HarmoFL: Harmonizing Local and Global Drifts in Federated Learning on Heterogeneous Medical Images

Meirui Jiang[1][*]**, Zirui Wang**[2][*]**, Qi Dou**[1]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong
[2] School of Biological Science and Medical Engineering, Beihang University

Abstract

Multiple medical institutions collaboratively training a model using federated learning (FL) has become a promising solution for maximizing the potential of data-driven models, yet the non-independent and identically distributed (non-iid) data in medical images is still an outstanding challenge in real-world practice. The feature heterogeneity caused by diverse scanners or protocols introduces a drift in the learning process, in both local (client) and global (server) optimizations, which harms the convergence as well as model performance. Many previous works have attempted to address the non-iid issue by tackling the drift locally or globally, but how to jointly solve the two essentially coupled drifts is still unclear. In this work, we concentrate on handling both local and global drifts and introduce a new harmonizing framework called HarmoFL. First, we propose to mitigate the local update drift by normalizing amplitudes of images transformed into the frequency domain to mimic a unified imaging setting, in order to generate a harmonized feature space across local clients. Second, based on harmonized features, we design a client weight perturbation guiding each local model to reach a flat optimum, where a neighborhood area of the local optimal solution has a uniformly low loss. Without any extra communication cost, the perturbation assists the global model to optimize towards a converged optimal solution by aggregating several local flat optima. We have theoretically analyzed the proposed method and empirically conducted extensive experiments on three medical image classification and segmentation tasks, showing that HarmoFL outperforms a set of recent state-of-the-art methods with promising convergence behavior. Code is available at: https://github.com/med-air/HarmoFL

Introduction

Multi-site collaborative training of deep networks is increasingly important for maximizing the potential of data-driven models in medical image analysis (Shilo, Rossman, and Segal 2020; Peiffer-Smadja et al. 2020; Dhruva et al. 2020), however, the data sharing is still restricted by some legal and ethical issues for protecting patient data. Federated learning recently allows a promising decentralized privacy-preserving solution, in which different institutions
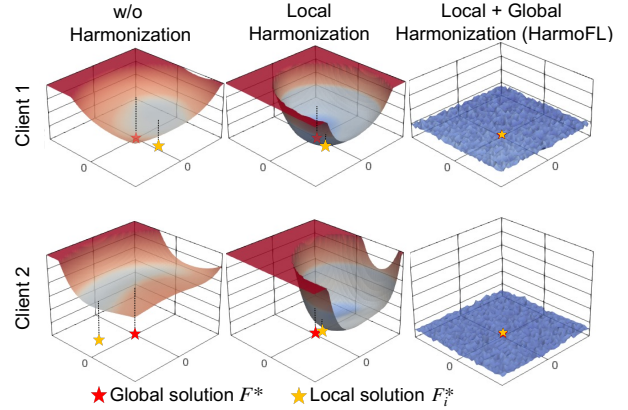

Figure 1: Loss landscape visualization of two clients without harmonization (left), with only local harmonization (middle) and our complete method of HarmoFL (right). The vertical axis shows the loss (denoting the solution for global objective as F^* and each local objective as F_i^*), and the horizontal plane represents a parameter space centered at the global model weight. (*See Introduction for detailed explanation.*)

can jointly train a model without actual data sharing, i.e., training client models locally and aggregating them globally (McMahan et al. 2017; Kaissis et al. 2020).

Despite the recent promising progress achieved by FL in medical image analysis (Dou et al. 2021; Rieke et al. 2020; Sheller et al. 2020; Roth et al. 2020; Ju et al. 2020), the non-independent and identically distributed (non-iid) data is still an outstanding challenge in real-world practice (Kairouz et al. 2019; Hsieh et al. 2020; Xu et al. 2021). Non-iid issue typically happens, and the device vendors or data acquisition protocols are responsible for heterogeneity in the feature distributions (Aubreville et al. 2020; Liu et al. 2020). For example, the appearance of histology images varies due to different staining situations, and MRI data of different hospitals suffer from feature distribution shifts associated with various scanners or imaging protocols.

Previous literature has demonstrated, both empirically and theoretically, that such data heterogeneity across clients introduces *drift* in both local (client) and global (server) optimizations, making the convergence slow and unstable (Zhao et al. 2018; Li et al. 2019; Karimireddy et al. 2020). Specif-

---

[*]Equal contribution. Corresponding: mrjiang@cse.cuhk.edu.hk

ically, in the local update, each client model will be optimized towards its own local optima (i.e., fitting its individual feature distribution) instead of solving the global objective, which raises a drift across client updates. Meanwhile, in the global update that aggregates these diverged local models, the server model is further distracted by the set of mismatching local optima, which subsequently leads to a global drift at the server model. Fig. 1 intuitively illustrates such local and global drifts via loss landscape visualization (Li et al. 2018), in an example of two non-iid clients. The vertical axis shows the loss at each client (denoting the solution for global objective as F^* and for each local objective as F_i^*), and the horizontal plane represents a parameter space centered at the specific parameters of global model . With the same objective function and parameter initialization for each client, we can see the local solution of two clients in the first column are significantly different, which indicates the drift across local client updates. Globally, a current good solution F^* may achieve a relatively low loss for both clients. However, since each client has its own shaped loss landscape, optimizing the current solution towards global optima is difficult and the aggregation of the two diverged local solutions further distracts the current good solution.

To address such non-iid problem in FL, existing works can be mainly divided into two groups, which correspond to tackling the drift locally or globally. The local side emphasizes how to better normalize those diverse data. For example, (Sheller et al. 2018) conducted a pioneer study and proposed using data pre-processing to reduce data heterogeneity. Recently, FedBN (Li et al. 2021) kept batch normalization layers locally to normalize local data distribution. As for the global server side, adjusting aggregation weight is a typical strategy, for instance, (Yeganeh et al. 2020) used inverse distance to adaptively reweight aggregation. Recently, the framework FedAdam (Reddi et al. 2021) is proposed for introducing adaptive optimization to stabilize server update in heterogeneous data. However, these methods tackle heterogeneity partially via either client or server update. The local and global drifts are essentially coupled, yet how to jointly solve them as a whole still remains unclear.

In this paper, we consider both client and server updates and propose a new harmonizing strategy to effectively solve the data heterogeneity problem for federated medical image analysis. First, to mitigate local drift at the local update, we propose a more effective normalization strategy via amplitude normalization, which unifies amplitude components from images decomposed in the frequency domain. Then for the global drift, although weighted parameter aggregation is widely adopted, the coordination can fail over heterogeneous clients with large parameter differences. Thus, we aim to promote client models that are easy to aggregate rather than design a new re-weighting strategy. Based on the harmonized feature space, we design a weight-perturbation strategy to reduce the server update drift. The perturbation is generated locally from gradients and applied on the client model to constrain a neighborhood area of the local converged model to have a uniformly low loss. With the perturbation, each client finds a shared flat optimal solution that can be directly aggregated with others, assisting the global

model to optimize towards a converged optimal solution. As can be observed from the last two columns of Fig. 1, the proposed amplitude normalization for local harmonization well mitigates the distance between global and local solutions, and with the weight perturbation for global harmonization, our approach achieves flat client solutions that can be directly aggregated to obtain an optimal global model.

Our main contributions are highlighted as follows:

- We propose to effectively mitigate the local update drift by normalizing the frequency-space amplitude component of different images into a unified space, which harmonizes non-iid features across clients.

- Based on the harmonized features, we further design a novel weight-perturbation based strategy to rectify global server update shift without extra communication cost.

- To the best of our knowledge, we are the first to simultaneously address both local and global update drifts for federated learning on heterogeneous medical images. We have also theoretically analyzed the proposed HarmoFL framework from the aspect of gradients similarity, showing that drift caused by data heterogeneity is bounded.

- We conduct extensive experiments on three medical image tasks, including breast cancer histology image classification, histology nuclei segmentation, and prostate MRI segmentation. Our HarmoFL significantly outperforms a set of latest state-of-the-art FL methods.

Related Work

There have been many methods proposed trying to improve federated learning on heterogeneous data and can be mainly divided into two groups, including improvements on local client training and on global server aggregation.

Local client training: Literature towards improving the client training to tackle local drift includes, using data pre-processing to reduce data heterogeneity (Sheller et al. 2018), introducing a domain loss for heterogeneous electroencephalography classification (Gao et al. 2019), using simulated CT volumes to mitigate scanner differences in cardiac segmentation (Li et al. 2020a), adding a proximal term to penalize client update towards smaller weight differences between client and server model (Li et al. 2020b), and learning affine transformation against shift under the assumption that data follows an affine distribution (Reisizadeh et al. 2020). Recently, both SiloBN (Andreux et al. 2020) and FedBN (Li et al. 2021) propose keeping batch normalization locally to help clients obtain similar feature distributions, and MOON (Li, He, and Song 2021) uses contrastive learning on latent feature representations at the client-side to enhance the agreements between local and global models.

Global server aggregation: Besides designs on the client-side, many methods are proposed towards reducing global drift with a focus on the server-side. (Zhao et al. 2018) create a small subset of data that is globally shared across clients, and many other methods improve the aggregation strategy, e.g. (Yeganeh et al. 2020) calculate the inverse distance to re-weight aggregation and FedNova (Wang et al. 2020) proposes to use normalized stochastic gradients to perform

global model aggregation rather than the cumulative raw local gradient changes. Very recently, FedAdam (Reddi et al. 2021) introduces adaptive optimization into federated learning to stabilize convergence in heterogeneous data.

However, the above-mentioned methods partially address the drift either from the local client or global server perspective. Instead, our approach aims jointly mitigating the two coupled drifts both locally and globally.

Methodology

To address the non-iid issues from both local and global aspects, we propose an effective new federated learning framework of HarmoFL. We start with the formulation of federated heterogeneous medical images analysis, then describe amplitude normalization and weight perturbation towards reducing local drift and global drift respectively. At last, we give a theoretical analysis of HarmoFL.

Preliminaries

Denote $(\mathcal{X}, \mathcal{Y})$ as the joint image and label space over N clients. A data sample is an image-label pair (x, y) with $x \in \mathcal{X}, y \in \mathcal{Y}$, and the data sampled from a specific i-th client follows data distribution $\mathcal{D}_i$. In this work, we focus on the non-iid feature shift. Given the joint probability $P(\mathbf{x}, y)$ of feature $\mathbf{x}$ of image x and label y, we have $P_i(\mathbf{x})$ varies even if $P(y \,|\, \mathbf{x})$ is the same or $P_i(\mathbf{x} \,|\, y)$ varies across clients while $P(y)$ is unchanged.

Our proposed HarmoFL aims to improve federated learning for both local client training and global aggregation. The federated optimization objective is formulated as follows:

$$\min_\theta \left[F(\theta) := \sum_{i=1}^{N} p_i F_i(\theta + \delta, \overline{\mathcal{D}_i}) \right], \tag{1}$$

where $F_i = \sum_{(x,y) \sim \mathcal{D}_i} \ell_i(\Psi(x), y; \theta + \delta)$ is the local objective function and ℓ_i is the loss function defined by the learned model θ and sampled pair (x, y). For the i-th client, p_i is the corresponding weight such that $p_i \geq 0$ and $\sum_{i=1}^{N} p_i = 1$. The δ is a weight perturbation term and $\overline{\mathcal{D}_i}$ is the harmonized feature distribution obtained by our amplitude normalization. Specifically, we propose a new amplitude normalization operator $\Psi(\cdot)$ harmonizing various client features distribution to mitigate client update drift. The normalizer manipulates amplitudes of data in the frequency space without violating the local original data preserving. Based on the harmonized features, we can generate a weight perturbation for each client without any extra communication cost. The perturbation forces each client reaching a uniformly low error in a neighborhood area of local optima, thus reducing the drift in the server.

Amplitude Normalization for Local Training

Structure semantic is an important factor to support medical imaging analysis while low-level statistics (e.g. color, contrast) usually help to differentiate the structure. By decomposing the two parts, we are able to harmonize low-level features across hospitals while preserving critical structures. Using the fast Fourier transform (Nussbaumer 1981), we can
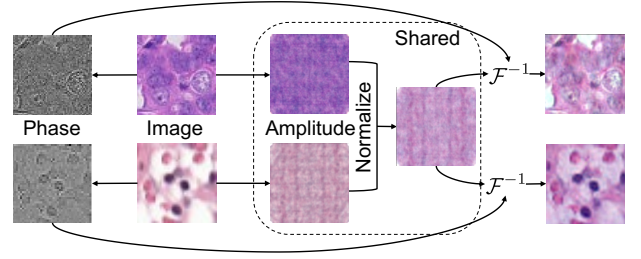

Figure 2: Amplitude normalization that harmonizes local client features. Phase components are strictly kept locally and only the average amplitude from each client is shared.

transform images into the frequency space signals and decompose out the amplitude spectrum which captures low-level features. Considering the patient level privacy sensitivity and forbiddance of sharing original images, we propose only using averaged amplitude term during communication to conduct the normalization.

More specifically, for input image x from the i-th client, we transform each channel of image $x \in \mathbb{R}^{H \times W \times C}$ into the frequency space signals $\mathcal{F}_i(x)(u, v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(h, w) e^{-j 2\pi \left(\frac{h}{H} u + \frac{w}{W} v \right)}$. Then we can split the real part $\mathcal{R}_i(x)$ and the imaginary part $\mathcal{I}_i(x)$ from the frequency signals $\mathcal{F}_i(x)$. The amplitude and phase component can be expressed as:

$$\begin{aligned} \mathcal{A}_i(x) &= \left[\mathcal{R}_i^2(x) + \mathcal{I}_i^2(x) \right]^{1/2} \\ \mathcal{P}_i(x) &= \arctan \left[\mathcal{I}_i(x) / \mathcal{R}_i(x) \right]. \end{aligned} \tag{2}$$

Next, we normalize the amplitude component of each image batch-wise with a moving average term. For the k-th batch of M sampled images, first we decompose the amplitude $\mathcal{A}_{i,x_m}$ and phase $\mathcal{P}_{i,x_m}$ for each single image x_m in the batch. We calculate the in-batch average amplitude and update the average amplitude with a decay factor v:

$$\overline{\mathcal{A}}_{i,k} = (1-v)\overline{\mathcal{A}}_{i,k-1} + v \frac{1}{M} \sum_{m=1}^{M} \mathcal{A}_{i,x_m}, \tag{3}$$

where $\overline{\mathcal{A}}_{i,k-1}$ is the average amplitude calculated from the previous batch and this term is set to zero for the first batch. The $\overline{\mathcal{A}}_{i,k}$ harmonizes low-level distributions inside a client and keeps tracking amplitudes of the whole client images. With the updated average term, each original image of the k-th batch is normalized using the average amplitude $\overline{\mathcal{A}}_{i,k}$ and its original phase component $\mathcal{P}_{i,x_m}$ in below:

$$\Psi(x_m) = \mathcal{F}^{-1}(\overline{\mathcal{A}}_{i,k}, \mathcal{P}_{i,x_m}), \tag{4}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform. After client training, the amplitude normalization layer which only contains the average amplitude information will be sent to the central server and generate a global amplitude. This global amplitude compromises various low-level visual features across clients and can help reduce client update drift at the next federated round. In practice, we find that only communication at the first round and fix the global amplitude can well harmonize non-iid features as well as saving the communication cost.

Weight Perturbation for Global Aggregation

The above-proposed amplitude normalization allows each client to optimize within a harmonized feature space, mitigating the local drift. Based on the harmonized feature space, we further aim to promote client models that are easy to aggregate, thus rectifying the global drift. As the global aggregation is typically the weighted averaging, a client with flat optima can well coordinate with other clients than sharp local minima (Keskar et al. 2016), which shows large error increases even with minor parameter change. Motivated from adversarial training, we propose a local optimization objective of each client as below:

$$\min_{\theta} \sum_{(x,y)\sim \mathcal{D}_i} \max_{\|x'-x\|_p \leq \delta} \ell(x',y;\theta), \qquad (5)$$

where x' is the adversarial image within a L_p-norm bounded δ-ball centered at original image x, the adversarial image is generated with the same label of x but different feature shift. However, the generating process has an extra communication burden since we have to transfer feature distribution information across clients. Based on the harmonized features, we carefully design a weight perturbation to effectively solve the Eq. (5). With a little abuse of notation, instead of generating adversarial images bonded by the term δ, we propose a new δ as a perturbation that is directly applied to model parameters. The perturbation is self-generated using gradients from the harmonized feature and has no extra communication cost. Formally, for the k-th batch, we first calculate the gradients of client model $\theta_{i,k-1}$ on the amplitude normalized feature $\Psi(x)$, which is calculated from Eq. (4). Then we use the Euclidean norm $\|\cdot\|_2$ to normalize the gradients and obtain the perturbation term for the current iteration of gradient descent:

$$\delta_k = \alpha \frac{\nabla \ell_i(\Psi(x),y;\theta_{i,k-1})}{\|\ell_i(\Psi(x),y;\theta_{i,k-1})\|_2}, \qquad (6)$$

where α is a hyper-parameter to control the degree of perturbation. The flat area around the local client optimum can be expanded with the larger α, and the choice of α is studied in the experiment section. After obtaining the perturbation term δ_k, we minimize the loss on the parameter-perturbated model as below:

$$\theta_{i,k} \leftarrow \theta_{i,k-1} - \eta_l \nabla \ell_i(\Psi(x),y;\theta_{i,k-1}+\delta_k), \qquad (7)$$

where $\theta_{i,k-1}$ is the model from the previous batch and η_l is the local client learning rate. After iteratively update, each local client model is gradually driven towards an optimal solution that holds a uniformly low loss around its neighborhood area (i.e. flat optimum), thus promoting the aggregation and getting rid of trapped into the local sharp minima.

Theoretical Analysis for HarmoFL

With the help of amplitude normalization and weight perturbation, we constrain the client update with less dissimilarity and achieve a model having a relatively low error with a range of parameter changes. To theoretically analyze HarmoFL, we transform our approach into a simplified setting where we interpret both the amplitude normalization

---

Algorithm 1: Harmonizing Local and Global Drifts in FL

Input: communication rounds T, number of clients N, mini-batch steps K, client learning rate η_l, global learning rate η_g, hyper-parameter α

Output: The final global model $\theta^{(T)}$

1: Initialize server model $\theta^{(1)}$
2: **for** $t = 1, 2, \cdots, T$ **do**
3:   **for** $i = 1, 2, \cdots, N$ **in parallel do**
4:     $\theta_{i,1}^t \leftarrow \theta^t$ ▷ send the global model θ^t to client i
5:     **for** $k = 1, 2, \cdots, K$ **do**   ▷ client training
6:       sample a batch of data pairs (x,y) of $\mathcal{D}_i$
7:       $\overline{\mathcal{A}}_{i,k}^t = (1-v)\overline{\mathcal{A}}_{i,k-1}^t$
8:         $+v\frac{1}{M}\sum_{m=1}^{M}\mathcal{A}_{x_m}$ ▷ Eq. (3)
9:       $\Psi(x) = \mathcal{F}^{-1}(\overline{\mathcal{A}}_{i,k}^t, \mathcal{P}_{x_m})$ ▷ Eq. (4)
10:      $\delta_k = \alpha \frac{\nabla \ell_i(\Psi(x),y;\theta)}{\|\ell_i(\Psi(x),y;\theta)\|_2}$ ▷ Eq. (6)
11:      $g_\delta = \nabla \ell_i(\Psi(x),y;\theta_i^t + \delta_k)$
12:      $\theta_{i,k+1}^t \leftarrow \theta_{i,k}^t - \eta_l g_\delta$ ▷ Eq. (7)
13:    **end for**
14:    **return** $\theta_{i,K}^t$ ▷ send client model to server
15:   **end for**
16:   $\theta^{t+1} \leftarrow \theta^t - \eta_g \sum_{i=1}^{N} p_i(\theta_{i,K}^t - \theta_{i,1}^t)$
17:   $\overline{\mathcal{A}}^{t+1} = \frac{1}{N}\sum_{i=1}^{N}\overline{\mathcal{A}}_{i,K}^t$
18: **end for**
19: **return** $\theta^{(T)}$

---

and weight perturbation from the gradients perspective. For amplitude normalization, it reduces gradients dissimilarity between clients by harmonizing the non-iid features. In the meanwhile, the weight perturbation forces clients to achieve flat optima, in which the loss variation is constrained, making the gradients change mildly. So both amplitude normalization and weight perturbation bound gradient differences across clients, and this assumption has also been widely explored in different forms (Yin et al. 2018; Li et al. 2020b; Vaswani, Bach, and Schmidt 2019; Karimireddy et al. 2019). Based on the standard federated optimization objective, we formulate a new form in below:

$$\min_{\theta} \left[F(\theta) := \sum_{i=1}^{N} p_i F_i(\theta, \mathcal{D}_i) \right]$$

$$s.t. \sum_{\substack{(x_i,y_i)\sim\mathcal{D}_i, \\ (x_j,y_j)\sim\mathcal{D}_j}} |\nabla \ell(x_i,y_i;\theta) - \nabla \ell(x_j,y_j;\theta)| \leq \epsilon, \qquad (8)$$

where $j \in \{1, \cdots, N\}\backslash\{i\}$ and ϵ is a non-negative constant and (x_i, y_i) is the image-label pair from the i-th client's distribution. To quantify the overall drift between client and server models at the t-th communication round, we define the overall drift term as below:

$$\Gamma = \frac{1}{KN} \sum_{k=1}^{K} \sum_{i=1}^{N} \mathbb{E}[\|\theta_{i,k}^t - \theta^t\|^2], \qquad (9)$$

where K is the mini-batch steps. We theoretically show that with the bounded gradient difference, our HarmoFL strategy

is guaranteed to have an upper bound on the overall drift caused by non-iid data.

Theorem 0.1 *With the shift term Γ defined in Eq. (9), assume the gradient dissimilarity and variance are bounded and the functions F_i are β-smooth, denote the effective step-size $\tilde{\eta} = K\eta_g\eta_l$, we have the upper bound for the overall drift of our HarmoFL below:*

$$\Gamma \leq \frac{1}{N}\sum_{i=1}^{N}\frac{4\tilde{\eta}^2}{\eta_g^2}\|\nabla F_i(\theta)\|^2 + \frac{4\tilde{\eta}^2\epsilon^2(N-1)^2}{\eta_g^2 N^2} + \frac{2\tilde{\eta}^2\sigma^2}{K\eta_g^2}$$

$$\leq \frac{4\tilde{\eta}^2(G^2 + B^2C^2)}{\eta_g^2} + \frac{4\tilde{\eta}^2\epsilon^2(N-1)^2}{\eta_g^2 N^2} + \frac{2\tilde{\eta}^2\sigma^2}{K\eta_g^2},$$

where $C := F(\theta^0) - F^*$ when functions F_i are convex, and for non-convex situation, we have $C := \|\nabla F(\theta)\|$. This theorem gives the upper bound with both the convex and non-convex assumptions for F_i.

Please find the notation table in Appendix A. All assumptions and proofs are formally given in Appendix B. The proof sketch is applying our extra gradient differences constraint in the subproblem of one round optimization and using the bounded dissimilarity. The shift bound is finally unrolled from a recursion format of one round progress.

Experiments

In this section, we extensively evaluate our method to demonstrate that harmonizing local and global drifts are beneficial for clients with heterogeneous features. Our harmonizing strategy, HarmoFL, achieves higher performance as well as more stable convergence compared with other methods on feature non-iid datasets. This is shown on the breast cancer histology image classification, histology nuclei segmentation, and prostate MRI segmentation. All results reported are the average of three repeating runs with a standard deviation of different random seeds. More results please refer to Appendix C.

Dataset and Experimental Settings

Breast cancer histology image classification. We use the public tumor dataset Camelyon17, which contains 450,000 histology images with different stains from 5 different hospitals (Bandi et al. 2018). As shown in Fig. 3, we take each hospital as a single client, and images from different clients have heterogeneous appearances but share the same label distribution (i.e. normal and tumor tissues). We use a deep network of DenseNet121 (Huang et al. 2017) and train the model for 100 epochs at the client-side with different communication frequencies. We use cross-entropy loss and SGD optimizer with a learning rate of $1e^{-3}$.

Histology nuclei segmentation. For the cell nuclei segmentation, we gather three public datasets, including MoNuSAC2020 (Verma et al. 2021), MoNuSAC2018 (Kumar et al. 2019) and TNBC (Naylor et al. 2018). For data from MoNuSAC2020, we divide them into 4 clients according to different hospitals they come from and form 6 clients in total. We use U-Net (Ronneberger, Fischer, and Brox 2015) and train the model for 500 communication rounds

with 1 local update epoch for each communication round. We use segmentation Dice coefficient (Dice) loss and SGD optimizer with a learning rate of $1e^{-4}$.

Prostate MRI segmentation. For the prostate segmentation, we use a multi-site prostate segmentation dataset (Liu et al. 2020) which contains 6 different data sources from 3 public datasets (Nicholas et al. 2015; Lemaître et al. 2015; Litjens et al. 2014). We regard each data source as a client and train the U-Net using Adam optimizer with a learning rate of $1e^{-4}$, momentum of 0.9 and 0.99.

We report the performance of global models, i.e., the final results of our overall framework. The model was selected using the separate validation set and evaluated on the testing set. To avoid distracting focus on the feature non-iid problem due to data imbalance, we truncate the sample size of each client to their respective smaller number in histology image classification and prostate MRI segmentation task, but we keep the client data imbalance in the nuclei segmentation task to demonstrate the performance with data quantity difference. If not specified, our default setting for the local update epoch is 1. We use the momentum of 0.9 and weight decay of $1e^{-4}$ for all optimizers in three tasks. We empirically set the decay factor v to 0.1 and set the degree of perturbation term to $5e^{-2}$ by grid search. For more dataset and implementation details, please refer to Appendix C.
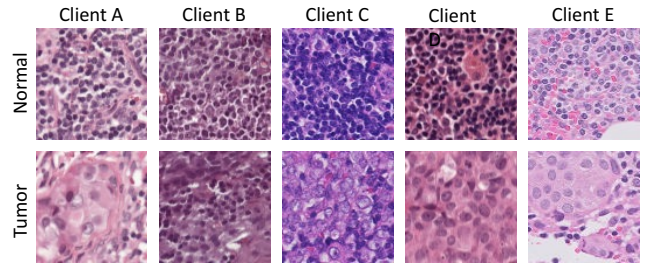

Figure 3: Examples of breast histology images of normal and tumor tissues from five clients, showing large heterogeneity.

Comparison with the State-of-the-arts

We compare our approach with recent state-of-the-art (SOTA) FL methods towards solving the non-iid problem. For local drifts, **FedBN** (Li et al. 2021) focuses on the non-iid feature shift with medical image applications, and both **FedProx** (Li et al. 2020b) and a recent method **MOON** (Li, He, and Song 2021) tackle the non-iid problem by constraining the dissimilarity between local and global models to reduce global aggregation shifts. **FedAdam** (Reddi et al. 2021) and **FedNova** (Wang et al. 2020) are proposed as general methods to tackle global drifts. For the breast cancer histology image classification shown in Table 2, we report the testing accuracy on five different clients and the average results. FedProx only achieves minor improvements than FedAvg, showing that only reducing global aggregation drift may not achieve promising results when local clients shifted severely. Another recent representation dissimilarity constraining method, MOON, boosts the accuracy on clients B, C, D but suffers from a large drop on client E. The reason

Method	Histology Nuclei Segmentation (Dice %)							Prostate MRI Segmentation (Dice %)						
	A	B	C	D	E	F	Avg.	A	B	C	D	E	F	Avg.
FedAvg (PMLR2017)	73.44 (0.02)	73.06 (0.24)	72.52 (0.85)	68.91 (0.34)	67.33 (0.86)	49.69 (0.34)	67.49 (9.06)	90.04 (1.27)	94.31 (0.28)	92.60 (0.66)	92.21 (0.71)	90.14 (0.27)	89.36 (1.76)	91.44 (1.91)
FedProx (MLSys2020)	73.49 (0.07)	73.11 (0.19)	72.45 (0.94)	69.01 (0.34)	67.33 (0.86)	49.56 (0.34)	67.49 (9.12)	90.65 (1.95)	94.60 (0.30)	92.64 (1.03)	92.19 (0.15)	89.36 (0.97)	87.07 (1.53)	91.08 (2.66)
FedNova (NeurIPS2020)	73.40 (0.05)	73.01 (0.38)	71.50 (1.35)	69.23 (0.34)	67.46 (1.07)	50.68 (0.34)	67.55 (8.57)	90.73 (0.41)	94.26 (0.08)	92.73 (1.29)	91.91 (0.61)	90.01 (0.87)	89.94 (1.54)	91.60 (1.70)
FedAdam (ICLR2021)	73.53 (0.08)	72.91 (0.24)	71.74 (1.33)	69.26 (0.50)	66.69 (1.18)	49.72 (0.11)	67.31 (8.98)	90.02 (0.29)	94.84 (0.11)	93.30 (0.79)	91.70 (0.16)	90.17 (1.46)	87.77 (1.35)	91.30 (2.53)
FedBN (ICLR2021)	72.50 (0.81)	72.51 (0.13)	74.25 (0.28)	64.84 (0.93)	68.39 (1.13)	69.11 (0.94)	70.27 (3.47)	92.68 (0.52)	94.83 (0.47)	93.77 (0.41)	92.32 (0.19)	93.20 (0.45)	89.68 (0.60)	92.75 (1.74)
MOON (CVPR2021)	72.85 (0.46)	71.92 (0.37)	69.23 (2.29)	69.00 (0.71)	65.08 (0.73)	48.26 (0.66)	66.06 (9.13)	91.79 (1.64)	93.63 (0.21)	93.01 (0.75)	92.61 (0.53)	91.22 (0.61)	91.14 (0.88)	92.23 (1.01)
HarmoFL (**Ours**)	**74.98 (0.36)**	**75.21 (0.57)**	**76.63 (0.20)**	**76.59 (0.77)**	**73.94 (0.13)**	**69.20 (1.23)**	**74.42 (2.76)**	**94.06 (0.47)**	**95.26 (0.38)**	**95.28 (0.33)**	**93.51 (0.79)**	**94.05 (0.50)**	**93.53 (1.02)**	**94.28 (0.80)**

Table 1: Results for histology nuclei segmentation and prostate MRI segmentation. The results of the Dice coefficient are reported. Each column represents one client and the Avg. is abbreviated for the average Dice.

Method	Breast Cancer Histology Image Classification (Accuracy %)					
	A	B	C	D	E	Avg.
FedAvg (PMLR2017)	91.10 (0.46)	83.12 (1.58)	82.06 (8.52)	87.49 (2.49)	74.78 (3.19)	83.71 (6.16)
FedProx (MLSys2020)	91.03 (0.50)	82.88 (1.63)	82.78 (8.56)	87.07 (1.76)	74.93 (3.05)	83.74 (5.99)
FedNova (NeurIPS2020)	90.99 (0.54)	82.97 (1.76)	82.40 (9.21)	86.93 (1.58)	74.86 (3.12)	83.61 (6.00)
FedAdam (ICLR2021)	87.45 (0.77)	80.38 (2.03)	76.89 (14.03)	89.27 (1.28)	77.86 (2.68)	82.37 (5.65)
FedBN (ICLR2021)	89.35 (8.50)	90.25 (1.66)	94.16 (1.00)	94.04 (2.32)	68.87 (22.14)	87.33 (10.55)
MOON (CVPR2021)	88.92 (1.54)	83.52 (0.31)	84.71 (5.14)	90.02 (1.56)	67.79 (2.06)	82.99 (8.93)
HarmoFL (**Ours**)	**96.17 (0.56)**	**93.60 (0.67)**	**95.54 (0.32)**	**95.58 (0.27)**	**96.50 (0.46)**	**95.48 (1.13)**

Table 2: Results for breast cancer histology images classification of different methods. Each column represents one client and the Avg. is abbreviated for the average accuracy.

may come from that images in client E appear differently as shown in Fig. 3, making the representations of client E failed to be constrained towards other clients. With the harmonized strategy reducing both local and global drifts, our method consistently outperforms others, reaching an accuracy of 95.48% on average, which is 8% higher than the previous SOTA (FedBN) for the feature non-iid problem. Besides, our method can help all heterogeneous clients benefit from the federated learning, where clients show a testing accuracy with a small standard deviation on average, while other methods show larger variance across clients.

For segmentation tasks, the experimental results of Dice are shown in Table 1 in the form of single client and average performance. On histology nuclei segmentation, HarmoFL significantly outperforms the SOTA method of FedBN and improves at least 4% on mean accuracy compared with all other methods. On the prostate segmentation task, as MRI images show fewer non-iid feature shifts than histology images, the performance gap of each client is not as large as nuclei segmentation. However, our method still consistently achieves the highest Dice of 94.28% and has a smaller standard deviation over different clients. Besides, we visualize the segmentation results to demonstrate a qualitative comparison, as shown in Fig. 4. Comparing with the first ground-truth column, due to the heterogeneous features, other federated learning methods either cover more or fewer areas in both prostate MRI and histology images. As can be observed from the second and fourth row, the heterogeneity in features also makes other methods fail to obtain an accurate boundary. But with the proposed harmonizing strategy, our approach shows more accurate and smooth boundaries.

Ablation Study

We further conduct ablation study based on the breast cancer histology image classification to investigate key properties of our HarmoFL, including the convergence analysis, influence of different local update epochs, and the effects of weight perturbation degree.

Convergence analysis. To demonstrate the effectiveness of our method on reducing local update and server update drifts, we plot the testing accuracy curve on the average of five clients for 100 communication rounds with 1 local update epoch. As shown in Fig. 5(a), the curve of HarmoFL increases smoothly with communication rounds increasing, while the state-of-the-art method of FedBN (Li et al. 2021), shows unstable convergence as well as lower accuracy. From the curve of FedBN (Li et al. 2021), we can see the there are almost no improvements at the first 10 communication rounds, the potential reason is that the batch normalization layers in each local client model are not fully trained to normalize feature distributions.

Influence of local update epochs. Aggregating with different frequencies may affect the learning behavior,
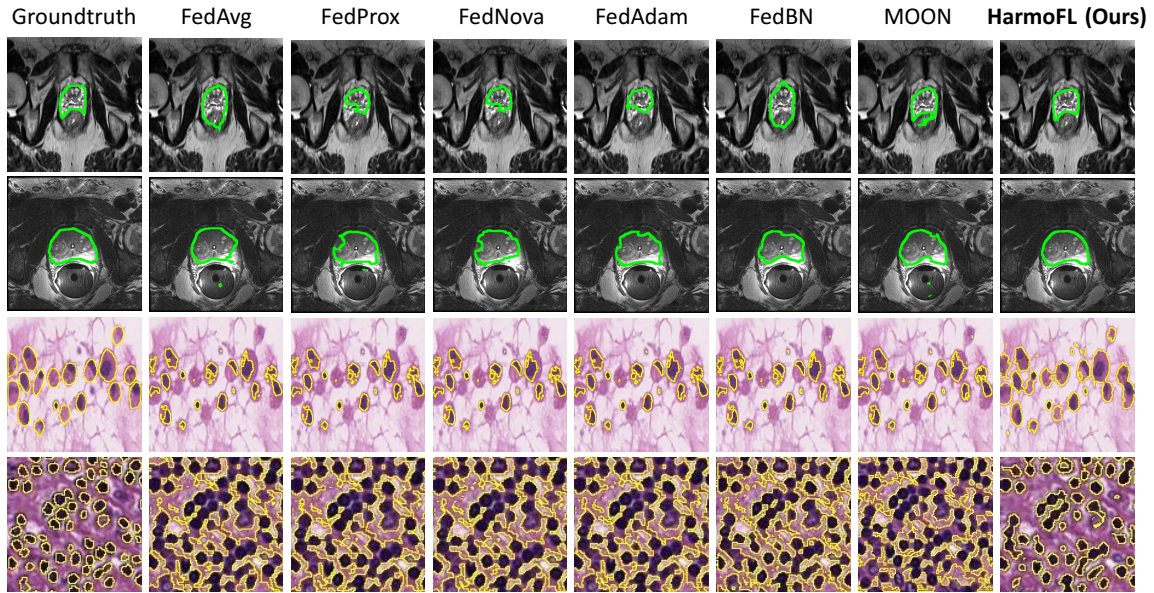
Figure 4: Qualitative comparison on segmentation results with our method and other state-of-the-art methods. Top two rows for the task of prostate MRI segmentation and the bottom two rows for the task of histology nuclei segmentation.
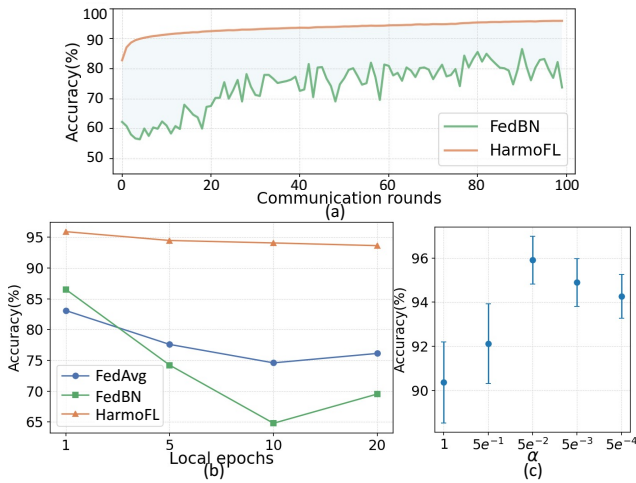

Figure 5: (a) Convergence in terms of testing accuracy with communication rounds. (b) Comparison of FedAvg, FedBN, and our HarmoFL with different local training epochs. (c) Performance of HarmoFL with different perturbation radius.

significantly outperforms other approaches and shows robustness to larger drift brought by more local update epochs.

Effects of weight perturbation degree. We further analyze how the weight perturbation degree control hyper-parameter α affects the performance of our method. Intuitively, the value of α indicates the radius of the flat optima area. A small radius may hinder clients to find a shared flat area during aggregation, while a large radius creates difficulties in the optimization of reaching such a flat optimum. As shown in Fig 5(c), we plot the average testing accuracy with standard error across clients by searching $\alpha \in \{1, 5e^{-1}, 5e^{-2}, 5e^{-3}, 5e^{-4}\}$. Our method reaches the highest accuracy when $\alpha = 5e^{-2}$ and the performance decreases with α reducing. Besides, we can see our method can achieve more than 90% accuracy even without degree control (i.e. $\alpha = 1$).

Conclusion

This work proposes a novel harmonizing strategy, HarmoFL, which uses amplitude normalization and weight perturbation to tackle the drifts that exist in both local client and global server. Our solution gives inspiration on simultaneously solving the essentially coupled local and global drifts in FL, instead of regarding each drift as a separate issue. We conduct extensive experiments on heterogeneous medical images, including one classification task and two segmentation tasks, and demonstrates the effectiveness of our approach consistently. We further provide theoretical analysis to support the empirical results by showing the overall non-iid drift caused by data heterogeneity is bounded in our proposed HarmoFL. Overall, our work is beneficial to promote wider impact of FL in real world medical applications.

since less frequent communication will further enhance the drift due to the non-iid feature, and finally obtaining a worse global model. We study the effectiveness of HarmoFL with different local update epochs and the results are shown in Fig. 5(b). When the local epoch is 1, each client communicates with others frequently, all methods have a relatively high testing accuracy. With more local epochs added, the local client update drift is increased and the differences between local and global models become larger. Both FedAvg and FedBN suffer from large drifts and show severe performance drops. However, because our method enables each client to train the model using weight perturbation with harmonized features, HarmoFL

References

Andreux, M.; du Terrail, J. O.; Beguier, C.; and Tramel, E. W. 2020. Siloed federated learning for multi-centric histopathology datasets. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 129–139. Springer.

Aubreville, M.; Bertram, C. A.; Donovan, T. A.; Marzahl, C.; Maier, A.; and Klopfleisch, R. 2020. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Scientific data*, 7(1): 1–10.

Bandi, P.; Geessink, O.; Manson, Q.; Van Dijk, M.; Balkenhol, M.; Hermsen, M.; Bejnordi, B. E.; Lee, B.; Paeng, K.; Zhong, A.; et al. 2018. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*.

Dhruva, S. S.; Ross, J. S.; Akar, J. G.; Caldwell, B.; Childers, K.; Chow, W.; Ciaccio, L.; Coplan, P.; Dong, J.; Dykhoff, H. J.; et al. 2020. Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. *NPJ digital medicine*, 3(1): 1–9.

Dou, Q.; So, T. Y.; Jiang, M.; Liu, Q.; Vardhanabhuti, V.; Kaissis, G.; Li, Z.; Si, W.; Lee, H. H.; Yu, K.; et al. 2021. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *NPJ digital medicine*, 4(1): 1–11.

Gao, D.; Ju, C.; Wei, X.; Liu, Y.; Chen, T.; and Yang, Q. 2019. Hhhfl: Hierarchical heterogeneous horizontal federated learning for electroencephalography. *arXiv preprint arXiv:1909.05784*.

Hsieh, K.; Phanishayee, A.; Mutlu, O.; and Gibbons, P. 2020. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, 4387–4398. PMLR.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Ju, C.; Gao, D.; Mane, R.; Tan, B.; Liu, Y.; and Guan, C. 2020. Federated transfer learning for eeg signal classification. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 3040–3045. IEEE.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2019. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*.

Kaissis, G. A.; Makowski, M. R.; Rückert, D.; and Braren, R. F. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6): 305–311.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 5132–5143. PMLR.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S. J.; Stich, S. U.; and Suresh, A. T. 2019. SCAFFOLD: Stochastic Controlled Averaging for On-Device Federated Learning.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kumar, N.; Verma, R.; Anand, D.; Zhou, Y.; Onder, O. F.; Tsougenis, E.; Chen, H.; Heng, P.-A.; Li, J.; Hu, Z.; et al. 2019. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 39(5): 1380–1391.

Lemaître, G.; Martí, R.; Freixenet, J.; Vilanova, J. C.; Walker, P. M.; and Meriaudeau, F. 2015. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Computers in biology and medicine*, 60: 8–31.

Li, D.; Kar, A.; Ravikumar, N.; Frangi, A. F.; and Fidler, S. 2020a. Federated Simulation for Medical Imaging. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 159–168.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the Loss Landscape of Neural Nets. In *Neural Information Processing Systems*.

Li, Q.; He, B.; and Song, D. 2021. Model-Contrastive Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020b. Federated optimization in heterogeneous networks. In *Conference on Machine Learning and Systems*.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2019. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; and Dou, Q. 2021. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In *International Conference on Learning Representations*.

Litjens, G.; Toth, R.; van de Ven, W.; Hoeks, C.; Kerkstra, S.; van Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J.; et al. 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical image analysis*, 18(2): 359–373.

Liu, Q.; Dou, Q.; Yu, L.; and Heng, P. A. 2020. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*.

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

Naylor, P.; Laé, M.; Reyal, F.; and Walter, T. 2018. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38(2): 448–459.

Nicholas, B.; Anant, M.; Henkjan, H.; John, F.; Justin, K.; et al. 2015. Nci-proc. ieee-isbi conf. 2013 challenge: Automated segmentation of prostate structures. *The Cancer Imaging Archive*.

Nussbaumer, H. J. 1981. The fast Fourier transform. In *Fast Fourier Transform and Convolution Algorithms*, 80–111. Springer.

Peiffer-Smadja, N.; Maatoug, R.; Lescure, F.-X.; D'ortenzio, E.; Pineau, J.; and King, J.-R. 2020. Machine learning for COVID-19 needs global collaboration and data-sharing. *Nature Machine Intelligence*, 2(6): 293–294.

Reddi, S. J.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; and McMahan, H. B. 2021. Adaptive Federated Optimization. In *International Conference on Learning Representations*.

Reisizadeh, A.; Farnia, F.; Pedarsani, R.; and Jadbabaie, A. 2020. Robust federated learning: The case of affine distribution shifts. *arXiv preprint arXiv:2006.08907*.

Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H. R.; Albarqouni, S.; Bakas, S.; Galtier, M. N.; Landman, B. A.; Maier-Hein, K.; et al. 2020. The future of digital health with federated learning. *NPJ digital medicine*, 3(1): 1–7.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Roth, H. R.; Chang, K.; Singh, P.; Neumark, N.; Li, W.; Gupta, V.; Gupta, S.; Qu, L.; Ihsani, A.; Bizzo, B. C.; et al. 2020. Federated learning for breast density classification: A real-world implementation. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 181–191. Springer.

Sheller, M. J.; Edwards, B.; Reina, G. A.; Martin, J.; Pati, S.; Kotrotsou, A.; Milchenko, M.; Xu, W.; Marcus, D.; Colen, R. R.; et al. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1): 1–12.

Sheller, M. J.; Reina, G. A.; Edwards, B.; Martin, J.; and Bakas, S. 2018. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, 92–104. Springer.

Shilo, S.; Rossman, H.; and Segal, E. 2020. Axes of a revolution: challenges and promises of big data in healthcare. *Nature medicine*, 26(1): 29–38.

Vaswani, S.; Bach, F.; and Schmidt, M. 2019. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1195–1204. PMLR.

Verma, R.; Kumar, N.; Patil, A.; Kurian, N. C.; Rane, S.; Graham, S.; Vu, Q. D.; Zwager, M.; Raza, S. E. A.; Rajpoot, N.; et al. 2021. MoNuSAC2020: A Multi-organ Nuclei Segmentation and Classification Challenge. *IEEE Transactions on Medical Imaging*.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. *Advances in Neural Information Processing Systems*, 33.

Xu, J.; Glicksberg, B. S.; Su, C.; Walker, P.; Bian, J.; and Wang, F. 2021. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1): 1–19.

Yeganeh, Y.; Farshad, A.; Navab, N.; and Albarqouni, S. 2020. Inverse distance aggregation for federated learning with non-iid data. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, 150–159. Springer.

Yin, D.; Pananjady, A.; Lam, M.; Papailiopoulos, D.; Ramchandran, K.; and Bartlett, P. 2018. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, 1998–2007. PMLR.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.