3. Duong, M.T., et al.: Artificial intelligence for precision education in radiology. Br. J. Radiol. **92**(1103), 20190389 (2019). https://doi.org/10.1259/BJR.20190389
4. Fetty, L., et al.: Latent space manipulation for high-resolution medical image synthesis via the StyleGAN. Z. Med. Phys. **30**(4), 305–314 (2020). https://doi.org/10.1016/j.zemedi.2020.05.001
5. Goodfellow, I., et al.: Generative Adversarial Nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems (NIPS). vol. 27 (2014). https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
6. Hyun Cho, J., Mall, U., Bala, K., Hariharan, B.: PiCIE: Unsupervised Semantic Segmentation using Invariance and Equivariance in Clustering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16789–16799 (2021). https://doi.org/10.1109/CVPR46437.2021.01652
7. Ji, X., Vedaldi, A., Henriques, J.: Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9864–9873 (2019). https://doi.org/10.1109/ICCV.2019.00996
8. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal Frequency Loss for Image Reconstruction and Synthesis. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 13899–13909 (2021). https://doi.org/10.1109/ICCV48922.2021.01366
9. Kaiser, L., Roy, A., Vaswani, A., Parmar, N., Bengio, S., Uszkoreit, J., Shazeer, N.: Fast Decoding in Sequence Models using Discrete Latent Variables. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, 80, pp. 2390–2399 (2018). https://proceedings.mlr.press/v80/kaiser18a.html
10. Li, H., Wei, D., Cao, S., Ma, K., Wang, L., Zheng, Y.: Superpixel-guided label softening for medical image segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12264, pp. 227–237. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59719-1_23
11. Ling, H., Kreis, K., Li, D., Kim, S.W., Torralba, A., Fidler, S.: EditGAN: High-Precision Semantic Image Editing. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (NeurIPS) vol. 34, pp. 16331–16345 (2021). https://proceedings.neurips.cc/paper/2021/file/880610aa9f9de9ea7c545169c716f477-Paper.pdf
12. Paszke, A.,et al.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems (NeurIPS), vol. 32, pp. 8024–8035 (2019). https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
13. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
14. Saboo, A., Gyawali, P.K., Shukla, A., Sharma, M., Jain, N., Wang, L.: Latent-optimization based disease-aware image editing for medical image augmentation. In: 32nd British Machine Vision Conference (BMVC). p. 181 (2021). https://www.bmvc2021-virtualconference.com/assets/papers/0840.pdf

15. Sasuga, S., et al.: Image Synthesis-Based Late Stage Cancer Augmentation and Semi-supervised Segmentation for MRI Rectal Cancer Staging. In: Nguyen, H.V., Huang, S.X., Xue, Y. (eds.) Data Augmentation, Labelling, and Imperfections. pp. 1–10. Springer Nature Switzerland, Cham (2022). https://link.springer.com/chapter/10.1007/978-3-031-17027-0_1
16. Schonfeld, E., Schiele, B., Khoreva, A.: A U-Net based discriminator for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8207–8216 (2020). https://doi.org/10.1109/CVPR42600.2020.00823
17. Thermos, S., Liu, X., O'Neil, A., Tsaftaris, S.A.: Controllable cardiac synthesis via disentangled anatomy arithmetic. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 160–170. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_15
18. Tiago, C., Snare, S.R., Šprem, J., McLeod, K.: A domain translation framework with an adversarial denoising diffusion model to generate synthetic datasets of echocardiography images. IEEE Access **11**, 17594–17602 (2023). https://doi.org/10.1109/ACCESS.2023.3246762
19. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068

Rethinking Semi-Supervised Federated Learning: How to Co-train Fully-Labeled and Fully-Unlabeled Client Imaging Data

Pramit Saha(✉), Divyanshu Mishra, and J. Alison Noble

Department of Engineering Science, University of Oxford, Oxford, UK
pramit.saha@eng.ox.ac.uk

Abstract. The most challenging, yet practical, setting of semi-supervised federated learning (SSFL) is where a few clients have fully labeled data whereas the other clients have fully unlabeled data. This is particularly common in healthcare settings where collaborating partners (typically hospitals) may have images but not annotations. The bottleneck in this setting is the joint training of labeled and unlabeled clients as the objective function for each client varies based on the availability of labels. This paper investigates an alternative way for effective training with labeled and unlabeled clients in a federated setting. We propose a novel learning scheme specifically designed for SSFL which we call Isolated Federated Learning (IsoFed) that circumvents the problem by avoiding simple averaging of supervised and semi-supervised models together. In particular, our training approach consists of two parts - (a) isolated aggregation of labeled and unlabeled client models, and (b) local self-supervised pretraining of isolated global models in all clients. We evaluate our model performance on medical image datasets of four different modalities publicly available within the biomedical image classification benchmark MedMNIST. We further vary the proportion of labeled clients and the degree of heterogeneity to demonstrate the effectiveness of the proposed method under varied experimental settings.

1 Introduction

Federated Learning (FL) [10–12,27] is a distributed learning approach that allows the collaborative training of machine learning models using data from decentralized sources while preserving data privacy. However, most current FL methods have limitations, including assuming fully annotated and homogeneous data distribution among local clients. In a practical scenario, like a multi-institutional healthcare collaboration, the participating clients (*i.e.*, medical institutions and hospitals) may not have the incentive or resources to annotate their data [16]. To address this, semi-supervised federated learning (SSFL)

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_39.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 414–424, 2023.
https://doi.org/10.1007/978-3-031-43895-0_39

[4,16,28] methods have been proposed to utilize unlabeled data and integrate semi-supervised learning algorithms [2,19–21,26] into federated settings.

Based on the availability of labeled data, the existing SSFL studies can be classified into two main scenarios: (a) labels-at-client, with each client having some labeled and some unlabeled data [9,15], (b) labels-at-server, with each client possessing only unlabeled data and the server possessing some labeled data [4,7,9,28]. We argue that a more realistic SSFL scenario which is highly challenging but rarely explored in the literature is where some clients have labeled data, and others have completely unlabeled data [14,16,24].

The classic federated averaging scheme aggregates weights of all labeled and unlabeled client models trained in parallel. The labeled clients typically use cross-entropy-based loss functions while the unlabeled clients primarily use consistency regularization loss [19] or pseudo-labeling-based [1,23] semi-supervised learning schemes. This results in high gradient diversity [28] between the supervised and unsupervised models particularly in heterogeneous client settings, as these are targeted to optimize separate objective functions. As a result, the aggregated global model is weak and unable to capture a strong representation of either group of clients. This, in turn, leads to the generation of noisy targets for unlabeled clients and hence the global model fails to converge. The situation is further aggravated under non-IID data distribution conditions where the labeled client class distribution varies greatly from that of unlabeled clients. This naturally poses the following important question: *"How can we effectively co-train supervised and unsupervised models under FL setting that aim to optimize separate objective functions at their respective heterogeneous labeled data or unlabeled data clients?"*

To address this question, we present a novel SSFL algorithm which we call IsoFed that effectively improves client training by isolating the model aggregation of labeled and unlabeled client groups while still leveraging one group of models to improve another. In summary, the primary contributions of this paper are:

1. We propose IsoFed, a novel SSFL framework, that realizes isolated aggregation of labeled and unlabeled client models in the server followed by federated self-supervised pretraining of the global model in each individual site.
2. This is the first work to reformulate model aggregation for fully labeled and fully unlabeled clients under SSFL settings. To the best of our knowledge, we are the first to isolate the aggregation of labeled and unlabeled client models while switching between the two client groups.
3. This work bridges the gap between Federated Learning and Transfer Learning (TL) [22] by combining the best of both worlds for learning across sites. First, we conduct federated model aggregation among the labeled or unlabeled client groups. Next, we leverage Transfer Learning to allow knowledge transfer between the two groups. Therefore, we avoid the issue of averaging the supervised and unsupervised models with high gradient diversity in the context of SSFL while also being unaffected by catastrophic forgetting encountered in multi-domain transfer learning.

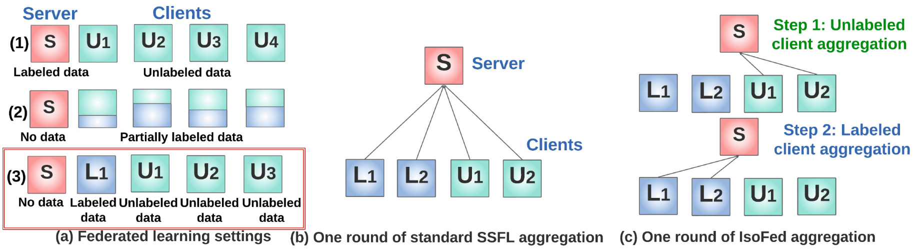

Fig. 1. Problem settings and aggregation schemes for semi-supervised federated learning. (a) Three plausible semi-supervised federated learning settings. We address the unique condition (3) with fully labeled and fully unlabeled clients. (b) One round of a standard FL aggregation scheme. (c) One round of our proposed two-step isolated aggregation scheme for labeled clients and unlabeled clients.

4. We, for the first time, extensively evaluate SSFL methods on multiple medical image benchmarks with a varying proportion of clients and degree of heterogeneity. Our results show that the proposed isolated aggregation followed by federated pretraining outperforms the state-of-the-art method, *viz.*, RSCFed [14] by **6.91**% in terms of accuracy and achieves near-supervised learning performance.

2 Methods

2.1 Problem Description

Assume a federated learning setting with m fully labeled clients denoted as $\{C_1, C_2, ..., C_m\}$ each possessing a labeled dataset $D^l = \{(X_i^l, y_i^l)\}_{i=1}^{N^l}$ and n fully unlabeled clients defined as $\{C_{m+1}, C_{m+2}, ..., C_{m+n}\}$ each possessing an unlabeled dataset $D^u = \{(X_i^u)\}_{i=1}^{N^u}$. Our objective is to learn a global model θ_{glob} via decentralized training.

2.2 Local Training

We adopt mean-teacher-based semi-supervised learning [12,14,20] to train each unlabeled client. At the beginning of each round, the global model W_{glob} is used to initialize the teacher model W_t. At the end of each communicating round, the student model W_s is returned to the server as the local model. Each batch of images undergoes two types of augmentations. The teacher model receives weakly augmented data whereas the student model receives strongly augmented data in each local iteration. In order to decrease entropy of model output, the temperature of predictions is further increased via sharpening operation [2,3,5,14] as $\hat{p}_{t,i} = Sharpen\ (p_t, \tau)_i = p_{t,i}^{\frac{1}{\tau}} / \sum_j p_{t,j}^{\frac{1}{\tau}}$ where $p_{t,i}$ and $\hat{p}_{t,i}$ denote each element in p_t before and after sharpening, respectively. τ denotes the temperature parameter. The student model is trained on the local data (D^u) via consistency regularization with the teacher model output. The consistency regularization loss is defined as $\mathcal{L}_{MSE} = \|\hat{p_t} - p_s\|_2^2$ where $\hat{p_t}$ and p_s are teacher and student predictions, respectively. $\|.\|_2^2$ denotes $L2$-norm. The student model weights are optimized via backpropagation whereas the teacher model weights are updated by exponential moving averaging (EMA) after each local iteration, as in Eq. 1:

$$W_{t+1} = \alpha W_s + (1 - \alpha) W_t \tag{1}$$

where α denotes momentum parameter. We optimize cross-entropy loss for local training on labeled clients defined as $\mathcal{L}_{CE} = -y_i \log\ p_i$, where y_i denotes labels.

2.3 Isolated Federated Aggregation

In this section, we explain the proposed isolated aggregation of labeled and unlabeled client models. Each communication round is composed of two consecutive substeps. First, the server initializes the global model W_{glob}^t and sends it to **unlabeled** clients (U_i). The global model is used to initialize the teacher model W_t in each client. At this stage, only the unlabeled clients perform local training on the global model by minimizing the consistency regularization loss. The updated semi-supervised models obtained after running the local epochs are then uploaded to the server. We adopt a dynamically weighted Federated Averaging scheme [14] to aggregate the model parameters of all unlabeled clients W_u at the server. For this, we first obtain the averaged model by performing Fed-Avg as in Eq. 2.

$$W_{avg} = \frac{\sum_{k=1}^{k=K} n_k W_k}{\sum_{k=1}^{k=K} n_k} \tag{2}$$

where K is the total number of clients. n_k is the number of samples in each client. The client models are then dynamically scaled using coefficients c_k designed as functions of the individual distances from the averaged model as denoted in Eq. 3. The global model (W_{glob}) is updated by re-aggregating the client weights scaled by new coefficients c_k. In Eq. 3, λ_c is a hyperparameter.

$$c_k = \frac{n_k \exp(-\lambda_c \frac{\|W_k - W_{avg}\|_2}{n_k})}{\sum_{k=1}^{k=K} n_k}, W_{glob} = \frac{\sum_{k=1}^{k=K} c_k W_k}{\sum_{k=1}^{k=K} c_k} \tag{3}$$

The updated global model parameters are then communicated to each **labeled** client which initializes its models using these weights and trains the local model via minimization of the standard cross-entropy loss. After a predefined number of local epochs, each labeled client uploads its local model to the server. The server then aggregates all the supervised models employing the aforementioned weighting scheme and the resultant global model W_{glob}^{t+1} is then sent to each unlabeled client at the beginning of the next round.

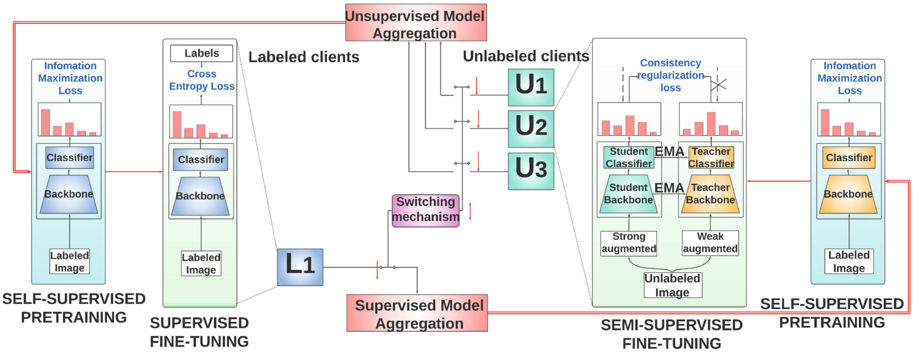

Fig. 2. Overview of our proposed methodology (IsoFed) with 1 labeled and 3 unlabeled clients. The unlabeled clients are trained using a mean-teacher-based SSL model. A switching mechanism swaps between labeled and unlabeled clients for isolated model aggregation in each round. After isolated model aggregation, an information maximization loss is used for client-adaptive pretraining to enhance the certainty and diversity of predictions of the global model for each client before actual local training.

2.4 Client-Adaptive Pretraining

Motivated by the recent success of continued pretraining in Natural Language Processing [6,8,17], we present a client-adaptive pretraining strategy as the second part of our proposed method. If we view the isolated FL from a transfer learning perspective, the global model received in one group of clients from the server can be regarded as an averaged model pretrained on the other group of clients. To improve client-specific model performance, we conduct a second phase of in-client federated pretraining on the global model before initializing it as a teacher model.

For self-supervised pretraining, we jointly learn the client-invariant features and client-specific classifier by optimizing an information-theoretic metric called information maximization (IM) loss denoted as $\mathcal{L}_{inf}$ in Eq. 4. It acts as an estimate of the expected misclassification error of the global model for each client. Optimizing the IM loss makes the global model output predictions that are individually certain but collectively diverse. With the help of a diversity preserving regularizer (first component in Eq. 4), IM avoids the trivial solution of entropy minimization where all unlabeled data collapses to the same one-hot encoding. The joint optimization is done by reducing the entropy of the output probability distribution of global model (p_i) in conjunction with maximizing the mutual information between the data distribution and the estimated output distribution yielded by the global model.

$$\mathcal{L}_{inf} = \mathbb{E}_{x \in D}\left[\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right)\log\left(\frac{1}{N}\sum_{i=1}^{N} p_i\right) - \frac{1}{N}\sum_{i=1}^{N} p_i \log p_i\right] \tag{4}$$

where N is the number of classes. x denotes any instance belonging to a dataset D. The entropy minimization leads to the least number of confused predictions whereas the regularizer avoids the degenerate solution where every data sample is assigned to the same class [13,18]. The pretrained model is then initialized as the teacher model to train the local student model in each round.

3 Experiments and Results

3.1 Datasets and FL Settings

To evaluate the performance and generalisability of the proposed method, we conduct experiments on four publicly available medical image benchmark datasets with different modalities [25], *viz.*, BloodMNIST (microscopic peripheral blood cell images), PathMNIST (colon pathology), PneumoniaMNIST (chest X-ray), and OrganAMNIST (abdominal CT - axial view). Each image resolution is 28 × 28 pixels and is normalized before feeding it to the network. BloodMNIST contains a total of 17,092 images and is organized into 8 classes. PathMNIST has 107,180 images and has 9 types of tissues. PneumoniaMNIST is a collection of 5,856 images and the task is binary classification (diseased vs normal). OrganAMNIST is comprised of 58,850 images and the task is multi-class classification of 11 body organs. We split each training dataset between 4 clients to mimic a practical collaborative setting in healthcare. To testify the versatility of the models, we study two challenging non-IID data partition strategies with 0.5 and 0.8-Dirichlet (γ). As a result, the number of samples per class and per client widely vary from each other. Additionally, we show the impact of varying the proportion of labeled clients ($75\%, 50\%, 25\%$) on model performance. See **Suppl. Sec 1** for more details.

3.2 Implementation and Training Details

For all datasets, we employ a simple CNN comprising of two 5 × 5 convolution layers, a 2 × 2 max-pooling layer, and two fully-connected layers as the feature extraction backbone followed by a two-layer MLP and a fully-connected layer as the classification network. Our model is implemented with PyTorch. We follow the settings prescribed for a training RSCFed to enable a fair comparison. See **Suppl. Sec 2** for more training details.

3.3 Results and Discussion

We use the standard metrics - accuracy, area under a ROC curve (AUC), Precision, and Recall to evaluate performance. We observe that the dynamically weighted version of Fed-Avg (discussed in Sect. 2.3) outperforms standard Fed-Avg and hence use it as a baseline in this paper instead of vanilla Fed-Avg. In order to fairly evaluate IsoFed, we compare with the following state-of-the-art SSFL benchmarks: (a) MT+wFed-Avg: a combination of Mean Teacher and dynamically weighted Fed-Avg, (b) RSCFed: Random sampling consensus-based FL [14]. Since RSCFed has already been shown to significantly outperform FedIRM [16] and Fed-Consist [24] on multiple datasets, we exclude those

Table 1. Comparison with baselines on BloodMNIST and PathMNIST. wFedAvg refers to dynamically weighted Federated averaging. UB implies Upper Bound. MT refers to Mean teacher-based SSL. Acc. and Prec. denote Accuracy and Precision. L and U denote the number of labeled and unlabeled clients respectively.

Labeling	**Method**	**Client**		**Metrics** (%)				**Metrics** (%)			
		L	U	**Acc.**	**AUC**	**Prec.**	**Recall**	**Acc.**	**AUC**	**Prec.**	**Recall**
				$\gamma = 0.8$ (less non-IID)				$\gamma = 0.5$ (more non-IID)			
Dataset 1 : BloodMNIST, Task : 8-class classification											
Fully supervised	wFed-Avg (UB)	4	0	79.57	96.61	77.65	75.70	79.45	96.80	78.28	73.31
Semi supervised	MT+wFed-Avg	3	1	77.32	96.70	74.16	73.79	70.89	95.11	73.46	65.06
	RSCFed	3	1	76.94	95.54	75.11	71.18	75.18	94.99	76.55	68.96
	IsoFed	3	1	**79.43**	**97.32**	**76.70**	**76.67**	**76.10**	**95.88**	**77.13**	**72.29**
	MT+wFed-Avg	2	2	75.88	96.56	72.85	71.94	58.29	88.35	57.85	60.46
	RSCFed	2	2	75.97	95.30	73.58	72.77	61.18	**91.50**	54.85	60.79
	IsoFed	2	2	**80.47**	**97.25**	**77.11**	**78.11**	**64.05**	90.01	**60.26**	**64.03**
	MT+wFed-Avg	1	3	75.24	95.13	72.43	70.37	52.56	89.39	57.89	55.81
	RSCFed	1	3	71.88	93.96	70.47	67.75	19.35	64.31	07.05	23.62
	IsoFed	1	3	**79.23**	**96.43**	**76.68**	**77.00**	**63.70**	**90.58**	**70.22**	**63.81**
Dataset 2 : PathMNIST, Task : 9-class classification											
Fully supervised	wFed-Avg (UB)	4	0	70.45	94.92	72.13	69.84	68.97	94.93	68.05	67.58
Semi supervised	MT+wFed-Avg	3	1	60.97	93.60	68.14	62.00	57.92	92.93	**67.20**	59.98
	RSCFed	3	1	61.55	93.71	61.00	58.95	58.33	93.59	60.68	58.73
	IsoFed	3	1	**63.10**	**94.73**	**69.25**	**64.62**	**60.23**	**93.98**	52.80	**61.66**
	MT+wFed-Avg	2	2	67.10	**95.17**	**66.41**	**66.4**0	61.28	91.26	61.50	57.56
	RSCFed	2	2	64.18	93.17	60.79	58.89	58.83	90.35	58.88	55.02
	IsoFed	2	2	**70.32**	94.74	65.96	64.86	**64.00**	**93.46**	**63.88**	**61.22**
	MT+wFed-Avg	1	3	59.57	90.66	63.14	58.93	56.31	89.92	60.42	53.92
	RSCFed	1	3	64.75	**94.09**	**66.89**	**63.66**	57.42	89.43	54.96	53.53
	IsoFed	1	3	**66.48**	92.24	63.71	62.06	**64.02**	**93.99**	**66.12**	**62.39**

methods from our comparative study due to space constraints. We consider fully-supervised FL as an upper bound and report the results for both the non-IID settings on each dataset. Tables 1-2 show that overall, IsoFed outperforms RSCFed by 6.91%, 4.15%, 7.28%, and 6.71% in terms of average accuracy, AUC, Precision, and Recall respectively.

Table 1 shows our method and our baselines on 8-class classification with BloodMNIST. L and U denote the number of labeled and unlabeled clients respectively. The average accuracy for fully-supervised FL is 79.51%. Among the baselines, MT+wFed-Avg has a higher overall accuracy score of 68.36% while RSCFed has an accuracy score of 63.41%. Particularly, we find RSCFed collapses under the most extreme case of $\gamma = 0.5$ and U=3. IsoFed improves the accuracy score to 73.83% and is stable for all evaluated conditions. Table 1 further reports performance on 9-class classification with PathMNIST. The fully-supervised FL achieves an overall accuracy of 69.71%. The baselines have very similar accuracy scores of 60.53% and 60.84% respectively. IsoFed improves it to 64.69%.

Table 2 shows binary classification results on PneumoniaMNIST. The fully-supervised FL has an overall accuracy of 87.18%. MT+wFed-Avg and RSCFed

Table 2. Performance comparison of IsoFed with baselines on PneumoniaMNIST and OrganAMNIST (with ablation study). PT refers to the federated pretraining step.

Labeling	Method	Client		Metrics (%)				Metrics (%)			
		L	U	Acc.	AUC	Prec.	Recall	Acc.	AUC	Prec.	Recall
				$\gamma = 0.8$ (less non-IID)				$\gamma = 0.5$ (more non-IID)			
Dataset 3 : PneumoniaMNIST, Task : Binary classification											
Fully supervised	wFed-Avg (UB)	4	0	87.34	95.32	86.71	89.02	87.02	95.64	86.45	88.76
Semi supervised	MT+wFed-Avg	3	1	86.54	95.20	85.94	88.21	86.86	94.85	85.92	87.86
	RSCFed	3	1	86.58	95.63	**89.02**	88.68	86.70	94.50	85.75	87.65
	IsoFed	3	1	**87.10**	**95.04**	86.45	**89.00**	**89.26**	**95.80**	**88.26**	**89.44**
	MT+wFed-Avg	2	2	83.65	89.74	82.45	82.99	82.21	96.17	83.20	85.26
	RSCFed	2	2	78.37	87.36	77.31	78.76	**84.46**	**95.58**	**84.58**	**86.88**
	IsoFed	2	2	**84.70**	**90.75**	**83.56**	**84.64**	82.68	95.15	83.34	85.41
	MT+wFed-Avg	1	3	81.41	89.84	82.05	77.69	**79.97**	**94.45**	**81.28**	**83.12**
	RSCFed	1	3	78.85	86.66	77.56	76.84	62.50	50.00	31.25	50.00
	IsoFed	1	3	**85.00**	**91.68**	**83.98**	**83.95**	77.12	93.65	80.47	81.40
Dataset 4 : OrganAMNIST, Task: 11-class classification											
Fully supervised	wFed-Avg (UB)	4	0	69.72	94.41	67.44	69.60	69.50	94.63	68.12	69.60
Semi supervised	MT+wFed-Avg	3	1	68.36	93.72	68.02	69.38	66.49	93.69	67.51	68.25
	RSCFed	3	1	68.14	94.26	67.44	69.53	67.08	93.82	68.82	68.36
	IsoFed w/o PT	3	1	68.98	94.32	**68.83**	69.88	67.45	93.98	67.85	69.35
	IsoFed	3	1	**69.47**	**95.05**	68.04	**70.85**	**68.65**	**94.88**	**68.64**	**69.77**
	MT+wFed-Avg	2	2	66.28	92.77	66.12	67.63	61.71	**92.55**	**65.79**	62.66
	RSCFed	2	2	66.68	92.42	66.90	66.56	62.51	91.89	64.09	63.35
	IsoFed w/o PT	2	2	68.67	93.25	67.65	68.50	**64.37**	92.11	65.70	65.17
	IsoFed	2	2	**68.95**	**93.95**	**68.32**	**69.83**	64.08	92.45	64.56	**65.47**
	MT+wFed-Avg	1	3	57.75	90.95	61.50	55.68	50.84	87.65	60.07	48.51
	RSCFed	1	3	58.50	90.86	63.48	55.76	54.90	89.58	50.53	53.41
	IsoFed w/o PT	1	3	62.03	91.36	64.50	61.44	56.40	89.79	61.61	55.72
	IsoFed	1	3	**62.77**	**91.48**	**64.52**	**61.79**	**61.90**	**91.55**	**62.39**	**60.21**

achieve average accuracy scores of 83.44% and 79.57%. IsoFed has the best accuracy of 85.45%. Furthermore, the results of 11-class anatomy classification task on OrganAMNIST are also reported in Table 2. The upper bound accuracy is 69.61% and the baseline accuracies are 61.91% and 62.97% respectively. IsoFed achieves an overall accuracy score of 65.97%. In general, the performance of all methods decreases with γ changing from 0.8 to 0.5. It is expected as the clients become more label-skewed due to higher non-IID data partition. However, our approach is least affected by this which is reflected in its accuracy decrease by 2.19% as opposed to 4.45% and 2.94% incurred by baselines. As foreseen, performance also deteriorates with decrease in the number of labeled clients. For L:U = 3:1, 2:2, 1:3, the baseline accuracies degrade by 2.16%, 5.61%, 15.31% and 2%, 5.01%, 12.91% w.r.t. fully supervised FL setting. However, for IsoFed, the decrease in accuracy is only 0.55%, 3.09%, and 7.28%, respectively. This proves the near-supervised learning performance of the proposed training method.

The superior performance of IsoFed over the baselines and closer performance to the upper bound demonstrates better learning and generalization. This is achieved by the isolated aggregation strategy and federated pretraining on all datasets.

3.4 Ablation Study

Owing to space constraints, we show ablation experiments only on OrganAMNIST, which provides the most challenging classification task, to evaluate the impact of IsoFed components. (More results in **Suppl. Sec 2**). Table 2 demonstrates that client-adaptive pretraining improves model accuracy by 5.50% for the most extreme condition of $\gamma = 0.5$ and L:U = 1:3.

4 Conclusion

We have introduced a novel SSFL framework called IsoFed, an isolated federated learning technique, to address joint training of labeled and unlabeled clients in the context of decentralized semi-supervised learning. It opens a new research direction in learning across domains by unifying two dominant approaches - Federated Learning (among labeled or unlabeled clients) and Transfer Learning (between labeled and unlabeled clients). Our results challenge the conventional strategy of co-training fully labeled and fully unlabeled clients in SSFL. Experimental results on 4 different medical imaging datasets with varied proportion of labeled clients $(25, 50, 75\%)$ and varied non-IID distribution (0.5 & 0.8-Dirichlet) show that IsoFed achieves a considerable boost compared to current state-of-the-art SSFL method. IsoFed can be easily incorporated into other federated learning-based aggregation schemes as well as used in conjunction with any other semi-supervised learning framework in federated learning setting.

Acknowledgement. This work was supported in part by the UK EPSRC (Engineering and Physical Research Council) Programme Grant EP/T028572/1 (VisualAI), a UK EPSRC Doctoral Training Partnership award, and the InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE) Project 2.1 (Cardiovascular risks in early life and fetal echocardiography).

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: MixMatch: a holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
4. Diao, E., Ding, J., Tarokh, V.: Semifl: communication efficient semi-supervised federated learning with unlabeled clients. arXiv preprint arXiv:2106.01432 3 (2021)

5. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
6. Gururangan, S., et al.: Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964 (2020)
7. He, C., Yang, Z., Mushtaq, E., Lee, S., Soltanolkotabi, M., Avestimehr, S.: SSFL: tackling label deficiency in federated learning via personalized self-supervision. arXiv preprint arXiv:2110.02470 (2021)
8. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018)
9. Jeong, W., Yoon, J., Yang, E., Hwang, S.J.: Federated semi-supervised learning with inter-client consistency & disjoint learning. arXiv preprint arXiv:2006.12097 (2020)
10. Ji, S., Saravirta, T., Pan, S., Long, G., Walid, A.: Emerging trends in federated learning: From model fusion to federated x learning. arXiv preprint arXiv:2102.12920 (2021)
11. Kairouz, P., et al.: Advances and open problems in federated learning. Found. Trends® Mach. Learn. **14**(1–2), 1–210 (2021)
12. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. IEEE Sig. Process. Mag. **37**(3), 50–60 (2020)
13. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International Conference on Machine Learning, pp. 6028–6039. PMLR (2020)
14. Liang, X., Lin, Y., Fu, H., Zhu, L., Li, X.: RSCFed: random sampling consensus federated semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10154–10163 (2022)
15. Lin, H., Lou, J., Xiong, L., Shahabi, C.: Semifed: Semi-supervised federated learning with consistency and pseudo-labeling. arXiv preprint arXiv:2108.09412 (2021)
16. Liu, Q., Yang, H., Dou, Q., Heng, P.-A.: Federated semi-supervised medical image classification via inter-client relation matching. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 325–335. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_31
17. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
18. Shi, Y., Sha, F.: Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. arXiv preprint arXiv:1206.6438 (2012)
19. Sohn, K., et al.: FixMatch: simplifying semi-supervised learning with consistency and confidence. Adv. Neural Inf. Process. Syst. **33**, 596–608 (2020)
20. Tarvainen, A., Valpola, H.: Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
21. Van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Mach. Learn. **109**(2), 373–440 (2020)
22. Weiss, K., Khoshgoftaar, T.M., Wang, D.D.: A survey of transfer learning. J. Big Data **3**(1), 1–40 (2016). https://doi.org/10.1186/s40537-016-0043-6
23. Yafen, L., Yifeng, Z., Lingyi, J., Guohe, L., Wenjie, Z.: Survey on pseudo-labeling methods in deep semi-supervised learning. J. Front. Comput. Sci. Technol. **16**(6), 1279 (2022)
24. Yang, D., et al.: Federated semi-supervised learning for covid region segmentation in chest CT using multi-national data from china, Italy, japan. Med. Image Anal. **70**, 101992 (2021)

25. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Sci. Data **10**(1), 41 (2023)
26. Zhang, B., et al.: FlexMatch: boosting semi-supervised learning with curriculum pseudo labeling. Adv. Neural Inf. Process. Syst. **34**, 18408–18419 (2021)
27. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y.: A survey on federated learning. Knowl.-Based Syst. **216**, 106775 (2021)
28. Zhang, Z., et al.: Improving semi-supervised federated learning by reducing the gradient diversity of models. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 1214–1225. IEEE (2021)