50.11 (3.85)	86.56 (1.64)	50.67 (2.93)	87.23 (1.18)
	q-FedAvg [20] (ICLR'20)	65.20 (1.26)	91.59 (0.67)	44.54 (0.80)	82.88 (0.61)	54.87 (0.58)	87.24 (0.47)	71.39 (0.73)	93.88 (0.30)	60.79 (1.38)	90.03 (0.39)	66.09 (0.96)	91.95 (0.30)
	FairFed [22] (AAAI'23)	60.36 (1.59)	90.29 (0.54)	49.04 (1.80)	84.86 (0.64)	54.70 (1.39)	87.58 (0.52)	63.94 (1.58)	91.89 (0.48)	56.27 (1.59)	89.58 (0.58)	60.11 (1.48)	90.73 (0.49)
	FedCE [19] (CVPR'23)	62.21 (1.27)	90.37 (0.56)	45.25 (2.65)	83.37 (1.72)	53.73 (1.11)	86.87 (0.91)	72.29 (0.64)	94.26 (0.28)	61.85 (1.28)	90.95 (0.40)	67.07 (0.74)	92.61 (0.31)
	FedGA [23] (CVPR'23)	59.56 (1.55)	89.53 (0.55)	48.16 (1.38)	84.64 (0.47)	53.86 (1.03)	87.08 (0.47)	66.35 (1.53)	92.78 (0.45)	59.22 (1.46)	90.20 (0.36)	62.79 (1.31)	91.49 (0.34)
Fair Generalization (1st-Order State)	FedISM [18] (IJCAI'24)	66.94 (0.84)	93.21 (0.32)	51.62 (1.39)	86.89 (0.49)	59.28 (0.52)	90.05 (0.18)	71.31 (0.51)	94.84 (0.15)	62.54 (0.83)	92.31 (0.18)	66.92 (0.36)	93.57 (0.12)
Ours (Multiple States)	FedISM+ (S)	**69.12 (1.00)**	**94.14 (0.26)**	50.79 (0.99)	87.37 (0.26)	**59.96 (0.81)**	**90.75 (0.15)**	72.25 (0.51)	95.18 (0.14)	**65.47 (1.28)**	93.14 (0.29)	68.86 (0.74)	94.16 (0.19)
	FedISM+ (L)	67.24 (1.14)	93.33 (0.46)	**52.08 (1.66)**	**88.02 (0.51)**	59.66 (0.89)	90.67 (0.27)	**72.70 (0.48)**	**95.41 (0.10)**	65.46 (0.83)	**93.16 (0.23)**	**69.08 (0.43)**	**94.29 (0.10)**

TABLE IV: Client-level fairness comparison. Numbers outside the parentheses correspond to the mean (%), while those within the parentheses are the standard deviation (%). Values in bold denote the best result, and values underlined indicate the second best.

Method	STD of AUC (%) among Clients ↓				
	RSNA ICH		ISIC 2019		Avg
	Gaussian Noise	Motion Blur	Gaussian Noise	Motion Blur	
FedAvg [1]	4.08(0.33)	1.51(0.12)	5.55(0.59)	1.78(0.16)	3.23
Agnostic-FL [21]	3.28(2.56)	1.47(1.19)	5.79(1.91)	1.02(0.61)	2.89
q-FedAvg [20]	3.02(0.46)	1.16(0.33)	3.49(0.35)	1.54(0.14)	2.30
FairFed [22]	2.72(0.25)	1.04(0.15)	2.17(0.24)	0.93(0.17)	1.72
FedCE [19]	2.95(0.27)	1.18(0.11)	2.80(0.72)	1.32(0.12)	2.06
FedGA [23]	_2.17(0.20)_	0.79(0.12)	**1.96(0.17)**	1.03(0.18)	_1.49_
FedISM [18]	2.18(0.12)	_0.64(0.05)_	2.53(0.30)	_1.01(0.10)_	1.59
FedISM+	**2.03(0.11)**	**0.57(0.06)**	_2.13(0.32)_	**0.82(0.10)**	**1.39**

distributions across different clients. The results demonstrate that FedISM+ surpasses FedISM in most metrics for RSNA ICH and all metrics for ISIC 2019; for instance, on the motion blur-corrupted distribution of the ISIC 2019 dataset, FedISM+ outperforms FedISM by more than 0.70% in AUC, exceeding the sum of standard deviations. These outcomes highlight the superiority of our design.

2) Client-Level Fairness Comparison: Rather than focusing solely on the worst-performing distributions, fairness in FL can also be evaluated by the uniformity of classification performance among clients [20]. To evaluate this, we report the standard deviation of performance among clients across multiple runs in Tab. IV. In our experiments, we simulate client-level testing by distributing the entire clean/corrupted testing set to specific clients. The results show that our proposed solution, FedISM+, achieves superior uniformity. While FedGA [23] also performs well, even outperforming FedISM+ in one case, it primarily achieves this uniformity by suppressing performance on clean images (see Tabs. II and III). Such an approach, though it may produce uniform performance, is less meaningful in practice because it does not truly enhance the model's ability to handle diverse distributions.

Remark: Compared to existing solutions, our proposed method, FedISM+, not only enhances classification perfor-

mance on the most challenging distributions but also improves the uniformity of performance across all clients.

D. Discussion

1) Discussion on Search Distance: In this section, we empirically evaluate our progressive scheme of search distance ρ from 0 to ρ_{max} (set as 0.1 in experiments) using the experimental setting mentioned in Sec. IV-B.

Firstly, we show that this scheme is not equivalent to simply tuning ρ for FedISM [18] that considers sharpness defined at a fixed search distance. As shown in the left of Fig. 4, FedISM shows a slower convergence rate with the increase of ρ in the early stage. This is because larger ρ can lead to more focus on the 1st-order state and slow down the decrease of loss. However, in the last stage, larger ρ helps the model converge on a flat minimum, which is better for generalization and achieves high classification performance on the testing set. Our method FedISM+, by increasing ρ as training progresses and considering both 0th- and 1st-order states, achieves a quicker convergence rate in the beginning and better performance in the later stage.

Secondly, we validate the effectiveness of τ, which controls the rate at which ρ evolves, as illustrated on the right side of Fig. 4. By decreasing τ, ρ reaches its maximum more quickly, exhibiting a trend similar to FedISM with a large ρ in the early stages. Because FedISM+ considers both 0th- and 1st-order states, it achieves relatively good convergence in the later stages, even with τ tuning.

Remark: Our proposed dynamic ρ strategy is not trivial as its effectiveness can not be achieved by simply tuning ρ in FedISM. The increase rate of ρ can impact the convergence rate of FedISM+.

2) Discussion on Parameter q: The parameter q plays a key role in determining how much emphasis FedISM+ places on clients with higher sharpness. To investigate its effect, we vary q over the values {0.1, 0.5, 1.0, 2.0, 5.0, 10.0} and evaluate the performance using the setup described in Sec. IV-B, with results shown in the left column of Fig. 5. For reference,
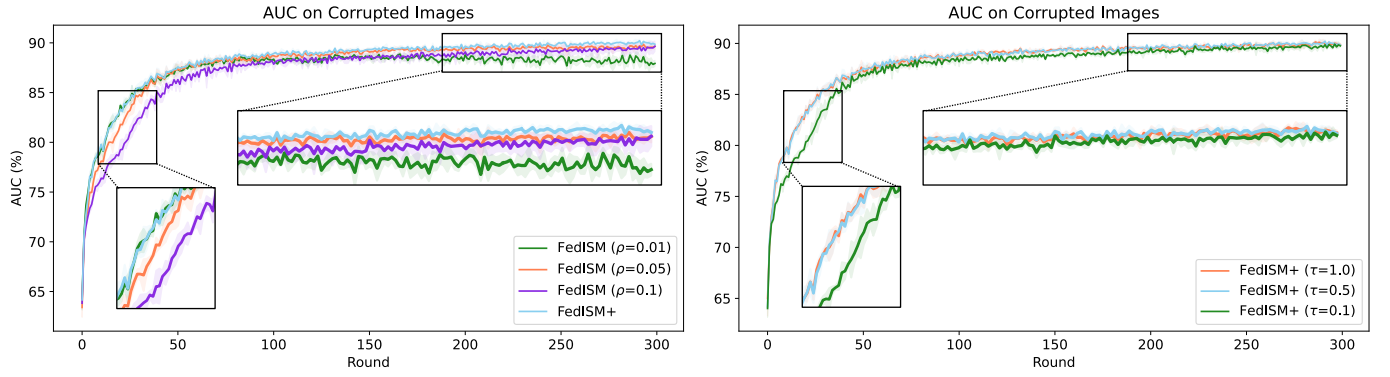
Fig. 4: **Left**: Comparison of FedISM+ with FedISM across different ρ values. **Right**: Evaluation across different τ settings. The solid lines indicate the mean values, and the shaded regions represent the standard deviations.
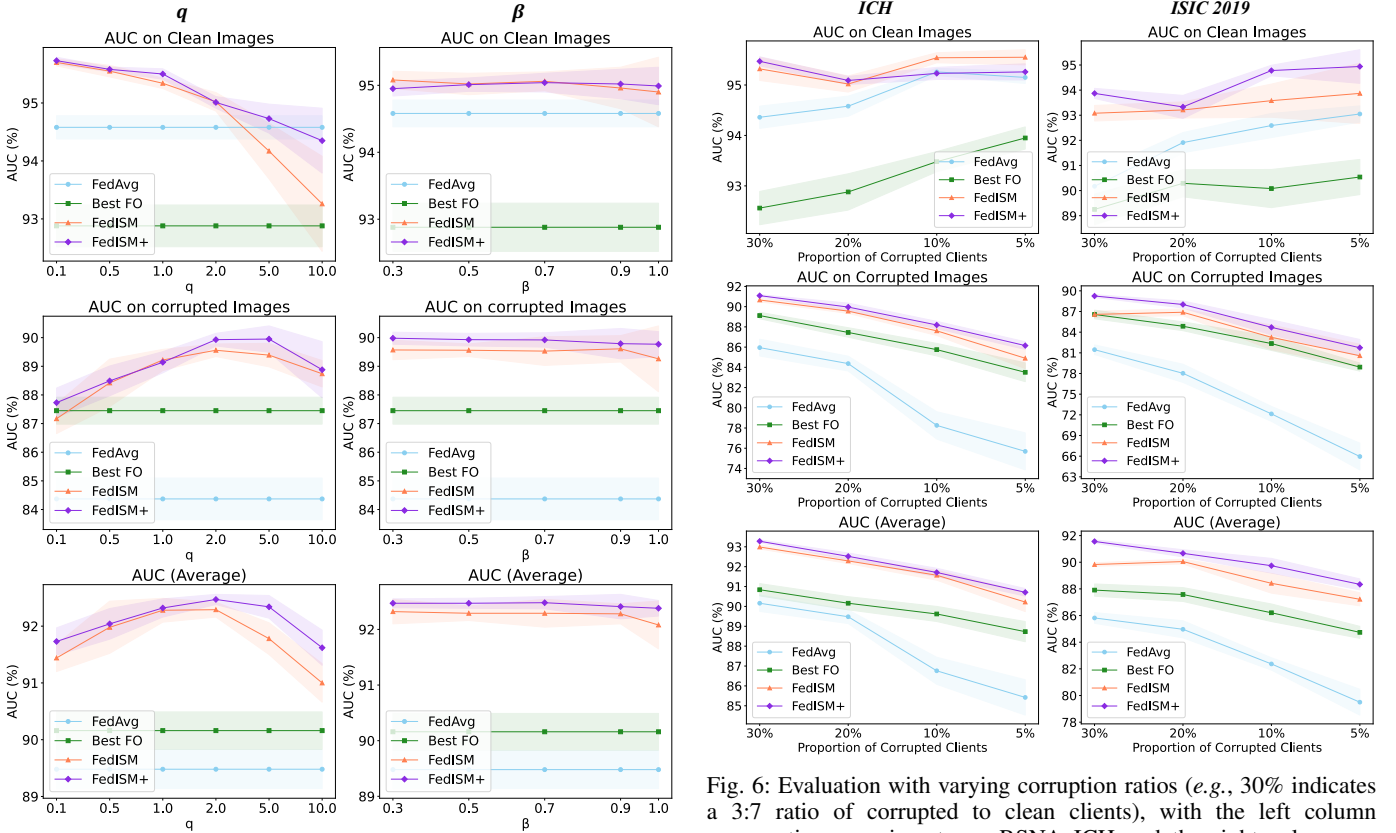

Fig. 5: **Left**: Evaluation on different q. **Right**: Evaluation on different β. The solid lines indicate the mean values, and the shaded regions represent the standard deviations. The best fair optimization method from Section IV-C is denoted as "Best FO" according to the performance on corrupted images.

Fig. 6: Evaluation with varying corruption ratios (*e.g.*, 30% indicates a 3:7 ratio of corrupted to clean clients), with the left column representing experiments on RSNA ICH and the right column on ISIC 2019. The solid lines indicate the mean values, and the shaded regions represent the standard deviations. The best fair optimization method from Section IV-C is denoted as "Best FO" according to the performance on corrupted images. We tune τ from $\{0.1, 0.5, 1.0\}$.

we also report the performance of FedAvg [1], the best fair optimization method discussed in Sec. IV-C and FedISM [18]. Both FedISM and FedISM+ exhibit similar trends as q varies. In particular, increasing q causes both methods to focus more on clients with higher sharpness (*i.e.*, clients with corrupted images), which leads to improved performance on corrupted images but a slight decline on clean images. Notably, very large values of q (such as 5 and 10) result in a minor performance drop on corrupted images, likely due to instability in the training process. On the whole, FedISM(+) performs

well across a wide range of q values, alleviating the need for extensive parameter tuning in practical applications. Additionally, FedISM+ generally outperforms FedISM, highlighting the effectiveness of the new design introduced in this paper.

3) Discussion on Parameter β: To maintain stability during training, we utilize a moving average scheme to process the aggregation weights in FedISM+ as follows:

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

In this section, we investigate the influence of the parameter β by experimenting with different values $\{0.3, 0.5, 0.7, 0.9,$
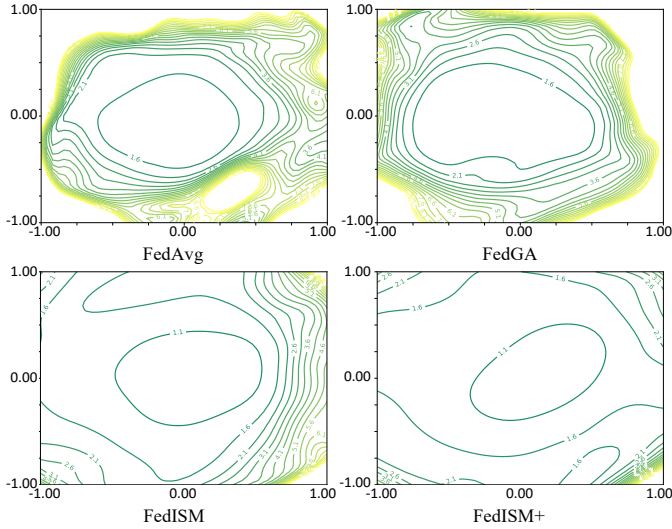
Fig. 7: Visualization of loss landscape on corrupted testing sets.

TABLE V: Communication overhead of each client in each round. 44.59MB is the size of ResNet-18 [44].

Method	FedAvg [1]	FedISM+
Overhead	2×44.59MB	2×44.59MB+4B (1.000000043×)

1.0}, using the setup described in Sec. IV-B. The results of these experiments are presented in the right column of Fig. 5, alongside the performance of FedAvg [1], the best fair optimization method, and FedISM [18] for a more thorough comparison. The figures show that, compared to cases where a moving average is either not applied or only weakly applied (*i.e.*, $\beta = 1.0$ or 0.9), incorporating a moderate moving average leads to better and more stable performance, often with a reduced standard deviation. Furthermore, the performance tends to be relatively unaffected by variations in β when the moving average is used.

4) Robustness across Different Shift Degrees: Our experiments also demonstrate the robustness of FedISM+, showcasing stable performance across various degrees of shift, specifically different ratios of corrupted clients. Quantitative results for two datasets at varying proportions (*i.e.*, 30%, 20%, 10% and 5%) of clients with Gaussian noise-corrupted images are illustrated in Fig. 6. The performance of FedAvg [1] reveals that reducing the ratio of corrupted clients complicates the maintenance of performance on corrupted images. FedISM consistently surpasses both FedAvg and previous fair optimization methods across all ratios. Notably, on both datasets, both the AUC on corrupted images and the average AUC for FedISM+ surpass those of FedISM, demonstrating the effectiveness of our new design.

5) Visualization of Loss landscape: Following [46], we visualize the loss landscape under model weight perturbation in Fig. 7. The experimental settings are the same as in Sec. IV-B. Compared with FedAvg [1] and FedGA [23], FedISM(+) converges to flatter regions, aligning with our aims and motivations. Notably, FedISM+ achieves smaller sharpness over a wider region compared to FedISM, as it considers more states.

6) Communication Overhead and Privacy: FedISM+ uploads sharpness/perturbed loss to the server, which can raise concerns regarding communication overhead and privacy. These concerns are addressed in this section. Regarding the former, we demonstrate that the additional communication overhead of FedISM+ is negligible, as shown in Tab. V, since only one extra float value needs to be uploaded. Concerning the latter, existing techniques in secure multi-party computation (*e.g.*, Homomorphic Encryption) ensure that we can compute the sum of all sharpness/perturbed loss (Eq. 16 and Eq. 17) without leaking any individual value [47]. This prevents third parties from inferring any information from the uploaded values.

V. CONCLUSION

In this paper, we explore how to address the fairness problem of FL raised by imaging quality shifts. Instead of promoting fairness through aligning a 0th- or 1st-order state of convergence across clients, we first propose capturing the full spectrum of convergence to build a better surrogate of fairness. Specifically, we generalize previous approaches by considering multiple states, and calculating sharpness/perturbed loss at different distances from zero up to a maximum. Building on this concept, we propose FedISM+, which effectively aligns the model's states across different clients. Comprehensive experiments conducted on the widely recognized RSNA ICH and ISIC 2019 datasets show that FedISM+ outperforms existing state-of-the-art fair FL approaches. We believe that our findings will inspire more studies on fair FL in medical applications and beyond, and generalization estimation with metrics on the training set.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017, pp. 1273–1282.

[2] Q. Dou, T. Y. So, M. Jiang, Q. Liu, V. Vardhanabhuti, G. Kaissis, Z. Li, W. Si, H. H. Lee, K. Yu, *et al.*, "Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study," *NPJ Digit. Medicine*, vol. 4, no. 1, p. 60, 2021.

[3] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous federated learning: State-of-the-art and research challenges," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–44, 2023.

[4] W. Huang, M. Ye, and B. Du, "Learn from others and be yourself in heterogeneous federated learning," in *CVPR*, 2022.

[5] W. Huang, M. Ye, Z. Shi, and B. Du, "Generalizable heterogeneous federated cross-correlation and instance similarity learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[6] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-iid features via local batch normalization," in *ICLR*, 2021.

[7] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *CVPR*, 2021, pp. 1013–1023.

[8] M. Jiang, H. Yang, C. Cheng, and Q. Dou, "IOP-FL: Inside-outside personalization for federated medical image segmentation," *IEEE Trans. Med. Imag.*, 2023.

[9] J. Zhang, Z. Li, B. Li, J. Xu, S. Wu, S. Ding, and C. Wu, "Federated learning with label distribution skew via logits calibration," in *ICML*, 2022.

[10] N. Wu, L. Yu, X. Yang, K. Cheng, and Z. Yan, "Fediic: Towards robust federated learning for class-imbalanced medical image classification," in *MICCAI*, 2023.

[11] N. Wu, L. Yu, X. Jiang, K. Cheng, and Z. Yan, "Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity," in *IJCAI*, 2023.

[12] Z. Chen, W. Li, X. Xing, and Y. Yuan, "Medical federated learning with joint graph purification for noisy label learning," *Medical Image Anal.*, vol. 90, p. 102976, 2023.

[13] N. Wu, Z. Sun, Z. Yan, and L. Yu, "Feda3i: Annotation quality-aware aggregation for federated medical image segmentation against heterogeneous annotation noise," in *AAAI*, 2024.

[14] J. Wicaksana, Z. Yan, D. Zhang, X. Huang, H. Wu, X. Yang, and K.-T. Cheng, "FedMix: Mixed supervised federated learning for medical image segmentation," *IEEE Trans. Med. Imag.*, 2023.

[15] X. Fang, M. Ye, and X. Yang, "Robust heterogeneous federated learning under data corruption," in *ICCV*, 2023.

[16] Z. Huang, M. Zhu, X. Xia, L. Shen, J. Yu, C. Gong, B. Han, B. Du, and T. Liu, "Robust generalization against photon-limited corruptions via worst-case sharpness minimization," in *CVPR*, 2023.

[17] J. Rawls, "Justice as fairness: A restatement," *Erin Kelly/Harvard University*, 2001.

[18] N. Wu, Z. Kuang, Z. Yan, and L. Yu, "From optimization to generalization: Fair federated learning against quality shift via inter-client sharpness matching," in *IJCAI*, 2024.

[19] M. Jiang, H. R. Roth, W. Li, D. Yang, C. Zhao, V. Nath, D. Xu, Q. Dou, and Z. Xu, "Fair federated medical image segmentation via client contribution estimation," in *CVPR*, 2023, pp. 16 302–16 311.

[20] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," in *ICLR*, 2020.

[21] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *ICML*, 2019, pp. 4615–4625.

[22] Y. H. Ezzeldin, S. Yan, C. He, E. Ferrara, and A. S. Avestimehr, "Fairfed: Enabling group fairness in federated learning," in *AAAI*, 2023.

[23] R. Zhang, Q. Xu, J. Yao, Y. Zhang, Q. Tian, and Y. Wang, "Federated domain generalization with generalization adjustment," in *CVPR*, 2023.

[24] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *ICLR*, 2021.

[25] J. Zhuang, B. Gong, L. Yuan, Y. Cui, H. Adam, N. Dvornek, S. Tatikonda, J. Duncan, and T. Liu, "Surrogate gap minimization improves sharpness-aware training," in *ICLR*, 2022.

[26] W. Huang, M. Ye, Z. Shi, G. Wan, H. Li, B. Du, and Q. Yang, "Federated learning for generalization, robustness, fairness: A survey and benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[27] L. Lyu, X. Xu, Q. Wang, and H. Yu, "Collaborative fairness in federated learning," *Federated Learning: Privacy and Incentive*, pp. 189–204, 2020.

[28] X. Xu, L. Lyu, X. Ma, C. Miao, C. S. Foo, and B. K. H. Low, "Gradient driven rewards to guarantee fairness in collaborative machine learning," in *NeurIPS*, 2021.

[29] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *ICLR*, 2020.

[30] Y. Zhou, Y. Qu, X. Xu, and H. Shen, "Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition," in *ICCV*, 2023, pp. 11 345–11 355.

[31] Z. Qu, X. Li, R. Duan, Y. Liu, B. Tang, and Z. Lu, "Generalized federated learning via sharpness aware minimization," in *ICML*, 2022.

[32] D. Caldarola, B. Caputo, and M. Ciccone, "Improving generalization in federated learning by seeking flat minima," in *ECCV*, 2022.

[33] Y. Sun, L. Shen, S. Chen, L. Ding, and D. Tao, "Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape," in *ICML*, 2023.

[34] A. Papadaki, N. Martinez, M. Bertran, G. Sapiro, and M. Rodrigues, "Minimax demographic group fairness in federated learning," in *ACM FAccT*, 2022, pp. 142–159.

[35] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *ICLR*, 2021.

[36] V. Vapnik, "Statistical learning theory," *John Wiley & Sons google schola*, vol. 2, pp. 831–842, 1998.

[37] M. Jiang, H. Yang, X. Li, Q. Liu, P.-A. Heng, and Q. Dou, "Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance," in *MICCAI*, 2022, pp. 196–206.

[38] A. E. Flanders, L. M. Prevedello, G. Shih, S. S. Halabi, J. Kalpathy-Cramer, R. Ball, J. T. Mongan, A. Stein, F. C. Kitamura, M. P. Lungren, *et al.*, "Construction of a machine learning dataset through collaboration: The RSNA 2019 brain CT hemorrhage challenge," *Radiology: Artificial Intelligence*, vol. 2, no. 3, 2020.

[39] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 1–9, 2018.

[40] N. C. 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*, 2018, pp. 168–172.

[41] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, *et al.*, "BCN20000: Dermoscopic lesions in the wild," *arXiv:1908.02288*, 2019.

[42] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "FedCorr: Multi-stage federated learning for label noise correction," in *CVPR*, 2022, pp. 10 184–10 193.

[43] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2019.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[45] W. Huang, M. Ye, Z. Shi, H. Li, and B. Du, "Rethinking federated learning with domain shift: A prototype view," in *CVPR*, 2023.

[46] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *NeurIPS*, vol. 31, 2018.

[47] Z. Shen, J. Cervino, H. Hassani, and A. Ribeiro, "An agnostic approach to federated learning with class imbalance," in *ICLR*, 2022.