

# Federated Semi-supervised Medical Image Classification via Inter-client Relation Matching

Quande Liu<sup>1</sup>, Hongzheng Yang<sup>2</sup>, Qi Dou<sup>1</sup>, and Pheng-Ann Heng<sup>1,3</sup>

<sup>1</sup> Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China  
{qdliu, qdou}@cse.cuhk.edu.hk

<sup>2</sup> Department of Computer Science and Engineering, Beihang University, China

<sup>3</sup> Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

**Abstract.** Federated learning (FL) has emerged with increasing popularity to collaborate distributed medical institutions for training deep networks. However, despite existing FL algorithms only allow the supervised training setting, most hospitals in realistic usually cannot afford the intricate data labeling due to absence of budget or expertise. This paper studies a practical yet challenging FL problem, named *Federated Semi-supervised Learning* (FSSL), which aims to learn a federated model by jointly utilizing the data from both labeled and unlabeled clients (i.e., hospitals). We present a novel approach for this problem, which improves over traditional consistency regularization mechanism with a new inter-client relation matching scheme. The proposed learning scheme explicitly connects the learning across labeled and unlabeled clients by aligning their extracted disease relationships, thereby mitigating the deficiency of task knowledge at unlabeled clients and promoting discriminative information from unlabeled samples. We validate our method on two large-scale medical image classification datasets. The effectiveness of our method has been demonstrated with the clear improvements over state-of-the-arts as well as the thorough ablation analysis on both tasks<sup>4</sup>.

**Keywords:** Federated learning · Semi-supervised learning · Medical image classification

## 1 Introduction

Data collaboration across medical institutions is increasingly desired to mitigate the scarcity and distribution bias of medical images, thereby improving the model performance on important tasks such as disease diagnosis [6,27]. Recently, federated learning (FL) has emerged as a privacy-preserving solution for this, which allows to learn from distributed data sources by aggregating the locally learned model parameters without exchanging the sensitive health data [8,12,14,16,23]. However, despite progress achieved, existing FL algorithms typically only allow the supervised training setting [3,13,15,24,26], which has limited the local clients

<sup>4</sup> Code will be made available at <https://github.com/liuquande/FedIRM>

(i.e., hospitals) without data annotations to join the FL process. Yet, in realistic scenarios, most hospitals usually cannot afford the intricate data labeling due to lack of budget or expertise [22]. How to utilize these widely-existing unlabeled datasets to further improve the FL models is still an open question to be solved.

To this end, we study a practical FL problem which involves only several labeled clients while most of the participating clients are unlabeled, namely *federated semi-supervised learning (FSSL)*. This is also noted by Yang et al. [30] very recently in COVID-19 lesion segmentation, which halts in an extremely simple case containing only one labeled and one unlabeled client. In contrast to their work, we for the first time broaden this problem to a more practical yet complex scenario in which multiple distributed labeled and unlabeled clients are involved. To address this problem, a naive solution is to simply integrate the off-the-rack semi-supervised learning (SSL) methods onto the federated learning paradigm. However, previous SSL methods are typically designed for centralized training setting [2, 10, 17, 29], which rely heavily on the assumption that the labeled data is accessible to provide necessary assistance for the learning from unlabeled data [1, 4]. In consistency-based methods [5, 31], for instance, the regularization of perturbation-invariant model predictions needs the synchronous labeled data supervision, in order to obtain the necessary task knowledge to produce reliable model predictions for unlabeled data where the consistency regularization is imposed on. Unfortunately, such close assistance from labeled data is lost in FSSL scenario, where the local dataset could be completely unlabeled. This will make the local model aloof from original task as the consistency-based training goes on, hence fail to fully exploit the knowledge at unlabeled clients.

Based on the above issues, compared with traditional SSL problem, the main challenge in FSSL lies in how to build the interaction between the learning at labeled and unlabeled clients, given the challenging constraint of data decentralization. In this work, our insight is to communicate their knowledge inherent in disease relationships to achieve this goal. The idea is motivated by an observation that the relationships exist naturally among different categories of disease and reflect the structural task knowledge in the context of medical image classification, as evidenced by disease similarity measure [18, 20]. More importantly, such disease relationships are independent of the observed hospitals, i.e., similar disease at one hospital should also be high-related at others. We may consider extracting such client-invariant disease relation information from labeled clients to supervise the learning at unlabeled clients, thereby mitigating the loss of task knowledge at unlabeled clients and effectively exploiting the unlabeled samples.

In this paper, we present to our knowledge the first FSSL framework for medical image classification, by exploring the client-independent disease relation information to facilitate the learning at unlabeled clients. Our method roots in the state-of-the-art consistency regularization mechanism, which enforces the prediction consistency under different input perturbations to exploit the unlabeled data. To address the loss of task knowledge at unlabeled clients which would lead to degenerated learning, we introduce a novel *Inter-client Relation Matching* scheme, by explicitly regularizing the unlabeled clients to capture similar disease relationships as labeled clients for preserving the discriminative task knowledge.

To this end, we propose to derive the disease relation matrix at labeled clients from pre-softmax features, and devise an uncertainty-based scheme to estimate reliable relation matrix at unlabeled clients by filtering out inaccurate pseudo labels. We validate our method on two large-scale medical image classification datasets, including intracranial hemorrhage diagnosis with 25000 CT slices and skin lesion diagnosis with 10015 dermoscopy images. Our method achieves large improvements by utilizing the unlabeled clients, and clearly outperforms the combination of federated learning with state-of-the-art SSL methods.

## 2 Method

In FSSL scenario, we denote  $\mathcal{D}_L = \{\mathcal{D}^1, \mathcal{D}^2, \dots, \mathcal{D}^m\}$  be the collection of  $m$  labeled clients, where each labeled client  $l$  contains  $N^l$  data and one-hot label pairs  $\mathcal{D}^l = \{(x_i^l, y_i^l)\}_{i=1}^{N^l}$ ; and let  $\mathcal{D}_U = \{\mathcal{D}^{m+1}, \mathcal{D}^{m+2}, \dots, \mathcal{D}^{m+n}\}$  be the  $n$  unlabeled clients, with each unlabeled client  $u$  containing  $N^u$  data samples  $\mathcal{D}^u = \{(x_i^u)\}_{i=1}^{N^u}$ . The goal of FSSL is to learn a global federated model  $f_\theta$  jointly utilizing the data from both labeled and unlabeled clients. Fig. 1 gives an overview of our proposed FSSL solution, i.e. FedIRM, in comparison with the naive FSSL solution.

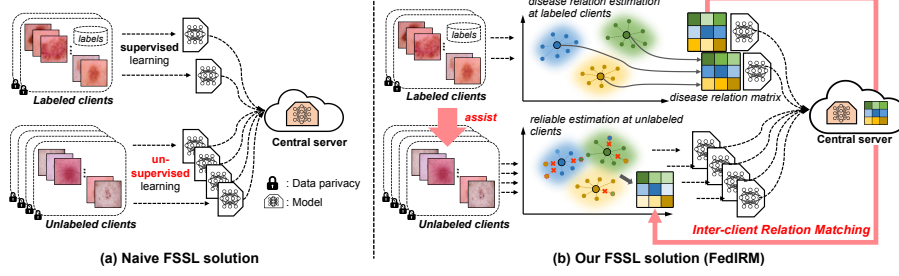
### 2.1 Backbone Federated Semi-supervised Learning Framework

Our method follows standard FL paradigm which involves the communication between a central server and local clients. Specifically, in each federated round, every client  $k$  will receive the same global model parameters  $\theta$  from the central server, and update the model with local learning objective  $\mathcal{L}^k$  for  $e$  epochs on its private data  $\mathcal{D}^k$ . The central server then collects the local model parameters  $\theta^k$  from all clients and aggregate them to update the global model. Such process repeats until the global model converges stably. In this work, we adopt the well-established federated averaging algorithm [19] (FedAvg) to update the global model, by aggregating the local model parameters with weights in proportional to the size of each local dataset, i.e.,  $\theta = \sum_{k=1}^K \frac{N^k}{N} \theta^k$ , where  $N = \sum_{k=1}^K N^k$ .

In our FSSL solution, the local learning objective at labeled clients adopts the cross entropy loss for capturing the discriminative task knowledge. At unlabeled clients, we preserve the state-of-the-art consistency regularization mechanism, which exploits the unlabeled data in an unsupervised manner by enforcing the consistency of model predictions under input perturbations. Formally, this learning objective at each unlabeled client  $u$  could be expressed as:

$$\mathcal{L}_c(\mathcal{D}^u, \theta^u) = \sum_{i=1}^{N^u} \mathbb{E}_{\xi, \xi'} \|f_{\theta^u}(x_i^u, \xi), f_{\theta^u}(x_i^u, \xi')\|_2^2 \quad (1)$$

where  $\xi$  and  $\xi'$  denote different input perturbations (e.g., adding Gaussian noise).



**Fig. 1.** (a) Naive FSSL solution simply performs unsupervised learning (e.g. consistency regularization) at unlabeled clients, hence the local model is prone to forget the original task knowledge as the training goes on. (b) **Our FedIRM explicitly utilizes the knowledge from labeled clients to assist the learning at unlabeled clients by aligning their extracted disease relationships, thereby mitigating the loss of task knowledge at unlabeled clients and promoting discriminative information from unlabeled data.**

## 2.2 Disease Relation Estimation at Labeled and Unlabeled Clients

Without the assistance from labeled data supervision, the local learning at unlabeled clients solely with consistency regularization is prone to forget the original task knowledge, therefore failing to fully exploit the information from unlabeled samples. To tackle this problem, we introduce a novel *inter-client relation matching* (IRM) scheme, which explicitly extracts the knowledge from labeled clients to assist the learning at unlabeled clients, by exploiting the rich information inherent in disease relationships. Specifically, the relationships exist naturally across different categories of disease and reflect structural task knowledge in medical image classification, *independent of the changes of observed hospitals*. In light of this, we aim to enforce the alignment of such disease relations across labeled and unlabeled clients, thereby promoting the learning of discriminative information at unlabeled clients for preserving such structural task knowledge.

**Disease Relation Estimation at Labeled Clients.** Inspired by knowledge distillation from deep networks, we estimate the disease relationships from the class ambiguity captured by deep models, i.e., per-class soft labels, and enforce them to be consistent between labeled and unlabeled clients. Formally, we first consider the relation estimation at labeled clients. For each labeled client  $\mathcal{D}^l$ , we summarize the model’s knowledge on each class  $c$  by computing per-category mean feature vectors  $\mathbf{v}_c^l \in \mathbb{R}^C$  (with  $C$  denoting total class number):

$$\mathbf{v}_c^l = \frac{1}{N_c^l} \sum_{i=1}^{N_c^l} \mathbb{1}_{[y_i^l=c]} \hat{f}_{\theta^l}(x_i^l) \quad (2)$$

where  $N_c^l$  is the number of samples with class  $c$  at labeled client  $\mathcal{D}^l$ ;  $\mathbb{1}_{[\cdot]}$  denotes the indicator function,  $\hat{f}$  denotes the model without last softmax layer. The obtained  $\mathbf{v}_c^l$  is then **scaled to a soft label distribution**, with a softened softmax function under temperature  $\tau > 1$  [7]:

$$\mathbf{s}_c^l = \text{softmax}(\mathbf{v}_c^l / \tau) \quad (3)$$

This distilled knowledge of soft label  $\mathbf{s}_c^l$  conveys how the network predictions of samples on certain class generally distribute across all classes, reflecting the relationships across different classes captured by the deep model. Consequently, the collection of soft labels from all classes could form a soft confusion matrix  $\mathcal{M}^l = [\mathbf{s}_1^l, \dots, \mathbf{s}_C^l]$ , which encodes the inter-class relationships among different categories of disease hence serve as the disease relation matrix.

**Reliable Disease Relation Estimation at Unlabeled Clients.** Since the data annotations are unavailable at unlabeled clients, we utilize the pseudo labels generated from model predictions to estimate the disease relation matrix. However, without sufficient task knowledge provided at unlabeled clients, the model predictions on unlabeled data could be noisy and inaccurate. We therefore employ an uncertainty-based scheme to filter out the unreliable model predictions, and only preserve the trustworthy ones to measure the reliable relation matrix.

Specifically, we take the local training at unlabeled client  $\mathcal{D}^u$  for instance. Given an input mini-batch  $\mathbf{x}^u$  of  $B$  image, we denote  $\mathbf{p}^u$  as the corresponding predicted probability and  $\mathbf{y}^u$  as the pseudo labels, i.e.,  $\mathbf{y}^u = \text{argmax}(\mathbf{p}^u)$ . Following the literature on uncertainty estimation, we approximate the uncertainty of model predictions with dropout of Bayesian networks [9]. Concretely, we perform  $T$ -time forward propagation for the input mini-batch  $\mathbf{x}^u$  under random dropout, obtaining a set of predicted probability vectors  $\{\mathbf{q}_t^u\}_{t=1}^T$ . The uncertainty  $\mathbf{w}^u$  is then estimated as the predictive entropy, which is computed from the averaged probability from the  $T$ -time forward passes as:

$$\mathbf{w}^u = - \sum_{c=1}^C \bar{\mathbf{q}}_{(c)}^u \log(\bar{\mathbf{q}}_{(c)}^u), \text{ with } \bar{\mathbf{q}}_{(c)}^u = \frac{1}{T} \sum_{t=1}^T \mathbf{q}_{t(c)}^u \quad (4)$$

where  $\mathbf{q}_{t(c)}^u$  is the value of the  $c$ -th class of  $\mathbf{q}_t^u$ . Since the predictive entropy has a fixed range, we can filter out the relatively unreliable predictions and only select the certain ones to compute the disease relation matrix. Hence, the per-category mean feature vectors  $\mathbf{v}_c^u$  (c.f. Eq. 2) at unlabeled clients are computed as:

$$\mathbf{v}_c^u = \frac{\sum_{i=1}^B \mathbb{1}_{[(\mathbf{y}_i=c) \cdot (\mathbf{w}_i^u < h)]} \cdot \mathbf{p}_i^u}{\sum_{i=1}^B \mathbb{1}_{[(\mathbf{y}_i=c) \cdot (\mathbf{w}_i^u < h)]}} \quad (5)$$

where  $h$  is the threshold to select the certain predictions from  $\mathbf{w}^u$ . Then, following the same operation as Eq. 3, the disease relation matrix at unlabeled client  $u$  is estimated as  $\mathcal{M}^u = [\mathbf{s}_1^u, \mathbf{s}_2^u, \dots, \mathbf{s}_C^u]$ .

### 2.3 Objective of Inter-client Relation Matching

With the above basis, we aim to enforce the unlabeled clients to produce similar disease relationships as labeled clients to preserve such discriminative task knowledge. Specifically, at the end of each federated round, the central server collects the relation matrix  $\mathcal{M}^l$  from each labeled client, and average them to compute a matrix representing the general disease relation information captured from all labeled data, i.e.,  $\mathcal{M} = \frac{1}{m} \sum_{l=1}^m \mathcal{M}^l$ . This obtained  $\mathcal{M}$  is then delivered to unlabeled clients to supervise their next round of local training. To establish

the supervision online, the relation matrix at unlabeled clients  $\mathcal{M}^u$  is estimated from each mini-batch during training. Finally, the inter-client relation matching loss is designed by minimizing the KL divergence between  $\mathcal{M}$  and  $\mathcal{M}^u$  as:

$$\mathcal{L}_{\text{IRM}} = \frac{1}{C} \sum_{c=1}^C (\mathcal{L}_{\text{KL}}(\mathcal{M}_c || \mathcal{M}_c^u) + \mathcal{L}_{\text{KL}}(\mathcal{M}_c^u || \mathcal{M}_c)), \quad (6)$$

$$\text{with } \mathcal{L}_{\text{KL}}(\mathcal{M}_c || \mathcal{M}_c^u) = \sum_j \mathcal{M}_{c(j)} \log \frac{\mathcal{M}_{c(j)}}{\mathcal{M}_{c(j)}^u}$$

where  $\mathcal{M}_c \in \mathbb{R}^C$  denote the relation vector of class  $c$ , i.e.,  $\mathcal{M}_c = \mathbf{s}_c$ ; and  $\mathcal{M}_{c(j)}$  denote its  $j$ -th entry. Overall, the local learning objectives at labeled ( $\mathcal{L}^l$ ) and unlabeled ( $\mathcal{L}^u$ ) clients are respectively expressed as:

$$\mathcal{L}^l = \mathcal{L}_{ce}(\mathcal{D}^l, \theta^l) \quad \text{and} \quad \mathcal{L}^u = \lambda(\omega)(\mathcal{L}_c + \mathcal{L}_{\text{IRM}}) \quad (7)$$

where  $\mathcal{L}_{ce}$  is the cross entropy loss;  $\mathcal{L}_c$  is the traditional consistency regularization loss (c.f. Eq. 1);  $\lambda(\omega)$  is a warming up function regarding federated round  $\omega$ , which helps to reduce the effect of the learning at unlabeled clients when the model is underfitting at earlier federated rounds.

### 3 Experiments

#### 3.1 Dataset and Experimental Setup

We validate our method on two important medical image classification tasks, including: intracranial hemorrhage (ICH) diagnosis from brain CT and skin lesion classification from dermoscopy images.

**Task 1 - Intracranial hemorrhage diagnosis.** We perform ICH diagnosis with the RSNA ICH Detection dataset[25], which aims to classify CT slices into 5 subtypes of ICH disease. Since most images in this dataset are healthy without any of the subtypes, we randomly sample 25000 slices from the dataset which contain one of the 5 subtypes of ICH for evaluation. These samples are then randomly divided into 70%, 10% and 20% for training, validation and testing. Since multiple slices may come from the same patient in this dataset, we have ensured no overlapped patients exist across the three split for a valid evaluation.

**Task 2 - Skin lesion diagnosis.** We employ ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection[21] dataset for skin lesion diagnosis, which contains 10015 dermoscopy images in the official training set labeled by 7 types of skin lesions. As the ground truth of official validation and testing set was not released, we randomly divide the entire training set to 70% for training, 10% for validation and 20% for testing. We perform the same data pre-processing for the two tasks. Specifically, we first resized the original images from  $512 \times 512$  to  $224 \times 224$ . To employ the pre-trained model, we then normalized the images with statistic collected from ImageNet dataset before feeding them into the network.

**Experiment Setup.** To simulate the FL setting, we randomly partition the training set into 10 different subsets serving as 10 local clients. Following the practice in SSL [28], we evaluate the model performance under 20% labeled data

**Table 1.** Quantitative comparisons with state-of-the-arts on two different tasks.

Method	Client num		Metrics				
	Label	Unlabel	AUC	Sensitivity	Specificity	Accuracy	F1
Task 1: Intracranial hemorrhage diagnosis							
FedAvg [19]	10	0	90.48±0.31	64.33±1.13	92.68±0.43	89.94±0.92	63.94±1.20
FedAvg [19]	2	0	83.40±0.87	57.88±1.68	90.48±0.79	87.45±1.08	57.10±1.29
Fed-SelfTraining [32]	2	8	84.32±0.82	57.94±1.66	90.22±0.74	87.90±1.81	57.48±1.14
Fed-Consistency [30]	2	8	84.83±0.79	57.26±1.93	90.87±0.62	88.35±1.32	57.61±1.08
FedIRM (ours)	2	8	87.56±0.56	59.57±1.57	91.53±0.81	88.89±1.29	59.86±1.65
Task 2: Skin Lesion Diagnosis							
FedAvg [19]	10	0	94.82±0.32	75.11±1.82	94.87±0.35	95.24±0.21	70.16 ±1.21
FedAvg [19]	2	0	90.65±1.23	65.53±1.76	91.76±0.48	92.53±0.67	52.59±1.42
Fed-SelfTraining [32]	2	8	90.82±0.56	67.03±1.93	93.61±0.21	92.47±0.34	53.44±1.85
Fed-Consistency [30]	2	8	91.13±0.62	68.55±1.29	93.45±0.94	92.67±0.39	54.25±1.31
FedIRM (ours)	2	8	92.46±0.45	69.05±1.71	93.29±0.59	92.89±0.25	55.81±1.49

setting, i.e., two clients are labeled and the remaining eight are unlabeled in our case. Five metrics are used to extensively evaluate the classification performance, including AUC, Sensitivity, Specificity, Accuracy and F1 score. We report the results in form of average and standard deviation over three independent runs.

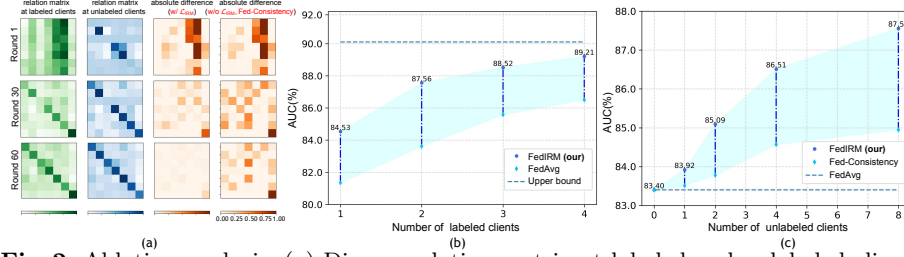
**Implementation Details.** We employ DenseNet121 [11] as the backbone for medical image classification. Two types of perturbations are utilized to drive the consistency regularization, including random transformation on input data (rotation, translation and flip) and dropout layer in the network. The temperature parameter  $\tau$  is empirically set as 2.0. The forward pass time  $T$  used to compute uncertainty is set as 8, and the threshold  $h$  to select reliable predictions is set as  $\ln 2$ . We follow [28] to apply a Gaussian warming up function  $\lambda(\omega) = 1 * e^{(-5(1-\omega/\Omega))}$ , where  $\Omega$  is set as 30. The local training adopts Adam optimizer with momentum of 0.9 and 0.99, and the batch size is 48 for both labeled and unlabeled clients. We totally train 100 federated rounds when the global model has converged stably, with the local training epoch  $e$  set as 1.

### 3.2 Comparison with State-of-the-arts.

We compare with recent FSSL methods, including **Fed-SelfTraining**[32], which performs self training at unlabeled clients by iteratively updating the model parameters and the pseduo labels of unlabeled data with expectation maximization; and **Fed-Consistency** [30], which employs the state-of-the-art SSL strategy, i.e., consistency regularization, to exploit the data at unlabeled clients (*without inter-client relation matching compared with our method*). We also compare with the **FedAvg** [19] model trained only with labeled clients or with all clients as labeled, which serve as the baseline and upperbound performance in FSSL.

The results on the two tasks are listed in Table 1. As observed, both Fed-SelfTraining and Fed-Consistency performs better than the baseline FedAvg model, which reflects the benefit to integrate the knowledge from additional unlabeled clients to improve FL models. Notably, compared with these methods, our FedIRM achieves higher performance on nearly all metrics, with 2.73%





**Fig. 2.** Ablation analysis. (a) Disease relation matrix at labeled and unlabeled clients under our method, as well as their absolute difference with or without  $\mathcal{L}_{IRM}$  (task 2); (b) Model performance under different labeled client number setting, using our approach and FedAvg (task 1); (c) Model performance as the number of unlabeled client increases (with labeled client number fixed), using our approach and Fed-Consistency (task 1).

and 1.33% AUC improvements on the two tasks over Fed-Consistency which does not employ our inter-client relation matching scheme. These clear improvements benefit from our FedIRM scheme which explicitly harnesses the discriminative relation information learned from labeled clients to facilitate the learning at unlabeled clients. Without the supervision from  $\mathcal{L}_{IRM}$ , the local training simply from consistency regularization is prone to forget the original task information, hence fail to fully exploit the discriminative information from unlabeled data.

### 3.3 Analytical Studies of Our Method

**Learning behavior under Inter-client Relation Matching.** Fig. 2(a) displays the disease relation matrix of labeled (first col.) and unlabeled clients (second col.) under our method, as well as their absolute difference under our method with (third col.) and without  $\mathcal{L}_{IRM}$  (i.e., Fed-Consistency, forth col.), at different federated rounds. As observed, the relationships between disease at labeled clients become increasingly clear as the federated training goes on, indicating that the model gradually captures such structural knowledge. Notably, the unlabeled clients in our method can well preserve such disease relationships, with highly consistent matrix patterns as labeled clients and low responses in the difference matrix. In contrast, the method without  $\mathcal{L}_{IRM}$  (i.e., Fed-Consistency) fails to do so and the responses in difference matrix are relatively high. This observation affirms the benefit to transfer such discriminative knowledge to facilitate the learning at unlabeled clients and also explains our performance gains.

**Effectiveness under different labeled client number.** We investigate the impact of different labeled client number in our method. As shown in the curve of Fig. 2(b), our method shows consistent improvements over the supervised-only FedAvg model under labeled client number from 1 to 4 (corresponds 10% to 40% labeled data setting in SSL). Importantly, using only 40% labeled client, our method achieves 89.21% AUC, which is very close to the upper-bound FedAvg model trained with 10 labeled clients (90.48%). This endorses the capability of our method to leverage the data from unlabeled clients for improving FL models.



**Effect of adding more unlabeled clients.** We finally analyze the effect of unlabeled client number on the performance of our FSSL method and Fed-Consistency, by fixing the labeled client number as 2 and gradually increasing the unlabeled client number in [1, 2, 4, 8]. As shown in Fig. 2(c), an interesting finding is the FSSL performance progresses as the unlabeled client number increases, indicating the potential in realistic scenarios to aggregate more widely-existing unlabeled clients to improve the FL models. Notably, our method consistently outperforms the Fed-Consistency method under different unlabeled client number, highlighting the stable capacity of our proposed FSSL learning scheme.

## 4 Conclusion

We present a new FSSL framework, which to our knowledge is the first method incorporating unlabeled clients to improve FL models for medical image classification. To address the deficiency of consistency regularization in FSSL, our method includes a novel inter-client relation matching scheme to explicitly utilize the knowledge of labeled clients to assist the learning at unlabeled clients. Experiments on two large-scale datasets demonstrate the effectiveness. Our method is extendable to non-IID scenario in FSSL setting, as the employed disease relations are independent of the observed clients and unaffected by image distributions.

## 5 Acknowledgement

The work described in this paper was supported in parts by the following grants: Key-Area Research and Development Program of Guangdong Province, China (2020B010165004), Hong Kong Innovation and Technology Fund (Project No. GHP/110/19SZ), Foundation of China with Project No. U1813204.

## References

1. Aviles-Rivero, A.I., Papadakis, N., Li, R., Sellars, P., Fan, Q., Tan, R.T., Schönlieb, C.: Graphx net – chest x-ray classification under extreme minimal supervision. In: Medical Image Computing and Comput.-Assisted Intervention. pp. 504–512. Springer (2019)
2. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–260. Springer (2017)
3. Chang, Q., Qu, H., Zhang, Y., Sabuncu, M., Chen, C., Zhang, T., Metaxas, D.N.: Synthetic learning: Learn from distributed asynchronous discriminator gan without sharing medical image data. In: CVPR. pp. 13856–13866 (2020)
4. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical image analysis **54**, 280–296 (2019)
5. Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C.: Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: Inf. Process. Med. Imaging. pp. 554–565 (2019)

6. 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.: Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. *NPJ digital medicine* **3**(1), 1–9 (2020)
7. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. *IEEE transactions on medical imaging* **39**(7), 2415–2425 (2020)
8. Dou, Q., So, T.Y., Jiang, M., Liu, Q., Vardhanabhuti, V., Kaissis, G., Li, Z., Si, W., Lee, H.H., Yu, K., 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)
9. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
10. Gyawali, P.K., Ghimire, S., Bajracharya, P., Li, Z., Wang, L.: Semi-supervised medical image classification with global latent mixing. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 604–613. Springer (2020)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
12. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* pp. 1–7 (2020)
13. Li, D., Kar, A., Ravikumar, N., Frangi, A.F., Fidler, S.: Federated simulation for medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 159–168. Springer (2020)
14. Li, W., Milletari, F., Xu, D., Rieke, N., Hancox, J., Zhu, W., Baust, M., Cheng, Y., Ourselin, S., Cardoso, M.J., et al.: Privacy-preserving federated brain tumour segmentation. In: International Workshop on Machine Learning in Medical Imaging. pp. 133–141. Springer (2019)
15. Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S.: Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical Image Analysis* **65**, 101765 (2020)
16. Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.A.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. *CVPR* (2021)
17. Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A.: Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging* **39**(11), 3429–3440 (2020)
18. Mathur, S., Dinakarpandian, D.: Finding disease similarity based on implicit semantic similarity. *Journal of biomedical informatics* **45**(2), 363–371 (2012)
19. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282 (2017)
20. Oerton, E., Roberts, I., Lewis, P.S., Guillems, T., Bender, A.: Understanding and predicting disease relationships through similarity fusion. *Bioinformatics* **35**(7), 1213–1220 (2019)
21. P. Tschandl, C. Rosendahl, H.K.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* **5** (2018)

22. Razzak, M.I., Naz, S., Zaib, A.: Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps* pp. 323–350 (2018)
23. 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.: The future of digital health with federated learning. *NPJ digital medicine* **3**(1), 1–7 (2020)
24. Roth, H.R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., Gupta, S., Qu, L., Ihsani, A., Bizzo, B.C., et al.: Federated learning for breast density classification: A real-world implementation. In: *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pp. 181–191. Springer (2020)
25. RSNA: Intracranial hemorrhage detection challenge. In: <https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/> (2019)
26. Sheller, M.J., Reina, G.A., Edwards, B., Martin, J., Bakas, S.: Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In: *Brainlesion Workshop, MICCAI*. pp. 92–104. Springer (2018)
27. Silva, S., Gutman, B.A., Romero, E., Thompson, P.M., Altmann, A., Lorenzi, M.: Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In: *ISBI*. pp. 270–274. IEEE (2019)
28. Tarvainen A., V.H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: *Adv. Neural Inf. Process. Syst.* (2017)
29. Wang, D., Zhang, Y., Zhang, K., Wang, L.: Focalmix: Semi-supervised learning for 3d medical image detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3951–3960 (2020)
30. Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H.R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., et al.: Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from china, italy, japan. *Medical Image Analysis* p. 101992 (2021)
31. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: *Medical Image Computing and Comput.-Assisted Intervention*. pp. 605–613. Springer (2019)
32. Zhang, Z., Yao, Z., Yang, Y., Yan, Y., Gonzalez, J.E., Mahoney, M.W.: Benchmarking semi-supervised federated learning. *arXiv preprint arXiv:2008.11364* (2020)