RSCFed: Random Sampling Consensus Federated Semi-supervised Learning

Xiaoxiao Liang[1], Yiqun Lin[1], Huazhu Fu[2], Lei Zhu[3,1], Xiaomeng Li[1‡]
[1] The Hong Kong University of Science and Technology, [2] IHPC, A*STAR
[3] The Hong Kong University of Science and Technology (Guangzhou)
{xliangak, ylindw}@connect.ust.hk, hzfu@ieee.org, {leizhu, eexmli}@ust.hk

Abstract

*Federated semi-supervised learning (FSSL) aims to derive a global model by training fully-labeled and fully-unlabeled clients or training partially labeled clients. The existing approaches work well when local clients have independent and identically distributed (IID) data but fail to generalize to a more practical FSSL setting, i.e., Non-IID setting. In this paper, we present a **R**andom **S**ampling **C**onsensus **Fed**erated learning, namely **RSCFed**, by considering the uneven reliability among models from fully-labeled clients, fully-unlabeled clients or partially labeled clients. Our key motivation is that given models with large deviations from either labeled clients or unlabeled clients, the consensus could be reached by performing random sub-sampling over clients. To achieve it, instead of directly aggregating local models, we first distill several sub-consensus models by random sub-sampling over clients and then aggregating the sub-consensus models to the global model. To enhance the robustness of sub-consensus models, we also develop a novel distance-reweighted model aggregation method. Experimental results show that our method outperforms state-of-the-art methods on three benchmarked datasets, including both natural and medical images. The code is available at https://github.com/XMed-Lab/RSCFed.*

1. Introduction

The core idea of federated learning (FL) is to train machine learning models on separate datasets that are distributed across different places or devices, which can preserve local data privacy to a certain extent. Over the past few years, FL has emerged as an important research area and attracted many researchers' attention to study its application in medical image diagnosis [10, 14, 28], image classification [16] and object detection [22].

Considerable efforts have been devoted to design various FL methods, such as FedAvg [23], SCAFFOLD [12] and MOON [16]. Although the results are quite promising, these methods require fully labeled images on each local client, limiting its application in real practice.

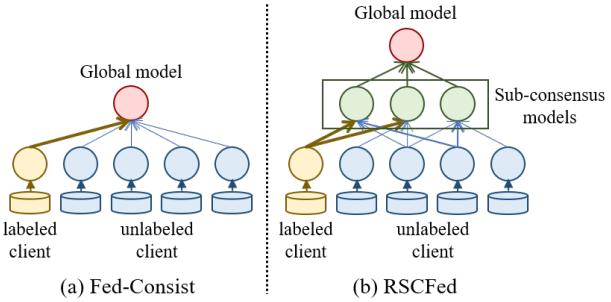

Figure 1. Illustration of the existing FSSL method, *i.e.*, [28] and our RSCFed. Existing methods simply perform the standard model aggregation in FedAvg [23], while our RSCFed distills multiple sub-consensus models from local models and updates the global model via aggregating sub-consensus models.

Recently, federated semi-supervised learning (FSSL) [8, 19, 21, 28] is becoming a new research topic, aiming at utilizing the unlabeled images to enhance the global model development. One line of the research studies FSSL by considering each client has partially labeled and unlabeled images. For example, Jeong *et al*. [8] introduced inter-client consistency loss to improve the global model by encouraging the consistent outputs from multiple clients. Another line of FSSL [21, 28] assumes that some local clients have fully labeled images while some clients contain unlabeled images, which we denote as labeled clients and unlabeled clients respectively. However, existing methods have two main limitations. First, they do not consider not independent and identically distributed data (Non-IID) among local clients, which is a key problem for FL and can cause a deterioration in accuracy [9, 15]. Second, some solutions [21] share the correlation matrix among local clients, which might cause information leakage.

This paper studies the FSSL with two widely used settings: (1) jointly training fully-labeled and fully-unlabeled clients; (2) jointly training partially-labeled clients. A straightforward solution is to extend existing FSSL meth-

‡Project lead and corresponding author.

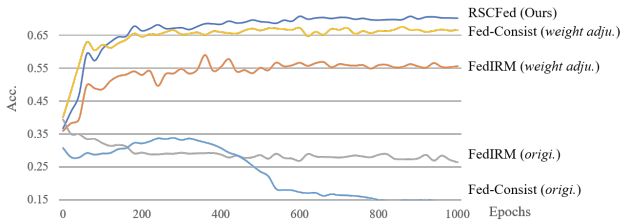

Figure 2. Comparisons of test accuracy curve of our RSCFed with FedIRM [21] and Fed-Consist [28] under original setting (*origi.*) and under our weight adjusting setting (*weight adju.*).

ods [21, 28] to the Non-IID setting. However, both FedIRM [21] and Fed-Consist [28] fail to generalize to the Non-IID setting. This is because FedIRM [21] proposed to share an inter-class correlation matrix among clients based on the assumption that each client has the same class relationship. However, the class relationship can not be correctly learned due to the heterogeneous data among local clients in the Non-IID setting, thus hurting the model performance. Fed-Consist [28] proposed to equally average the model weights from labeled and unlabeled clients. However, the performance significantly decreases when unlabeled clients increases since the global model may be dominated by the unlabeled clients. Adjusting aggregation weights for labeled and unlabeled clients is one solution, *i.e.*, increasing the weights for labeled clients while decreasing the weights for unlabeled ones. Nevertheless, this result achieves limited performance; see *(weight adju.)* in Fig. 2.

To this end, we present **R**andom **S**ampling **C**onsensus **Fed**erated learning, namely **RSCFed**, by considering the ***uneven reliability*** among models either from fully-labeled and fully-unlabeled clients or from several partially-labeled clients under the Non-IID setting without ***any information leakage among clients***. For example, labeled clients are easily biased to local data, while unlabeled clients are difficult to achieve the high accuracy, leading to uneven model reliability among clients. On the other hand, training with several partially-labeled clients may also cause uneven model reliability because images in each client are heterogeneously distributed in quantity skew and label skew. To achieve a robust global model, our key idea is *to regard the local models as noisy models and distill several consensus models via random sampling before aggregating to the global model*, as shown in Fig. 1. Specifically, in each synchronization round, we randomly sub-sample clients and record the averaged weights from the sub-sampled models as ***a sub-consensus model.*** By performing the operation multiple times, we update the global model via aggregating multiple sub-consensus models. To distill a robust sub-consensus model from randomly sampled local clients, we introduce a distance-reweighted model aggregation (DMA) module, which dynamically increases the weights for models that are close to the sub-consensus model and vice versa. The idea shares a similar spirit with random sample consensus (RANSAC) [5], which identifies points as outliers if they are far from the model. We conduct extensive experiments on natural image classification datasets (e.g., SVHN and CIFAR-100) and medical dataset (i.e., ISIC 2018 Skin) to demonstrate the effectiveness of RSCFed. Overall, our main contributions can be summarized as follows:

- In this paper, we present a novel FSSL method, named RSCFed, to address the uneven reliability of Non-IID local clients. Unlike existing FSSL frameworks that directly aggregate local clients, RSCFed proposes the concept of updating the global model via aggregating multiple sub-consensus models.
- To improve the sub-consensus model, we introduce a novel distance-reweighted model aggregation (DMA) module, which dynamically adjusts the weights of each sampled local client to the sub-consensus model.
- Experiments on three public datasets demonstrate that our RSCFed significantly outperforms the other state-of-the-art FSSL methods. We further show that with larger ratio of unlabeled data involved, the better improvement RSCFed can achieve.

2. Related Work

2.1. Federated Learning with Non-IID

Federated learning provides multi-institutional data collaboration solutions for model training under a data-decentralized scheme [13, 29]. Two common problems in this field are system heterogeneity and statistical heterogeneity, which refer to the inconsistency of computational abilities and data distribution among clients. A pioneering work provided the most widely recognized FL baseline, FedAvg [23], followed by many heterogeneous FL solutions, which could be categorized into two branches: local training-oriented methods [16] and model aggregation-oriented methods.

Local Training-oriented Methods As for local training-oriented methods, Li *et al*. [17] add an additional regularization term in local objectives, representing the distance between the global model and local model, thus giving constraints on the model drift. Besides, Karimireddy *et al*. [12] prove control variates to correct local model update, and Li *et al*. [16] introduce a contrastive loss term to prevent local models from their local minimum. Several other methods perform inter-client privacy-invariant information exchange [20, 30]. However, most existing methods for Non-IID data fail under the FSSL setting due to the uneven model reliability from labeled and unlabeled clients. Besides, some methods exchange the information among clients [20, 30], which may have the potential for information leakage. Unlike these methods, we do not share any information among clients.

Model Aggregation-oriented Methods As for improvements on model aggregation, Wang *et al*. [27] normalize the received local gradients before averaging; Wang *et al*. [26] perform layer-wise averaging with Bayesian nonparametric methods; Chen *et al*. [2] regard the known global and local models as samples from an assumed distribution, where another set of models are sampled as teacher models and are later utilized in server-side knowledge distillation under the assumption that unlabeled data could be kept at the server. Zhang *et al*. [33] further extends a single global model to multiple global models, in which affinity towards all global model candidates are computed at each client. Finally, global models are weighted averaged by affinity in a personalized manner according to each client. However, these methods are developed for supervised federated learning, while our work focuses on FSSL with uneven model reliability from labeled and unlabeled clients.

2.2. Semi-Supervised Learning

Standard semi-supervised learning aims to optimize a model with both labeled and unlabeled data in a centralized manner. The learning paradigm usually involves smoothness-based consistency regularization [4, 18, 25, 31], entropy minimization-based self-training methods, [3, 35], and their combinations [1, 7, 24]. For instance, Zou *et al*. [35] fuse the decoder prediction and self-attention Grad-CAM from weakly augmented images to obtain a reliable pseudo label, with which the prediction of strongly augmented image could be supervised. Self-training and co-training-based methods also gained popularity in data-centralized data schemes. However, these methods require labeled images and unlabeled images during the training process. While, in FSSL setting, the labeled and unlabeled images are decentralized to labeled and unlabeled clients, respectively. Instead of studying how to get a good model with labeled and unlabeled images, this paper presents a novel method on model aggregation with uneven model reliability from labeled and unlabeled clients.

2.3. Federated Semi-Supervised Learning

FSSL can be broadly classified into two categories. One category assumes that every local client contains partially labeled images. For instance, Jeong *et al*. [8] and Lin *et al*. [19] let each client hold labeled and unlabeled data simultaneously. Besides, Jeong *et al*. [8] and Zhang *et al*. [34] assume labeled data is available only at the server, and Kang *et al*. [11] assume labeled and unlabeled data are isolated but inter-client sample overlapping exists.

Another category considers that some clients are fully labeled while some clients contain unlabeled images. For example, Liu *et al*. [21] propose to learn inter-class relationship, which is learned from labeled clients and shared among labeled and unlabeled clients. However, this method fails under the Non-IID setting, as inter-class correlations are no longer similar among clients due to data heterogeneity. Besides, Yang *et al*. [28] introduce a consistency-based method, in which different augmentations were applied to unlabeled images with their predictions similarity maximized. While the consistency loss still works with heterogeneous data, only one unlabeled client was involved in their method. However, we found that these methods fail to generalize to the Non-IID setting. Our proposed RSCFed shows its robustness towards uneven model reliability under FSSL.

3. Methodology

Fig. 3 shows an overview of our RSCFed. With some labeled and unlabeled local clients, our RSCFed respectively performs the following steps in each round: (1) Randomly sample local clients; (2) Assign current global model to selected clients as initialization, and conduct local training on selected clients; (3) Collect models from selected clients, execute distance-reweighted model aggregation (DMA) to obtain a sub-consensus model; (4) Repeat step (1)-(3) multiple times to obtain a set of sub-consensus models; (5) Aggregate a new model from the sub-consensus models set to be the next global model.

3.1. FSSL Setting

In the methodology, we consider the FSSL with fully-labeled and fully-unlabeled clients. Assume there are m labeled clients denoted as $\{C_1, ..., C_m\}$, and each of them has a local dataset, $\mathcal{D}^l$, defined as $\mathcal{D}^l = \left\{\left(X_i^l, y_i^l\right)\right\}_{i=1}^{N^l}$. Similarly, there are n unlabeled clients denoted as $\{C_{m+1}, ..., C_{m+n}\}$, and each has a dataset $\mathcal{D}^u$ containing N^u unlabeled data $\mathcal{D}^u = \{(X_i^u)\}_{i=1}^{N^u}$. Our goal is to derive a good global model θ_{glob} by utilizing both labeled and unlabeled data in a decentralized scheme.

3.2. Local Training

All local models are initialized with the current global model θ_{glob}^t at the beginning of t^{th} synchronization round. Our proposed RSCFed adopts standard supervised and unsupervised training on labeled and unlabeled clients, respectively. For simplification, we default all representations in this section occur in t^{th} synchronization round.

Labeled clients For local training on labeled clients, we adopt cross-entropy loss $\mathcal{L}_{CE}$ as the main objective:

$$\mathcal{L}_{CE} = -y_i \log(\hat{y}_i), \tag{1}$$

where $\hat{y}_i$ is the prediction of local data from the local model. The client then returns θ_l to server after training.

Unlabeled clients Unlabeled clients adopt mean-teacher-based consistency regularization framework, and regard student model as the local model. The teacher model θ_{tea} is

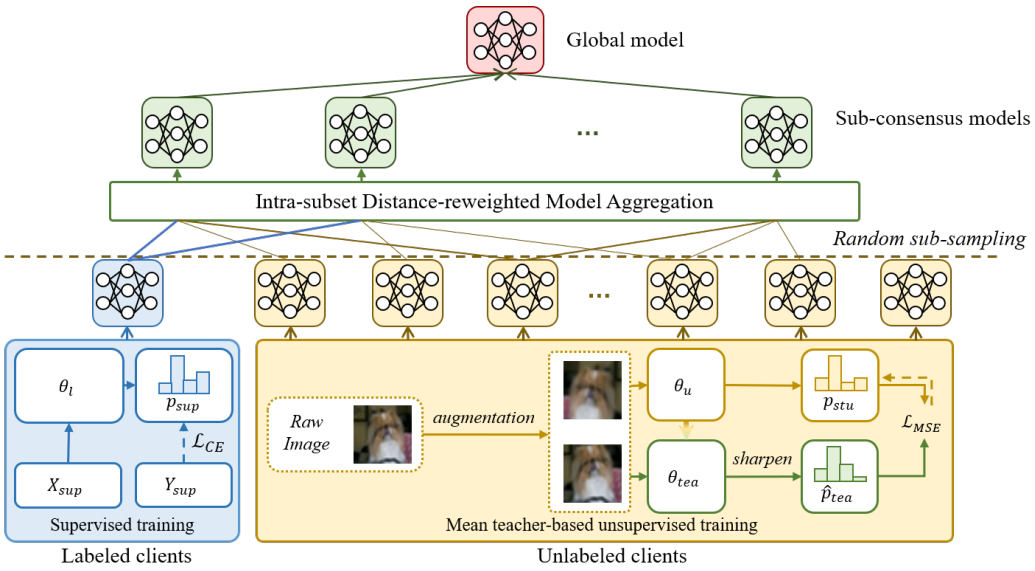

Figure 3. An overview of our proposed RSCFed. The labeled and unlabeled clients optimize by supervised cross-entropy loss $\mathcal{L}_{CE}$ and mean-teacher-based consistency loss $\mathcal{L}_{MSE}$, respectively. Our RSCFed performs multiple random sub-sampling among all clients with distance-reweighted model aggregation (DMA) to increase the weights for clients that are close to the sub-consensus model and visa versa. This module can help avoid the influence of a deviated local model to the global model.

initialized with θ^0_{glob} when this client is the first time selected. In each local iteration on unlabeled clients, a batch of input images is augmented twice and separately fed into the student and the teacher models. After their predictions p_{stu} and p_{tea} are generated, we utilize "sharpening" defined in [1] to increase the temperature of teacher's predictions:

$$\hat{p}_i = \text{Sharpen}(p_{tea}, \tau)_i = p_i^{\frac{1}{\tau}} \Big/ \sum_j p_j^{\frac{1}{\tau}}, \tag{2}$$

where p_i and $\hat{p}_i$ refer to each element in p_{tea} before and after sharpening respectively, and τ is the temperature parameter. Thus p_{tea} is "sharpened" to $\hat{p}_{tea}$, and the sample is pushed away from the decision boundary to generate better targets for consistency alignment. With the two predictions of differently augmented input, the mean-square-error loss is adopted as the local objective on unlabeled clients:

$$\mathcal{L}_{MSE} = \|\hat{p}_{tea} - p_{stu}\|_2^2 . \tag{3}$$

Note that only the student model is updated via Eqn. (3), and the teacher model receives student model parameters via exponential moving average after each local iteration:

$$\theta_{tea} = \alpha\theta_{stu} + (1-\alpha)\,\theta_{tea}, \tag{4}$$

where α is the momentum parameter. The unlabeled client finally return the student model as its local model θ_u.

3.3. Random Sampling Consensus FL

We propose RSCFed, a novel FSSL framework with random subset sampling and distance-reweight model aggregation, to obtain a more robust global model from heavily biased local models. To be more specific, we randomly subsample over all clients and collect models they uploaded to dig their underlying consensus. Then, we obtain a sub-consensus model by aggregating collected models, where a distance-reweighted model aggregation (DMA) strategy is introduced to dynamically adjust their weights. We repeat these two steps for M times to obtain a set of sub-consensus models. Finally, we aggregate the sub-consensus models set to obtain a global model in each round.

Multiple Random Sub-sampling. Random sub-sampling is proposed to distill a sub-consensus model. We propose to perform multiple random sub-sampling to get multiple sub-consensus models. To achieve it, at the beginning of the synchronization round t, we perform M times independent random subsampling to sample K clients. The server then sends global model θ^t_{glob} to sampled clients, followed by executing local training on sampled clients. Note that if the clients are sampled multiple times in a round, we do not need to send the global model for initialization again to save communication costs.

Distance-reweighted Model Aggregation. To enhance the robustness of the sub-consensus model, instead of aggregating multiple selected clients like FedAvg [23], we propose a novel distance-reweighted model aggregation (DMA). Our key idea is to dynamically increase the weights for models that are close to the average model and vice versa. For local models of sampled clients, we perform model aggregation with a model distance-based re-weighting strategy we design. For each subset, we firstly compute an intra-subset averaged model θ_{avg}:

$$N_{total} = \sum_{i=1}^{K} N_i, \text{ and } \theta_{avg} = \sum_{i=1}^{K} \frac{N_i}{N_{total}}\theta_i, \tag{5}$$

Algorithm 1: The RSCFed framework

Input: θ^t_{glob}: the global model from $t-1^{th}$ round; N: number of clients; M: number of subsets; K: number of clients in each subset

Output: θ^{t+1}_{glob} from t^{th} round

for $m \leftarrow 0$ **to** M **do**
  Randomly select $\{C_i\}_{i=1}^{K}$ from N clients
  for $k \leftarrow 0$ **to** K **do**
    send global model θ_{glob} to C_k;
    $\theta_k \leftarrow$ **LocalTraining**(k, θ_{glob})
  $\bar{\theta} \leftarrow$ **Avg**$(\theta_k, k = 0toK-1)$ Eqn. (5);
  $\bar{w}_k \leftarrow$ **ReWeight**$(\theta^m_k, \bar{\theta}^m)$ Eqn. (6);
  $\theta^m_{sub} \leftarrow \sum_{k=0}^{K-1} \bar{w}_k \theta^m_k$

Return $\theta^{t+1}_{glob} \leftarrow \frac{1}{M} \sum_{m=0}^{M-1} \theta^m_{sub}$

where θ_i represents the i^{th} local model of the subset, N_i stands for its local data amount, and K denotes the number of clients in a subset. Instead of simply averaging local clients, our DMA dynamically scales w_i for i^{th} client in each subset, as follows:

$$w_i = \frac{N_i}{N_{total}} \exp\left(-\beta \cdot \frac{\|\theta_i - \theta_{avg}\|_2}{N_i}\right), \text{ and } \bar{w}_i = \frac{w_i}{\sum_j w_j}, \tag{6}$$

where β is a hyper-parameter and $\|\theta_i - \theta_{avg}\|_2$ refers to L_2 Norm of the model gradient between i^{th} local model and temporal averaged model within the subset. The model distance is divided by local data quantity n_i to reduce the impact of local iterations on model drift. We then normalize the intra-subset model weight to [0, 1].

After obtaining a set of sub-consensus models, we denote their equally weighted average to be the final global model θ_{glob}:

$$\theta^{t+1}_{glob} = \frac{1}{M} \sum_{m=0}^{M-1} \theta^m_{sub}, \tag{7}$$

where θ^m_{sub} denotes the m^{th} sub-consensus model. Then $t+1^{th}$ synchronization round is executed with θ^{t+1}_{glob} as initialization. The whole updating in t^{th} synchronization round of our RSCFed is presented in Algorithm 1.

4. Experiments

To demonstrate the effectiveness and robustness of our proposed RSCFed, we conduct experiments on 3 benchmark datasets, and further evaluate RSCFed under intensive settings like different unlabeled data ratio, limited communication cost, *etc*.

4.1. Dataset and Experimental Setup

Benchmark Datasets. We evaluate the effectiveness of our proposed method on two natural image classification datasets, i.e., SVHN and CIFAR-100. Moreover, to simulate the realistic privacy data decentralized-distributed scenario, we evaluate our method on ISIC 2018 (Skin Lesion Analysis Towards Melanoma Detection) consisting of 10,015 dermoscopy images with seven types of skin lesions. For all three benchmark datasets, 80% images of each dataset are randomly selected for training, and the remaining images are for testing. For SVNH and CIFAR-100, we resize the original 32×32 images of these two datasets to 40×40 pixels, randomly crop a 32×32 region, and then utilize a normalization operation on the cropped region to generate the input of our network. Regarding ISIC 2018, we resize the spatial resolution of the original image from 600×450 to 240×240, randomly crop a 224×224 region, and normalize the cropped region as the network input.

Feature extraction backbone. When training on SVHN and CIFAR-100, we follow [16] to employ a simple CNN as the feature extraction backbone, which contains two 5×5 convolution layers, a 2×2 max-pooling layer, and two fully-connected layers. For the ISIC 2018 dataset, we utilize ResNet-18 [6] as the feature extraction backbone. After that, we employ a two-layer MLP and a fully-connected layer to formulate a classification network at each client for all datasets. Moreover, the same classification network is also utilized at each client of the compared methods for a fair comparison.

Federated Learning setting. We follow existing methods [16,26,32] to use a Dirichlet distribution $Dir(\gamma)$ (γ=0.8 for all three benchmark datasets) to generate the non-IID data partition in clients. After such a Non-IID data partition strategy, the number of classes and samples at each client differ from each other, and thus not all clients contain samples from all classes.

Implementation Details. We utilize the SGD optimizer, and implement our method with PyTorch. The learning rates in the labeled client and the unlabeled clients are empirically set to 0.03 and 0.021 for all methods on SVHN and CIFAR-100, and 0.002 and 0.001 for ISIC 2018. The batch size is set to 64 for SVHN and CIFAR-100, and 12 for ISIC 2018. We train 1000 synchronization rounds for all datasets to make the global model stably converged, and the local training epoch is set to 1. The number of sub-sampling operations M and the number of local clients used in each sub-sampling operation K are set as: M=3, and K=5. Our method has three parameters: the momentum parameter α of Eqn. (4), temperature parameter τ of Eqn. (2), and the scaling factor β of Eqn. (6). And we empirically set α=0.001, τ=0.5 for all three benchmark datasets. The scaling factor β is set to 10,000 for SVHN and CIFAR-100,

Table 1. Results on SVHN, CIFAR-100, and ISIC 2018 datasets under heterogeneous data partition. Note that FedIRM [21] and Fed-Consist [28] fail to generalize in Non-IID setting. The results reported in this Table are performed with weight adjusting; see Fig. 2

Labeling Strategy	Method	Client Num.		Metrics			
		labeled	unlabeled	Acc. (%)	AUC. (%)	Precision (%)	Recall (%)
		Dataset 1: SVHN					
Fully supervised	FedAvg [23] (upper-bound)	10	0	82.05	97.82	81.59	77.90
	FedAvg [23] (lower-bound)	1	0	60.54	91.23	64.38	57.34
Semi supervised	FedIRM [21]	1	9	55.69	91.19	66.78	56.40
	Fed-Consist [28]	1	9	66.94	94.19	68.92	66.75
	RSCFed (ours)	1	9	**70.26**	**95.54**	**73.36**	**68.46**
		Dataset 2: CIFAR-100					
Fully supervised	FedAvg [23] (upper-bound)	10	0	25.87	90.44	29.97	26.01
	FedAvg [23] (lower-bound)	1	0	12.02	76.03	10.76	11.58
Semi supervised	FedIRM [21]	1	9	14.11	79.22	14.64	14.03
	Fed-Consist [28]	1	9	13.89	78.31	15.12	12.95
	RSCFed (ours)	1	9	**15.82**	**81.41**	**15.85**	**16.37**
		Dataset 3: ISIC 2018: Skin Lesion Classification					
Fully supervised	FedAvg [23] (upper-bound)	10	0	84.07	95.64	76.68	62.97
	FedAvg [23] (lower-bound)	1	0	68.14	84.12	41.91	38.61
Semi supervised	FedIRM [21]	1	9	68.10	84.11	41.96	38.94
	Fed-Consist [28]	1	9	68.74	84.71	41.91	38.63
	RSCFed (ours)	1	9	**70.26**	**86.01**	**45.65**	37.91

Table 2. Quantitative results of our method and the backbone model [28] without the multiple sub-sampling operations and the distance-weighted aggregation mechanism on the three benchmark datasets. "SSO" denotes the multiple sub-sampling operation with model aggregation, while "DMA" represents the distance-weighted model aggregation mechanism.

	SSO	DMA	Metrics	
			Acc. (%)	AUC (%)
			Dataset 1: SVHN	
Basic	×	×	66.94	94.19
Basic + SSO	✓	×	69.15	95.2
RSCFed (ours)	✓	✓	**70.26**	**95.54**
			Dataset 2: CIFAR-100	
Basic	×	×	13.89	78.3
Basic + SSO	✓	×	14.92	81.8
RSCFed (ours)	✓	✓	**15.82**	**81.4**
			Dataset 3: ISIC 2018	
Basic	×	×	68.74	84.7
Basic + SSO	✓	×	69.85	85.5
RSCFed (ours)	✓	✓	**70.26**	**86.0**

and 0.01 for ISIC 2018.

4.2. Results with labeled and unlabeled clients

FSSL setting. In this setting, the training dataset contains ten clients: one labeled client with labeled images and nine unlabeled clients with only unlabeled samples. Furthermore, the same FSSL training dataset is utilized to train our network and state-of-the-art methods for a fair comparison.

Implementation details. Note that the original work in [21, 28] reach very limited result with enough labeled data when all local models are aggregated via FedAvg [23], see Fig. 2. Hence, we re-implement [28], try increased aggregation weight for labeled client from the set $\{20\%, 30\%, 50\%, 70\%\}$. Our experiments show that 50% achieves the best classification accuracy. Hence, we empirically enlarge the weight of labeled client to about 50%, and other nine unlabeled clients share the remaining 50% weight in each FSSL synchronization round. Such aggregating weight is also applied to guarantee the deep models performance when we re-implement FedIRM [21] and our RSCFed.

Table 3. Comparison of our method (RSCFed) against FedIRM [21] and Fed-Consist [28] with number of labeled and unlabeled clients set to 2 and 8.

Method	Metrics			
	Acc.(%)	AUC(%)	Precision(%)	Recall(%)
FedIRM(origi.) [21]	59.75	87.4	67.55	55.26
FedIRM(weight adju.) [21]	74.10	94.8	76.49	70.45
Fed-Consist [28]	75.52	96.4	77.75	70.30
RSCFed(ours)	**76.65**	**96.7**	**78.61**	**73.16**

Compared methods. We compare our network against state-of-the-art FSSL methods, including (1) FedIRM [21], which computes an inter-class relationship labeled clients and utilizes it as extra supervisions for unlabeled clients; (2) Fed-Consist [28], which computes a consistency loss on predictions from multiple augmented inputs for unlabeled data in a mean teacher framework [25]. We also compare our network against FedAvg [23] trained with all 10 labeled clients as the upper-bound classification result, and FedAvg [23] trained with all 1 labeled clients as the lower-bound classification result; see Table 1. Moreover, we introduce four widely-used metrics to compare different methods, and they are Accuracy, Area under the ROC Curve (AUC), Precision, and Recall.

Quantitative comparisons. Table 1 reports the quantitative results of our network and state-of-the-art methods on three benchmark datasets in terms of four metrics. Basically, we can find that the results of the two compared

Table 4. Ablation study on our method (RSCFed) in terms of different Unlabeled Client numbers and a comparison with a SOTA FSSL method (i.e., Fed-Consist [28]).

Total client numbers	Client splitting		Fed-Consist [28]		Our RSCFed		Improvements	
	Labeled	Unlabeled	Acc.(%)	AUC.(%)	Acc.(%)	AUC.(%)	Acc.(%)	AUC.(%)
5	1	4	67.82	95.3	69.33	95.8	**1.51**	**0.5**
10	1	9	66.94	94.2	70.26	95.5	**3.32**	**1.3**
15	1	14	69.65	94.3	73.19	95.6	**3.54**	**1.3**
25	1	24	60.28	89.3	63.79	90.9	**3.51**	**1.6**
35	1	34	56.08	90.6	59.82	92.8	**3.74**	**2.2**
50	1	49	56.20	88.0	60.18	91.5	**3.98**	**3.5**

FSSL methods (i.e., FedIRM [21] and Fed-Consist [28]) and our network are between the upper-bound results and the lower-bound result obtained by FedAvg [23] for all three benchmark datasets. From these quantitative results, we can observe that our proposed RSCFed has a superior metric performance over all competitors on the three benchmark datasets. Our superior performance over Fed-Consist indicates a generalization ability enhancement obtained by the aggregation strategy in our network. Moreover, our network also outperforms FedIRM in terms of four metrics on three datasets. The reason behind is that the consistent assumption of inter-class relationship among clients is not correct due to non-IID data distribution on all clients in our work.

Evaluation on SVHN. Regarding two compared methods, Fed-Consist has the best Accuracy performance of 66.94%, the best AUC performance of 94.19%, the best Precision performance of 68.92%, and the best Recall performance of 66.75%. More importantly, our method has larger metric scores than Fed-Consist, and achieves an Accuracy of 70.29% (3.32% improvement), an AUC of 95.54% (1.35% improvement), a Precision of 73.36% (4.44% improvement), and a Recall of 68.46% (1.71% improvement).

Evaluation on CIFAR-100. Regarding CIFAR-100, FedIRM has a larger Accuracy score of 14.11%, and a larger AUC score of 79.22%, and a larger Recall score of 14.03%, while Fed-Consist has a larger Precision score of 15.12%. Compared to these two state-of-the-art methods, our network improves the Accuracy score from 14.11% to 15.82%, the AUC score from 79.22% to 81.41%, the Precision score from 15.12% to 15.85%, and improves the Recall score from 14.03% to 16.37%.

Evaluation on ISIC 2018. Although Fed-Consist has a larger Recall score than our method, our method also achieves the best Accuracy score of 70.26%, and the best AUC score of 86.01%, and the best Precision score of 45.65% among all three compared methods. It indicates that our federated semi-supervised learning method has a higher classification accuracy for ISIC 2018.

4.3. Results with partially labeled clients

FSSL setting. To better elaborate the ability of RSCFed in solving uneven model reliability, we further extend RSCFed to another line of FSSL, where all local clients are partially labeled, *i.e.*, only 10% images are labeled. For this setting, we adopt same network backbone as in the previous setting. Since all clients are partially labeled, no weight scaling operation is performed.

Table 5. Results with partially labeled clients on SVHN dataset.

Method	Metrics			
	Acc.(%)	AUC(%)	Precision(%)	Recall(%)
Fed-Consist [28]	77.54	96.63	77.90	74.11
RSCFed(ours)	79.01	97.05	79.19	75.49

Results Table 5 shows our method and our baseline, *i.e.*, Fed-Consist [28] on SVHN dataset. Note that since the method in [21] requires extra supervision from fully labeled clients and cannot generalize to this setting, we do not list their results here. From Table 5 we can see that our RSCFed still outperforms Fed-Consist [28] by more than 1% in most metrics. To be more specific, our work made 1.47% improvement in Accuracy, 1.29% in Precision score, 1.38% in Recall score, and 0.42% in AUC score.

4.4. Ablation Studies

We further conduct ablative experiments to evaluate the effectiveness of the major components (sub-sampling and aggregation strategy) of our RSCFed, and further discuss its performance in terms of different unlabeled ratio, different communication cost limitations, and different hyper-parameters. All experimental results in this section are evaluated on SVHN dataset unless separately clarified.

Effectiveness of SSO and DMA. To evaluate the effectiveness of the multiple sub-sampling operations (SSO) and the distance-reweighted model aggregation (DMA), we perform an ablation study on three benchmark datasets. Table 2 compares the Accuracy and AUC scores of quantitative results of our method and two baseline networks (i.e., "Basic+SSO" and "Basic"). From these quantitative results, we can find that SSO and DMA have significant contributions to the success of our method in FSSL scenario. By observing the quantitative results of "Basic+SSO" and "Basic", we can find that our SSO increases the accuracy score of 2.21% and the AUC score of 1.01% on SVHN, the accuracy score of 1.03% and the AUC score of 3.5% on CIFAR-100, as well as the accuracy score of 1.11% and the AUC score of 0.8% on ISIC 2018. Moreover, the DMA of our method

Table 6. Ablation study of our method (RSCFed) in terms of different communication costs and comparisons them against Fed-Consist. "Com. cost" (stands for communication cost) denotes as how many times as much as that of the state-of-the-art method (i.e., Fed-Consist [28]).

Method	Client num.	Com. cost	Metrics	
			Acc. (%)	AUC (%)
Fed-Consist [28]	10	1.0×	66.94	94.2
Our RSCFed	8	0.8×	68.23	94.4
	9	0.9×	69.25	95.0
	10	1.0×	69.54	95.2
	15	1.5×	70.26	95.5

Table 7. Ablation study results in terms of different hyper-parameter values. M denotes the number of sub-sampling, K represents the number of clients in each sub-sampling.

Hyper-parameters	Metrics	
$M \times K$	Acc. (%)	AUC (%)
3×5	70.26	95.5
5×3	70.28	95.1
2×7	70.13	95.4
4×4	70.18	95.2

also helps to improve the accuracy score of 1.11% and the AUC score of 0.34% on SVHN, the accuracy score of 0.9% and the AUC score of -0.4% on CIFAR-100, as well as the accuracy score of 0.41% and the AUC score of 0.5% on ISIC 2018, as shown in the results of our method and "Basic+SSO".

Labeled Client Ratio. Note that FedIRM [21] focuses on ten clients with 2 labeled client and 8 unlabeled clients as the setting of FSSL, where more labeled data is involved. To evaluate the performance of RSCFed with increased labeled client ratio, we compare our work with previous arts under fixed number of clients. Following FedIRM [21], we empirically divides the whole training data into 10 clients consists of 2 labeled client and 8 unlabeled clients. Table. 3 lists extensive metrics of our method, FedIRM [21] and Fed-Consist [28]. Consistent improvements made by RSCFed can be observed in four metrics. Our improvement reaches 1.13% in accuracy, 0.3% in AUC, 0.86% in Precision, and notably 2.86% in Recall.

Unlabeled Client Ratio. To evaluate our performance under different unlabeled client ratios, we conduct an ablation study to compare different federated semi-supervised learning methods in terms of different numbers of clients, where the number of labeled client is set to 1 in all methods. Here, we consider the whole client number as 5, 10, 15, 25, 35, and 50, and Table 4 reports the results of our method and Fed-Consist [28]). As can be seen, our improvement of Accuracy and AUC scores over Fed-Consist are also enlarged when the number of unlabeled clients increases. The accuracy improvement is progressively increased from 1.51% to 3.98%, and the AUC improvement is from 0.5% to 3.5%, as the number of unlabeled clients grows from 4 to 49.

Communication Cost Limitations. Note that the compared two methods passed ten local client models in each synchronization round, while our method considers 15 local models since we utilize 3 sub-sampling operations, and 5 local clients is selected in each sub-sampling operation. Hence, the communication cost of our method is 1.5 times that of Fed-Consist [28]. We conduct an ablation study experiment to evaluate our method under different communication cost limitations. Specifically, we consider another three cases with 8 clients, 9 clients, and 10 clients, and thus the communication cost are 0.8, 0.9, and 1.0 times of the Baseline's communication cost. Table 6 lists the Accuracy and AUC scores of our method with different communication costs and Fed-Consist [28]. It shows that our network with 0.8 time of communication cost also outperforms the state-of-the-art method (Fed-Consist [28]) in terms of Accuracy and AUC scores.

Hyper-parameters. Note that our network has two major hyper-parameters, and they are the number (M) of sub-sampling operations and the number (K) of local clients used in each sub-sampling operation. Apparently, we empirically set $M = 3$, and $K = 5$. Here, we conduct an ablation study to study different choices of M and K, and report the Accuracy and AUC scores in Table 7. From the results, we can find that the Accuracy and AUC scores are only slightly different under different M and K values.

5. Conclusion

This work presents an important, practical but overlooked federated learning problem: federated semi-supervised learning with Non-IID local clients. Considering the uneven reliability among labeled and unlabeled clients, our key idea is that the consensus could be reached by performing multiple sub-sampling over clients. Instead of simply aggregating local models, we devise a sub-consensus model by randomly sub-sampling over clients and introduce a distance-reweighted model aggregation module to aggregate sub-sampled models in each synchronization round. Experimental results on three benchmark datasets show that our network consistently outperforms state-of-the-art methods, which proves the effectiveness of our method.

Acknowledgement

This work was supported by a research grant from HKUST Bridge Gap Fund (BGF.027.2021), a research grant from Shenzhen Municipal Central Government Guides Local Science and Technology Development Special Funded Projects (2021Szvup139), a research grant from the National Natural Science Foundation of China (Grant No. 61902275), and A*STAR AI3 HTPO Seed Fund (C211118012).

References

[1] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 3, 4

[2] Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning. In *International Conference on Learning Representations*, 2021. 3

[3] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 3

[4] Wenhui Cui, Yanlin Liu, Yuxing Li, Menghao Guo, Yiming Li, Xiuli Li, Tianle Wang, Xiangzhu Zeng, and Chuyang Ye. Semi-supervised brain lesion segmentation with an adapted mean teacher model. In *International Conference on Information Processing in Medical Imaging*, pages 554–565. Springer, 2019. 3

[5] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[7] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. Simple: Similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2021. 3

[8] Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *International Conference on Learning Representations*, 2021. 1, 3

[9] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019. 1

[10] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020. 1

[11] Yan Kang, Yang Liu, and Tianjian Chen. Fedmvt: Semi-supervised vertical federated learning with multiview training. *arXiv preprint arXiv:2008.10838*, 2020. 3

[12] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020. 1, 2

[13] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 2

[14] Rajesh Kumar, Abdullah Aman Khan, Jay Kumar, A Zakria, Noorbakhsh Amiri Golilarz, Simin Zhang, Yang Ting, Chengyu Zheng, and WenYong Wang. Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging. *IEEE Sensors Journal*, 2021. 1

[15] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021. 1

[16] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021. 1, 2, 5

[17] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018. 2

[18] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020. 3

[19] Haowen Lin, Jian Lou, Li Xiong, and Cyrus Shahabi. Semifed: Semi-supervised federated learning with consistency and pseudo-labeling. *arXiv preprint arXiv:2108.09412*, 2021. 1, 3

[20] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021. 2

[21] Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng. Federated semi-supervised medical image classification via inter-client relation matching. *arXiv preprint arXiv:2106.08600*, 2021. 1, 2, 3, 6, 7, 8

[22] Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. Fedvision: An online visual object detection platform powered by federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13172–13179, 2020. 1

[23] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 1, 2, 4, 6, 7

[24] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 3

[25] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017. 3, 6

[26] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020. 3, 5

[27] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*, 2020. 3

[28] Dong Yang, Ziyue Xu, Wenqi Li, Andriy Myronenko, Holger R Roth, Stephanie Harmon, Sheng Xu, Baris Turkbey, Evrim Turkbey, Xiaosong Wang, et al. Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from china, italy, japan. *Medical image analysis*, 70:101992, 2021. 1, 2, 3, 6, 7, 8

[29] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 2

[30] Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. Fedmix: Approximation of mixup under mean augmented federated learning. In *International Conference on Learning Representations*, 2021. 2

[31] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 3

[32] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pages 7252–7261. PMLR, 2019. 5

[33] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. Personalized federated learning with first order model optimization. In *International Conference on Learning Representations*, 2021. 3

[34] Zhengming Zhang, Zhewei Yao, Yaoqing Yang, Yujun Yan, Joseph E Gonzalez, and Michael W Mahoney. Benchmarking semi-supervised federated learning. *arXiv preprint arXiv:2008.11364*, 17, 2020. 3

[35] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 3