

# Fair Federated Medical Image Segmentation via Client Contribution Estimation

Meirui Jiang<sup>1</sup>, Holger R. Roth<sup>2</sup>, Wenqi Li<sup>2</sup>, Dong Yang<sup>2</sup>, Can Zhao<sup>2</sup>,  
 Vishwesh Nath<sup>2</sup>, Daguang Xu<sup>2</sup>, Qi Dou<sup>1,\*</sup>, Ziyue Xu<sup>2,\*</sup>

<sup>1</sup> The Chinese University of Hong Kong

<sup>2</sup> NVIDIA

## Abstract

*How to ensure fairness is an important topic in federated learning (FL). Recent studies have investigated how to reward clients based on their contribution (collaboration fairness), and how to achieve uniformity of performance across clients (performance fairness). Despite achieving progress on either one, we argue that it is critical to consider them together, in order to engage and motivate more diverse clients joining FL to derive a high-quality global model. In this work, we propose a novel method to optimize both types of fairness simultaneously. Specifically, we propose to estimate client contribution in gradient and data space. In gradient space, we monitor the gradient direction differences of each client with respect to others. And in data space, we measure the prediction error on client data using an auxiliary model. Based on this contribution estimation, we propose a FL method, federated training via contribution estimation (FedCE), i.e., using estimation as global model aggregation weights. We have theoretically analyzed our method and empirically evaluated it on two real-world medical datasets. The effectiveness of our approach has been validated with significant performance improvements, better collaboration fairness, better performance fairness, and comprehensive analytical studies. Code is available at <https://nvidia.github.io/NVFlare/research/fed-ce>*

## 1. Introduction

Recent development of federated learning (FL) facilitates collaboration for medical applications, given that multiple medical institutions can jointly train a consensus model without sharing raw data [1–6]. FL provides an opportunity to leverage larger and more diverse datasets to derive a robust and generalizable model [7, 8]. However, it is usually difficult to pool different institutions together

to train a FL model in practice. The challenges mainly lie in two aspects. First, it takes effort to set up and participate in federated training, medical institutions may not be sufficiently motivated to contribute to a FL study without a fair credit assignment and a fair reward allocation, i.e., *collaboration fairness* [9]. Second, medical data are heterogeneous in amounts and data-collection process [10–13], which may lead to inferior performance for clients with either less data or a data distribution deviating from others, harming *performance fairness* [14, 15]. It is critical to involve diverse datasets and improve individual prediction accuracy for building robust medical applications with low error tolerance [16]. Therefore, we argue that these two types of fairness need to be considered together.

Despite recent investigations on fairness-related topics, existing literature mostly addresses collaboration fairness and performance fairness separately. For example, methods for *collaboration fairness* aim to estimate client reward, by using the computation and communication cost of each client [17], evaluating local validation performance [18], and using cosine similarity between local and global updates [19]. Meanwhile, methods for *performance fairness* aim to mitigate performance disparities, by using minimax optimization to improve worst-performing clients [15, 20], re-weighting clients to adjust fairness/accuracy trade-off [14], or learning personalized models [21]. To adequately address concerns on these two fairness, we postulate that it is desirable to consider both simultaneously, because reward estimation and model performance could essentially be coupled during training. Solutions on how to tackle *collaboration fairness* and *performance fairness* together are still under-investigated, especially in medical domain.

To tackle this problem, our insight is to estimate the contribution of each client, and further use the contribution to promote training performance. The idea is inspired by Shapley Value (SV) [22], a classic approach to quantify the contribution of participants in cooperative game theory. SV proposes to permute all possible subsets of participants to calculate the contribution of a certain client. Some existing works have adopted SV for estimating client re-

\*Corresponding authors: Qi Dou (qidou@cuhk.edu.hk) and Ziyue Xu (ziyuex@nvidia.com)

ward [19, 23–25]. However, these methods mostly approximate SV by comparing local model updates or local model validations, which can be highly correlated with local sample numbers. A client with more samples can dominate the training, resulting in inaccurate estimation results. Therefore, finding a more accurate and robust estimation is imperative to break through this bottleneck.

In this work, we propose a novel client contribution estimation method to approximate SV by comparing a certain client with respect to all other clients. We further present a new FL algorithm, federated training via contribution estimation (*FedCE*), which uses client contributions as new weighting factors for global model aggregation. Specifically, since the fundamental setting of SV is to validate if a new client contributes to all possible combinations of existing clients, to effectively and efficiently approximate it, we propose to directly measure how a certain client contributes to all remaining clients, rather than computing all possible permutations. Our contribution measurement considers both gradient and data space to quantify the contribution of each client. In gradient space, we calculate the gradient direction differences between one client and all the other clients; and in data space, we measure the prediction error on client data by using an auxiliary model, which is calculated by excluding a client’s own parameters. By combining these two measurements, we are able to quantify each client’s contribution for collaboration fairness, and further use this estimation to promote training for performance fairness. Our main contributions are summarized as follows:

- We propose a novel method for client contribution estimation to facilitate *collaboration fairness*. We empirically and theoretically analyze the robustness of this estimation method under various FL data distributions.
- We propose a novel federated learning method, *FedCE*, based on the proposed client contribution estimation to help promote *performance fairness*, and we theoretically analyze the model’s convergence.
- We conduct extensive experiments on two medical image segmentation tasks. The proposed FedCE method significantly outperforms several latest state-of-the-art FL methods, and comprehensive analytical studies demonstrate the effectiveness of our method.

## 2. Related Works

### 2.1. Fairness in Federated Learning

Fairness has received much attention in machine learning area, it is a broad topic that studies unintended behaviors of machine learning models [26, 27]. Under the setting of FL, “individual/group fairness”, “collaboration fairness”, and “performance fairness” are three most commonly studied

types of fairness. The first one aims to mitigate model bias on specific protected attribute(s) [28–32], the second one expects each client to receive a reward that fairly reflects its contribution [9, 18], and the third one requires uniformity of the performance distribution across clients [33, 34]. In this paper, we mainly focus on the latter two - “collaboration fairness” and “performance fairness”. For collaboration fairness, Kang et al. [17] proposed using local computation and communication cost to estimate contribution; CFFL [18] investigated the fairness by evaluating the validation performance on each client; and Shi et al. [35] proposed to filter out low-quality local gradients based on loss measurement. For performance fairness, Mohri et al. [36] first proposed to optimize the performance of the single worst device by proposing a minimax optimization scheme. Later, q-FedAvg [14] was proposed with a more flexible optimization objective, which can be tuned based on the desired amount of fairness. Recently, Ditto [21] has been proposed to provide fairness by learning personalized models. However, current studies treat these two fairness as separate problems without utilizing their underlying connection. Also, most works are validated on common benchmark datasets (e.g., MNIST and CIFAR) with arbitrary client splits. It still remains a question of how to jointly tackle collaboration and performance fairness for real-world applications in medical imaging, where client data are multi-source, highly heterogeneous, and complicated.

### 2.2. Shapley Value based Client Valuation

Shapley value (SV) is a concept measuring importance of players in cooperative game theory [22, 37]. Based on this, Ghorbani et al. [38] proposed data SV to quantify the contribution of each data point in machine learning. Later on, Tang et al. [39] applied data SV on chest x-ray data. However, directly calculating SV is computationally expensive, and almost infeasible in FL with a decent number of participants. Under FL scenario, multiple studies have been performed aiming to efficiently approximate SV [23, 40]. For example, Kumar et al. [41] proposed to train linear models as proxies for client data, and used the model ensemble to approximate SV; Wang et al. [42] applied SV by considering clients in an ordered sequence rather than calculating all subsets. Song et al. [24] proposed to approximate SV by using validation accuracy of intermediate models during federated training; CGSV [19] approximated SV by using cosine similarity between local and global updates. Liu et al. [25] reconstructed FL models from gradients to approximate SV instead of repeat training with different permutations. However, these methods either require auxiliary validation data or solely rely on intermediate results. Our work aims to design a more comprehensive and practical measurement, which considers both intermediate status and actual performance without requiring extra validation data.

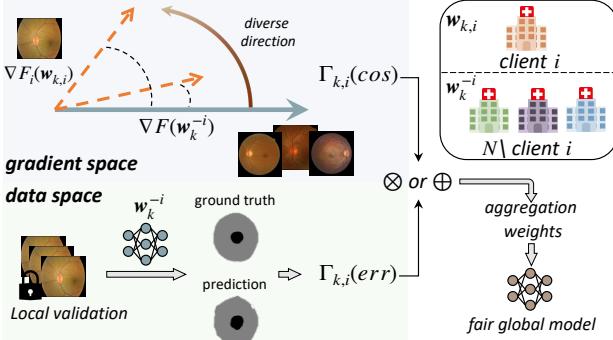


Figure 1. The proposed FedCE framework with client contribution estimation mechanism.

### 3. Methods

#### 3.1. Preliminary

Let  $\mathcal{D}$  denote a distribution supported on a space  $\mathcal{Z}$ , where  $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ ,  $\mathcal{X} \in \mathbb{R}^d$  and  $\mathcal{Y}$  are input and output respectively. For  $N \in \mathbb{N}$  clients, we have  $\mathcal{D}^N = \{\mathcal{D}_i\}_{i=1}^N$  as the set of local client distributions, and a coalition  $S \sim \mathcal{D}^M$  is a subset of clients, such that  $|S| = M$ , where  $M$  denotes number of clients in the coalition. Let  $U : \mathcal{Z} \rightarrow [0, 1]$  denote the utility function, where for any  $S \subseteq \mathcal{Z}$ ,  $U(S)$  represents the value of this subset. For example,  $U$  is typically chosen as the accuracy of an empirical risk minimizer when  $S$  are training clients. We define SV as below.

**Definition 3.1** (*Shapley Value* [43]) Given a utility function  $U$ , a distribution  $\mathcal{D}$  supported on  $\mathcal{Z}$ , and  $N \in \mathbb{N}$ , for all client  $i \in [N]$ , the Shapley Value (SV)  $\nu$  is defined as:

$$\nu(i; U, \mathcal{D}, N) = \mathbb{E}_{M \sim [N], S \sim \mathcal{D}^{M-1}} [U(S \cup \{i\}) - U(S)].$$

From this definition, the SV of a client is its expected marginal contribution in  $U$  to a set of client coalitions  $S$ . To calculate SV, we need to consider all possible client coalitions, i.e., all subsets of  $N$  clients. The cost will be exponentially increased with respect to the number of clients  $N$ , that is,  $\mathcal{O}(2^N)$ . Such computation is extremely expensive, even with a small number of clients. Therefore, it is critical to find an efficient solution for client valuation.

#### 3.2. Client Contribution Estimation

By analyzing the SV definition, we notice that the key is to measure the value with and without a certain client with respect to all possible combinations of other clients. In other words, validate if a new client contributes (adds value) to existing clients. Therefore, we propose an efficient approximation, by directly measuring the contribution of client  $i$  to others ( $N \setminus \{i\}$ ). We define our new value  $\hat{\nu}$  as:

$$\begin{aligned} \hat{\nu}(i; \Gamma, \mathcal{D}, N) &= \mathbb{E}_{S \sim \mathcal{D}^N} [\Gamma_i(S \setminus \{i\}, \{i\})] \\ &= \Gamma_i(N \setminus \{i\}, \{i\}), \end{aligned} \quad (1)$$

where  $\Gamma_i$  is our proposed function to measure the contribution of client  $i$ . Different from common implementations of the utility function  $U$  using accuracy, we propose a more comprehensive way to measure the contribution by considering both gradient and data space, as shown in Fig. 1.

We first introduce the measurement in gradient space by using cosine similarity (cos). For client  $i$  at  $k$ -th round:

$$\Gamma_{k,i}(\cos) \triangleq 1 - \cos (\nabla F_i(\mathbf{w}_{k,i}), \nabla F(\mathbf{w}_k^{-i})) , \quad (2)$$

where  $\nabla F_i(\mathbf{w}_{k,i})$  denotes the local client gradient, which is calculated by differences between global model parameter  $\mathbf{w}_k$  and local model parameter  $\mathbf{w}_{k,i}$ .  $\nabla F(\mathbf{w}_k^{-i}) = (\nabla F(\mathbf{w}_k) - p_i \nabla F_i(\mathbf{w}_{k,i})) / (1 - p_i)$  is aggregated gradients excluding client  $i$  and  $p_i \geq 0$  denotes client importance (e.g., proportional to client sample number). Global gradients is denoted by  $\nabla F(\mathbf{w}_k) = \sum_{i=1}^N p_i \nabla F_i(\mathbf{w}_{k,i})$  where  $\sum_{i=1}^N p_i = 1$ . Then we further normalize the term, i.e.,  $\Gamma_{k,i}(\cos) = \Gamma_{k,i}(\cos) / \sum_{i=1}^N \Gamma_{k,i}(\cos)$  (for ease of notation, we reuse  $\Gamma_{k,i}(\cos)$ ), to ensure the summation over clients adds up to 1.  $\Gamma_{k,i}(\cos)$  quantifies the contribution by measuring the optimization direction of client  $i$  compared with others. In particular, if the cosine similarity between client  $i$  and others is close to 1,  $\Gamma_{k,i}(\cos)$  becomes 0, indicating this client does not represent a new direction information. Hence, removing client  $i$  will have little impact on the global model's update direction. We argue that it is important to capture the large data variety and heterogeneity in FL for training a robust and generalizable global model. Therefore, we assign more weight to clients presenting different gradient directions.

However, as an indication in gradient space, a gradient direction that is different from others may not be sufficient to fully measure the contribution from a certain client to the overall FL model performance. Consequently, we further propose a measurement in data space by calculating the model error on the clients' data. Similar to the client exclusion setting in gradient space, we calculate the aggregated model parameters by excluding client  $i$  at  $k$ -th round, i.e.,  $\mathbf{w}_k^{-i} = (\mathbf{w}_k - p_i \mathbf{w}_{k,i}) / (1 - p_i)$ . Then we validate this new model on client  $i$ 's data samples:

$$\Gamma_{k,i}(\text{err}) \triangleq \mathcal{E}(\widehat{\mathcal{D}}_i; \mathbf{w}_k^{-i}), \quad (3)$$

where  $\mathcal{E}(\widehat{\mathcal{D}}_i; \mathbf{w}_k^{-i})$  denotes the error on the empirical distribution  $\widehat{\mathcal{D}}_i$ , here we use the validation samples in client  $i$ , i.e., the error is the full performance score “1” minus the obtained validation performance. We also normalize this term to ensure the summation over clients is 1. The intuition of measuring error is that, if a client presents a new data distribution, using model parameters from other clients only may not result in a good performance. Otherwise, if  $\mathbf{w}_k^{-i}$  already achieves low error on  $\widehat{\mathcal{D}}_i$ , incorporating updates from client  $i$  may not improve the overall performance significantly. By

further adding this assessment, we complement the previous findings in gradient space and better determine the contribution from a client to the overall model performance.

To combine these two factors for the contribution estimation, we choose multiplication and summation as two alternative mechanisms. We name them as  $\Gamma_{k,i}^m$  and  $\Gamma_{k,i}^s$ , respectively, and formulate them as:

$$\begin{cases} \Gamma_{k,i}^m = \Gamma_{k,i}(\cos) \times \Gamma_{k,i}(\text{err}) \\ \Gamma_{k,i}^s = \Gamma_{k,i}(\cos) + \Gamma_{k,i}(\text{err}). \end{cases} \quad (4)$$

These contribution estimation terms are calculated at each communication round, and by accumulating them over all  $K$  rounds, we can derive the final contribution estimations. We evaluate both combinations in our experiments.

### 3.3. Federated Training via Contribution Estimation – FedCE

From the proposed formulation of client contribution, it is natural to consider using these contribution estimation results to further improve federated training. In this regard, we propose a new federated algorithm, FedCE, by using client contributions as weighting factors for global model aggregation. Instead of using the standard federated averaging (FedAvg) weight  $p_i$ , which is typically proportional to the data amount of each client [44], we use our estimated client contribution as the new weight.

There are two benefits of using client contributions to promote fairness. First, client contributions are more comprehensive and fair than the standard weights based on sample numbers. The standard weights is a weak representation of client data distribution and could be vulnerable to data manipulations, such as increasing number by repeating. Second, contribution-based aggregation encourages the global model to cover a wide range of data distributions. Distributions sharing a common pattern are easy to fit. In contrast, clients with rare distribution are usually under-represented in training, which is a potential driver of model performance unfairness. Our contribution mechanism helps promote training on these clients, because they present different data information. As a result, this will facilitate the performance fairness of the global model.

Taking the multiplication-based combination as an example, at the  $k$ -th round, we calculate the aggregation weights as follows:

$$\rho_k^m = \frac{1}{Z_k} \left[ \sum_k \Gamma_{k,0}^m, \dots, \sum_k \Gamma_{k,N}^m \right], \quad (5)$$

where  $Z_k = k \sum_{i=1}^N \sum_k \Gamma_{k,i}^m$  is a normalization factor to ensure  $\sum_{i=1}^N \rho_{k,i}^m = 1$ . Then we obtain the global model by using the new weights to aggregate local client gradients:

$$\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k - \eta \sum_{i=1}^N \rho_{k,i}^m \cdot \nabla F_i(\mathbf{w}_{k,i}). \quad (6)$$

---

#### Algorithm 1 Our proposed method FedCE

**Input:** communication rounds  $K$ , number of clients  $N$ , local datasets  $\{\mathcal{D}_i\}_{i=1}^N$ , learning rate  $\eta$ , local steps  $\{\kappa_i\}_{i=1}^N$ .  
**Output:** final global model  $\mathbf{w}_K$ , contributions  $\{\rho_{K,i}\}_{i=1}^N$ .

```

1: Initialize server model  $\mathbf{w}_0$ 
2: for  $k = 1, \dots, K-1$  do
3:   Server:  $\mathbf{w}_{k,i}^0 \leftarrow \mathbf{w}_k$   $\triangleright$  distribute global model  $\mathbf{w}_k$ 
4:   for Client  $i = 1, 2, \dots, N$  in parallel do
5:      $\nabla F(\mathbf{w}_k) = \mathbf{w}_k - \mathbf{w}_{k-1}$ 
6:     for  $j = 1, 2, \dots, \kappa_i$  do  $\triangleright$  client training
7:        $\mathbf{w}_{(k,i)}^{j+1} \leftarrow \mathbf{w}_{(k,i)}^j - \eta \nabla F_i(\mathbf{w}_{(k,i)}^j)$ 
8:     end for
9:      $\nabla F_i(\mathbf{w}_{k,i}) = \mathbf{w}_{k,i}^{\kappa_i} - \mathbf{w}_{k,i}^0$ 
10:     $\nabla F(\mathbf{w}_k^{-i}) = \frac{\nabla F(\mathbf{w}_k) - \rho_{(k-1,i)} \nabla F_i(\mathbf{w}_{k,i})}{1 - \rho_{(k-1,i)}}$ 
11:     $\Gamma_{k,i}(\cos) = 1 - \cos(\nabla F_i(\mathbf{w}_{k,i}), \nabla F(\mathbf{w}_k^{-i})) \triangleright$  Eq.(2)
12:     $\mathbf{w}_k^{-i} = \frac{(\mathbf{w}_k - \rho_{(k-1,i)} \mathbf{w}_{k,i})}{(1 - \rho_{(k-1,i)})}$ .
13:     $\Gamma_{k,i}(\text{err}) = \mathcal{E}(\mathcal{D}_i; \mathbf{w}_k^{-i}) \triangleright$  Eq.(3)
14:    calculate  $\rho_{k,i}$   $\triangleright$  Eq.(5)
15:    return  $\mathbf{w}_{k,i}^{\kappa_i}, \rho_{k,i}$   $\triangleright$  send client model and
           contribution to server
16:  end for
17:  Server:  $\mathbf{w}_{k+1} \leftarrow \sum_{i=1}^N \rho_{k,i} \mathbf{w}_{k,i}^{\kappa_i}$ 
18: end for
19: return  $\mathbf{w}_K, \{\rho_{K,i}\}_{i=1}^N$ 

```

---

Similarly, for the summation-based combination (defined in Eq. (4)), we compute the aggregation weights of  $\rho_k^s$  by replacing  $\Gamma_{k,i}^m$  with  $\Gamma_{k,i}^s$ . We present the full algorithm in Algorithm 1. The final outputs are global model and client contribution estimations. Our aggregation weight is dynamical, and it considers all the historical information. Note that our method does not require any extra training compared with FedAvg [44]. Furthermore, the contribution estimation is performed locally, which helps reduce the communication burden and potential risk of information leakage.

### 3.4. Theoretical Analysis for FedCE

Since our contribution value is naturally parameterized by the underlying data distribution  $\mathcal{D}$ , it is helpful to investigate how the value will change if the data distribution changes. Here we formally quantify the value differences under distributional shift by presenting an upper bound.

**Theorem 3.2** Let  $\Gamma$  be  $\mathcal{B}$ -Lipschitz stable with respect to  $\mathcal{Z}$ . Suppose  $\mathcal{D}_s$  and  $\mathcal{D}_t$  are two distributions over  $\mathcal{Z}$ . Then, for all  $N \in \mathbb{N}$  and all  $i \in \mathcal{Z}$ ,

$$|\widehat{\nu}(i; \Gamma, \mathcal{D}_s, N) - \widehat{\nu}(i; \Gamma, \mathcal{D}_t, N)| \leq 2N\mathcal{B} \cdot W_1(\mathcal{D}_s, \mathcal{D}_t),$$

where  $W_1$  denotes the Wasserstein distance between two distributions. This theorem measures the values changes

under two different data distributions. If two distributions are similar, then similar values should be obtained. While for two different distributions, we can bound the difference in terms of the Wasserstein distance. For more details and proofs, please refer to Appendix A.

Then we analyze the convergence behavior of our method. To complete the analysis, we adopt assumptions on local function smoothness and gradient variance, which are classically used in optimization literature [45–49]. We present our results below.

**Theorem 3.3** *Assume the objective function is Lipschitz smooth and gradient variance is bounded. In the  $k$ -th round for  $k \in [0, K - 1]$ , let  $\beta_{(k,i)}$  and  $\delta_{(k,i)}$  be factors relate to variance bounding for client  $i$ ,  $L$  be factor of smoothness, and  $\eta$  be learning rate, when  $\eta L - 1 \geq 0$ , we have:*

$$F(\mathbf{w}_{k+1}) - F(\mathbf{w}_k) \leq \left( \frac{\eta}{4} (2\eta L + \sum_{i=1}^N p_i \rho_{(k,i)} \eta A_{(k,i)}) - \eta \right) \|\nabla F(\mathbf{w}_k)\|^2.$$

where  $A_{(k,i)} \triangleq \eta \sqrt{\kappa_{(k,i)}} (\kappa_{(k,i)} - 1) \beta_{(k,i)}^2 \delta_{(k,i)}$  is a variable that relates to  $\beta_{(k,i)}^2$ ,  $\delta_{(k,i)}$  and local iteration steps  $\kappa_{(k,i)}$ . The theorem analyzes the upper bound of the convergence with our method in the context of global model update. Furthermore, we present another corollary to determine the upper bound on our aggregation factor  $\rho_{(k,i)}$ .

**Corollary 3.4** *For  $k \in [0, K - 1]$ , considering the  $A_{(k,i)}$ -dominated convergence, the model converges when*

$$\rho_{(k,i)} \leq \frac{4A_{(k,i)}^{1/2}}{(A_{(k,i)} - 2L)} = \mathcal{O}\left(\frac{1}{\sqrt{A_{(k,i)}}}\right).$$

The proof can be found in Appendix B. According to this corollary, we could promote the convergence by minimizing the upper bound, that is, increasing  $A_{(k,i)}$ . Specifically, this term contains four terms. For  $\eta$  and  $\kappa$ , it is intuitive that increasing the learning rate or local iteration steps can increase model convergence speed at an early stage. However, it may let the model trap into a local optimum or suffer large client drifts when data are heterogeneous [48]. For the term  $\beta$ , it can be converted to a form related to  $\delta$ . So we discuss  $\delta$  here. The term  $\delta$  is a variable in one of our assumptions, which can be written as  $\|\sum_{s=0}^{\lambda-1} \nabla F_i(\mathbf{w}_{(k,i)}^s)\|^2 / (\sum_{s=0}^{\lambda-1} \|\nabla F(\mathbf{w}_k)\|^2)$ . This term quantifies the percentage of local gradients over the global(aggregated) gradients. That is, to increase  $\delta_{(k,i)}$ , we need to weigh more on local gradients from client  $i$ . Since the client with boundary data or different distribution is under-represented during training, which harms the overall convergence. We need to assign higher weights to promote training on this kind of client, thus improving convergence.

This well matches our contribution estimation method, i.e., allocating higher weight to clients presenting different information in gradient space or suffering high error on local data when their gradient is excluded.

## 4. Experiment

Our method is evaluated on two medical image segmentation tasks: retinal fundus image segmentation [50] and prostate MRI segmentation [51]. We compare our method with other methods on segmentation performance, performance fairness, and collaboration fairness. We also conduct in-depth analyses on our method, including convergence speed, robustness to free riders and distribution changes, and effectiveness of each component. For more results, please refer to Appendix C.

### 4.1. Experimental Settings

**Datasets.** We evaluate our approach on two medical image segmentation datasets: the prostate MRI dataset from 6 institutions [51–54], and the retinal fundus dataset from 6 different sources [50, 55–57]. Each institution/source is treated as a single client, and the data is randomly split into training, validation, and testing sets with a ratio of 0.5, 0.25, and 0.25 for each client. All images are resized to  $256 \times 256$ . Note that the data collected from different medical centers present realistic heterogeneous distributions, due to varying local devices and imaging protocols. As shown in Fig. 2, the retinal fundus dataset has lower data distribution similarity among clients, while the prostate dataset has a relatively higher data similarity.

**Evaluation metrics.** To comprehensively evaluate our approach, we adopt four different metrics. We use the Dice coefficient (Dice) to evaluate segmentation results. We use the Pearson correlation and Euclidean distance to measure the performance fairness, and further add cosine similarity to evaluate the accuracy of contribution estimation. Following the fairness definition from [14, 21], we also calculate the standard deviation of test performance among clients.

**Implementation details.** In our implementation, all methods use the same training settings. The loss function is dice loss [58], and the optimizer is Adam with  $\beta = (0.9, 0.99)$ . The learning rate is  $1e-3$  and the batch size is set to 8. We trained for 200 federated rounds to ensure that the model converged steadily, and the local update epoch is set to 1.

**Baseline methods.** We compare our approach with state-of-the-art (SOTA) methods targeting collaboration fairness and performance fairness, including: q-FedAvg [14], a method to learn fair performance distribution; CFFL [18], a method for collaboration fairness by evaluating local participant's validation accuracy; FedCI, a method we extend from a client valuation method CI [24], by using the valuation results as aggregation weights; CGSV [19], a method quantifying client reputation based on SV. Furthermore, we also

Table 1. Performance comparison using Dice score on image segmentation datasets of retinal fundus images and prostate MRI.

Task	Retinal Fundus Segmentation								Prostate MRI Segmentation							
	1	2	3	4	5	6	Avg.	Std.	1	2	3	4	5	6	Avg.	Std.
Client	86.69	85.51	86.21	89.91	79.77	90.98	86.51	3.95	91.23	84.59	87.57	87.37	86.70	89.25	87.79	2.26
Standalone	81.34	85.21	83.28	88.16	40.81	90.79	78.27	18.66	91.10	84.59	89.02	89.09	83.87	<b>89.27</b>	87.82	2.90
FedAvg	86.24	86.97	87.37	89.13	44.68	90.72	80.85	17.80	90.94	85.60	89.28	89.18	84.27	88.67	87.99	2.52
q-FedAvg	85.72	86.29	86.96	88.62	41.12	90.16	79.81	19.02	91.01	85.49	89.24	88.98	82.11	88.17	87.50	3.20
CFFL	87.02	86.93	87.35	88.53	40.99	90.22	80.17	19.24	91.21	85.40	89.49	88.37	83.96	88.49	87.82	2.68
FedCI	83.46	85.57	85.47	88.48	33.79	<b>91.01</b>	77.96	21.80	91.15	84.90	89.27	88.09	83.47	89.16	87.67	2.91
FedCE (Multi.)	86.73	<b>87.45</b>	87.51	89.26	<b>57.30</b>	90.25	<b>83.08</b>	<b>12.70</b>	<b>91.43</b>	<b>85.79</b>	89.21	89.13	<b>85.68</b>	88.62	<b>88.31</b>	<b>2.22</b>
FedCE (Sum.)	<b>87.22</b>	87.36	<b>87.93</b>	<b>89.66</b>	54.42	90.92	82.92	14.03	91.18	85.54	<b>89.59</b>	<b>89.22</b>	84.99	88.79	88.22	2.43

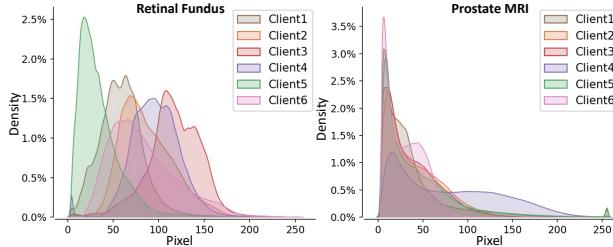


Figure 2. Pixel intensity distributions. Left denotes samples from retinal fundus dataset and right ones are from prostate MRI.

compared with the FedAvg [44] and Standalone (local training and evaluation on each client’s own data).

## 4.2. Experimental Results

**Segmentation performance.** We first present the comparison on segmentation performance. Table 1 lists all the quantitative results for two segmentation tasks, including performance on each client, the average performance, and the standard deviation across clients. The goal of performance fairness methods is to lower variance while maintaining the average performance. From the table, it can be observed that client 5 suffers a significant performance drop across FL methods due to less similar data distributions in retinal dataset. Since the compared methods may not specifically consider such large heterogeneity in training data, their performance on client 5 is lower than FedAvg and suffer a higher standard deviation. For other clients, these methods present higher or comparable results. Compared with them, our approaches achieve higher overall performance with an increase of 2.23% and 2.07% and make improvements on most clients (5 over 6). And the performance variance of our method is significantly lower than others (with a decrease of 5.10). For prostate MRI, most methods achieve higher average performance and lower variance than FedAvg. And our methods outperform all SOTA methods in terms of overall performance and variance.

**Performance fairness.** One aim of our study is to improve performance fairness, which we evaluate and compare our

Table 2. Fairness comparison with our method and others. We use Pearson correlation ( $\uparrow$ ) and Euclidean distance ( $\downarrow$ ) as metrics. Value in parentheses denotes the p-value.

Task	Retinal Fundus Segmentation		Prostate MRI Segmentation		
	Metric	Pearson Correlation	Euclidean Distance	Pearson Correlation	Euclidean Distance
FedAvg [44]		88.82 ( $1.8e^{-2}$ )	38.94	88.67 ( $1.9e^{-2}$ )	3.31
q-FedAvg [14]		87.02 ( $2.4e^{-2}$ )	38.96	91.69 ( $1.0e^{-2}$ )	2.61
CFFL [18]		84.53 ( $3.4e^{-1}$ )	38.96	92.47 ( $8.3e^{-3}$ )	2.46
FedCI [24]		85.69 ( $2.9e^{-2}$ )	48.42	93.57 ( $6.1e^{-3}$ )	2.25
CGSV [19]		87.47 ( $2.3e^{-2}$ )	49.47	87.71 ( $2.2e^{-2}$ )	3.45
FedCE (Multi.)		88.94 ( $1.8e^{-2}$ )	<b>24.57</b>	<b>98.25</b> ( $4.6e^{-4}$ )	<b>1.09</b>
FedCE (Sum.)		<b>89.11</b> ( $1.7e^{-2}$ )	24.69	97.15 ( $1.2e^{-3}$ )	1.41

method with others on fairness metrics. Besides comparing the standard deviation of performance in Table 1, we further consider using a scaled Pearson correlation and Euclidean distance, which are also adopted in [19, 40]. The results of our fairness comparison are presented in Table 2, where we calculate the Pearson correlation and Euclidean distance between the test Dice scores of standalone and other methods. Our methods consistently achieve a high degree of fairness compared to others, as indicated by the higher correlation value and lower distance to the Standalone results. The p-value of our results is smaller than 0.05 and also lower than others. Notably, for prostate segmentation, we observe that methods with better fairness than FedAvg also achieve higher accuracy and smaller variance, highlighting the importance of fairness metrics in performance evaluation.

**Collaboration fairness.** For collaboration fairness, our proposed method provides an indication of reward/profit distribution by measuring the contribution of clients. For methods in comparison, except FedAvg, the aggregation weights of others are also dynamic during training. So we take the aggregation weights as client contributions and perform the comparison. To obtain the "ground-truth" of a client’s contribution, we conduct leave-one-out experi-

Table 3. Client contribution estimation comparison by comparing the results of leave-one-out with that of our method and others. We use Pearson correlation ( $\uparrow$ ), Euclidean distance ( $\downarrow$ ), and cosine similarity ( $\uparrow$ ). Value in parentheses denotes the p-value.

Task	Retinal Fundus Segmentation			Prostate MRI Segmentation			
	Metric	Pearson Correlation	Euclidean Distance	Cosine Similarity	Pearson Correlation	Euclidean Distance	Cosine Similarity
FedAvg [44]	-39.76 ( $4.4e^{-1}$ )	0.55	0.26	3.01 ( $9.5e^{-1}$ )	0.62	0.52	
q-FedAvg [14]	63.28 ( $1.8e^{-1}$ )	0.31	0.57	63.35 ( $1.8e^{-1}$ )	0.59	0.57	
CFFL [18]	0.90 ( $9.9e^{-1}$ )	0.45	0.47	75.44 ( $8.3e^{-2}$ )	0.49	0.74	
FedCI [24]	-12.36 ( $8.2e^{-1}$ )	0.37	0.53	-0.31 ( $1.0e^0$ )	0.61	0.53	
CGSV [19]	-44.50 ( $3.8e^{-1}$ )	0.57	0.22	-1.85 ( $9.7e^{-1}$ )	0.63	0.50	
FedCE (Multi.)	94.93 ( $3.8e^{-3}$ )	<b>0.17</b>	<b>0.82</b>	93.12 ( $6.9e^{-3}$ )	<b>0.49</b>	<b>0.75</b>	
FedCE (Sum.)	<b>96.34</b> ( $2.0e^{-3}$ )	0.22	0.73	<b>93.53</b> ( $6.1e^{-3}$ )	0.53	0.69	

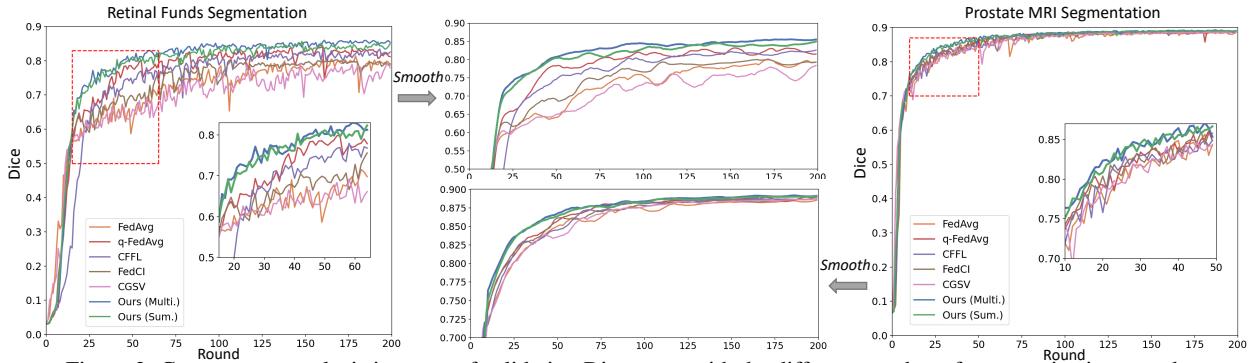


Figure 3. Convergence analysis in terms of validation Dice score with the different number of communication rounds.

ments, a popular and reliable way for data valuation [38]. This approach assesses how much performance we will lose if we remove a certain client. After obtaining leave-one-out results, we compare them with contributions estimated by our approach and other methods. Table 3 presents all the comparison results using three different metrics. It can be observed that, our methods achieve a higher correlation and cosine similarity, as well as lower distance compared with others. We notice that FedAvg presents a low correlation value, it is reasonable because the proportion of sample number may not correlate well with performance improvements. These three metrics comprehensively validate the accuracy of our client contribution estimation.

### 4.3. Analytical Studies

We further conduct in-depth analytical studies to investigate key properties of our method, including: i) the convergence speed; ii) the robustness against free riders; iii) the robustness against distribution changes; and iv) the contribution of each measurement metric.

**Convergence speed.** We first show the average validation Dice score across clients per communication round for different FL methods. As shown in Fig. 3, it can be observed that the curves of our methods converge to higher performance with faster speed than compared methods. This at-

tributes to our contribution-based aggregation, which involves diverse gradients to promote global model optimization on the overall data distribution. This observation also validates our theoretical analysis that using contribution estimation to aggregate models in FL helps promote convergence. In addition, we applied Savitzky–Golay filter [59] to smooth the curves to better present the overall trend.

**Robustness against free riders.** We then study a situation where a “free rider” joins the FL: if a client does not have enough data to participate in FL, it may cheat by repeating one image several times, and try to obtain the global model for its own use. However, in such case, the free rider has almost no contribution to the global training, and it should not enjoy such “free lunch” [60]. We hereby consider identifying the free rider by calculating a new value, which is the multiplication of local-global gradients cosine similarity and local-global model error difference. Note that these gradients and models are naturally generated during our method training. The results are shown in Fig. 4, we present snapshots on six communication rounds. Each row denotes an independent federated training, and the y-axis shows which client is the free rider. It can be observed that free riders are detected at very early stages, i.e., within 10 rounds, and as training goes on, the results become more significant. Note that when client 6 is the free rider, even

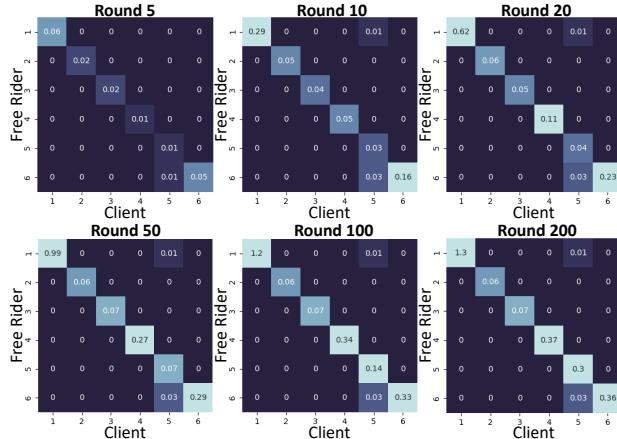


Figure 4. Free rider study. Each row denotes an independent federated training procedure, and the y-axis indicates which client is the free rider. A free rider is detected with a high value.

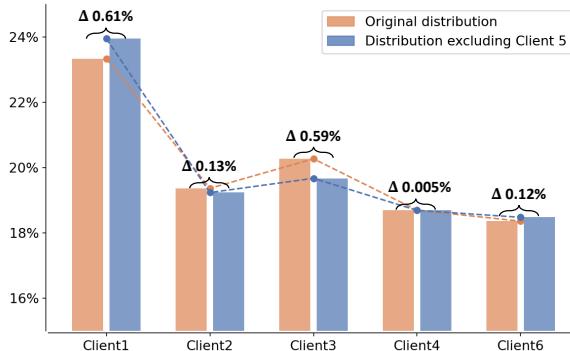


Figure 5. Study to validate the distribution shift robustness of our contribution estimation metric. The y-axis denotes the relative weight percentage, and  $\Delta$  denotes the differences.

though client 5 has a relatively high value of 0.03 at round 10, client 6 has a significantly higher value (over 5 times) than client 5. Our results present the potential of identifying free riders at an early stage to save time and development costs in real-world practice.

**Robustness against distribution changes.** For client contribution estimation, we may expect that the relative value among clients should be robust to distribution changes. Therefore, we form two distributions to investigate the estimation robustness of our method. We notice the overall clients’ distributions differ a lot by including/excluding client 5 on retinal fundus dataset. We use “original distribution” to denote all 6 clients and “distribution excluding Client 5” to denote the collection of other 5 clients. We present the estimation value in Fig. 5. The five estimation results from “original distribution” are re-normalized. Our metric presents similar estimation values for these 5 clients, even though the overall distributions are different. The value changes are smaller than 1%. And two different estimations have a similar trend, as shown by the curves.

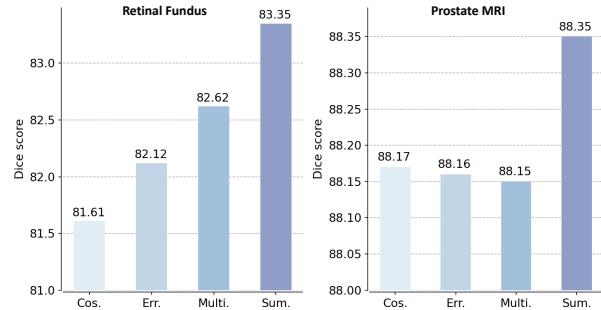


Figure 6. Ablation study on effects of two separate contribution quantification metrics and their combination on two datasets.

This study also empirically validates our theorem 3.2 for the upper bound of value changes under distribution shift.

**Contribution of each component.** We further conduct the ablation study for our two components (i.e.,  $\Gamma(\cos)$  and  $\Gamma(\text{err})$ ). As shown in Fig. 6, solely using either one will lead to a decrease in the performance on both segmentation tasks. This is reasonable because either individual metric may not be able to fully quantify the contribution. As for the combination, the performance improvements reflect how the two components play complementary roles in improving the global model. Please note, on the prostate dataset, the performance differences by single measurement and multiplication are marginal. This may be because both follow similar trends for this application, where the differences between clients are significantly smaller than in the retinal dataset. Even in this case, the summation combination still shows improved performance.

## 5. Conclusion

We have studied a novel and practical problem of jointly tackling *collaboration fairness* and *performance fairness*. We have proposed a novel method to estimate client contributions from both gradient and data space, followed by a fair reward allocation based on those estimates. We further design a novel fair FL algorithm by using the estimated contributions to re-weight the global aggregation. Our solution provides inspiration to motivate more clients to join a FL project, leveraging larger and diverse data, benefiting the acceptance of FL for medical imaging and healthcare applications. We conducted comprehensive experiments on two medical datasets and provided theoretical analysis for estimation robustness and model convergence. Our proposed estimation mechanism is extendable to other FL problems, such as Non-IID data and adversarial robustness. For future work, we plan to extend our method to detect adversarial clients and consider fairness on clients with noisy data.

## Acknowledgement

This work was supported by NVIDIA and National Natural Science Foundation of China (Project No. 62201485).

## References

- [1] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020. 1
- [2] Chuhan Wu, Fangzhao Wu, et al. Communication-efficient federated learning via knowledge distillation. *Nature communications*, 13(1):1–8, 2022. 1
- [3] Micah J Sheller, Brandon Edwards, G Anthony Reina, Jason Martin, Sarthak Pati, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020. 1
- [4] Ittai Dayan, Holger R Roth, et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature medicine*, 27(10):1735–1743, 2021. 1
- [5] Qi Dou, Tiffany Y So, Meirui Jiang, Quande Liu, et al. Federated deep learning for detecting covid-19 lung abnormalities in ct: a privacy-preserving multinational validation study. *NPJ digital medicine*, 4(1):1–11, 2021. 1
- [6] Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G Anthony Reina, Patrick Foley, et al. Federated learning enables big data for rare cancer boundary detection. *arXiv preprint arXiv:2204.10836*, 2022. 1
- [7] Liangqiong Qu, Yuyin Zhou, Paul Pu Liang, Yingda Xia, Feifei Wang, et al. Rethinking architecture design for tackling data heterogeneity in federated learning. In *CVPR*, pages 10061–10071, 2022. 1
- [8] Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M. Patel. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In *CVPR*, pages 2423–2432, June 2021. 1
- [9] Zirui Zhou, Lingyang Chu, Changxin Liu, Lanjun Wang, Jian Pei, and Yong Zhang. Towards fair federated learning. In *KDD*, pages 4100–4101, 2021. 1, 2
- [10] Marc Aubreville, Christof A Bertram, Taryn A Donovan, Christian Marzahl, Andreas Maier, and Robert Klopfleisch. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Scientific data*, 7(1):1–10, 2020. 1
- [11] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *ICLR*, 2021. 1
- [12] Jie Xu, Benjamin S Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1):1–19, 2021. 1
- [13] Lin Zhang, Li Shen, Liang Ding, Dacheng Tao, and Ling-Yu Duan. Fine-tuning global model via data-free knowledge distillation for non-iid federated learning. In *CVPR*, pages 10174–10183, June 2022. 1
- [14] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *ICLR*, 2020. 1, 2, 5, 6, 7
- [15] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *ICML*, pages 4615–4625. PMLR, 2019. 1
- [16] Xiaoxuan Liu, Ben Glocker, Melissa M McCradden, Marzyeh Ghassemi, Alastair K Denniston, and Lauren Oakden-Rayner. The medical algorithmic audit. *The Lancet Digital Health*, 2022. 1
- [17] Jiawen Kang, Zehui Xiong, Dusit Niyato, Shengli Xie, and Junshan Zhang. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journal*, 6(6):10700–10714, 2019. 1, 2
- [18] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. In *Federated Learning*, pages 189–204. Springer, 2020. 1, 2, 5, 6, 7
- [19] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. *NeurIPS*, 34:16104–16117, 2021. 1, 2, 5, 6, 7
- [20] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multi-objective optimization. *IEEE TNSE*, 2022. 1
- [21] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *ICML*, pages 6357–6368, 2021. 1, 2, 5
- [22] Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997. 1, 2
- [23] Shuyue Wei, Yongxin Tong, et al. Efficient and fair data valuation for horizontal federated learning. In *Federated Learning*, pages 139–152. Springer, 2020. 2
- [24] Tianshu Song, Yongxin Tong, and Shuyue Wei. Profit allocation for federated learning. In *IEEE Big Data*, pages 2577–2586. IEEE, 2019. 2, 5, 6, 7
- [25] Zelei Liu, Yuanyuan Chen, Han Yu, et al. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM TIST*, 13(4):1–21, 2022. 2
- [26] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017. 2
- [27] Shira Mitchell, Eric Potash, et al. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018. 2
- [28] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021. 2
- [29] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *SDM*, pages 181–189. SIAM, 2021. 2
- [30] Borja Rodríguez Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS Workshop Privacy in Machine Learning*, 2021. 2
- [31] Sen Cui, Weishen Pan, Jian Liang, et al. Addressing algorithmic disparity and performance inconsistency in federated learning. *NeurIPS*, 34:26091–26102, 2021. 2

- [32] Lingyang Chu, Lanjun Wang, Yanjie Dong, et al. Fedfair: Training fair models in cross-silo federated learning. *arXiv preprint arXiv:2109.05662*, 2021. 2
- [33] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *Big Data*, pages 1051–1060. IEEE, 2020. 2
- [34] Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *NeurIPS*, 33:15111–15122, 2020. 2
- [35] Zhuan Shi, Lan Zhang, Zhenyu Yao, et al. Fedfaim: A model performance-based fair incentive mechanism for federated learning. *IEEE Transactions on Big Data*, 2022. 2
- [36] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *ICML*, pages 4615–4625, 2019. 2
- [37] Roger B Myerson. *Game theory: analysis of conflict*. Harvard university press, 1997. 2
- [38] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *ICML*, pages 2242–2251. PMLR, 2019. 2, 7, 14
- [39] Siyi Tang, Amirata Ghorbani, et al. Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset. *Scientific reports*, 11, 2021. 2
- [40] Ruoxi Jia, David Dao, Boxin Wang, et al. Towards efficient data valuation based on the shapley value. In *AISTATS*, pages 1167–1176. PMLR, 2019. 2, 6
- [41] Sourav Kumar, A Lakshminarayanan, et al. Towards more efficient data valuation in healthcare federated learning using ensembling. In *DeCaF, FAIR workshops*, pages 119–129. Springer, 2022. 2
- [42] Tianhao Wang, Johannes Rausch, et al. A principled approach to data valuation for federated learning. In *Federated Learning*, pages 153–167. Springer, 2020. 2
- [43] Amirata Ghorbani, Michael Kim, and James Zou. A distributional framework for data valuation. In *ICML*, pages 3535–3544. PMLR, 2020. 3, 11
- [44] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017. 4, 6, 7, 14
- [45] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *ICLR*, 2019. 5, 12
- [46] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020. 5, 12
- [47] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, et al. Adaptive federated optimization. In *ICLR*, 2021. 5, 12
- [48] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, et al. SCAFFOLD: Stochastic controlled averaging for federated learning. In *ICML*, 2020. 5, 12, 13
- [49] Qianqian Tong, Guannan Liang, and Jinbo Bi. Effective federated adaptive gradient methods with non-iid decentralized data. *arXiv preprint arXiv:2009.06557*, 2020. 5, 12
- [50] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *MedIA*, 59:101570, 2020. 5
- [51] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *MedIA*, 18(2):359–373, 2014. 5
- [52] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE TMI*, 2020. 5
- [53] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. 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, 2015. 5
- [54] Bloch Nicholas, Madabhushi Anant, Huisman Henkjan, Freymann John, Kirby Justin, et al. Nci-proc. ieee-isbi conf. 2013 challenge: Automated segmentation of prostate structures. The Cancer Imaging Archive, 2015. 5
- [55] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, pages 1–6. IEEE, 2011. 5
- [56] Jayanthi Sivaswamy, S Krishnadas, Arunava Chakravarty, G Joshi, A Syed Tabish, et al. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1):1004, 2015. 5
- [57] Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Es-lam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhamad Alkatee, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Retinal fundus images for glaucoma analysis: the riga dataset. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, pages 55–62. SPIE, 2018. 5
- [58] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5
- [59] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964. 7
- [60] Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*, 2020. 7
- [61] Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002. 11