FedCD: Federated Semi-Supervised Learning with Class Awareness Balance via Dual Teachers

Yuzhi Liu[1], **Huisi Wu**[1*], **Jing Qin**[2]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] Centre for Smart Health, The Hong Kong Polytechnic University
hswu@szu.edu.cn

Abstract

Recent advancements in deep learning have greatly improved the efficiency of auxiliary medical diagnostics. However, concerns over patient privacy and data annotation costs restrict the viability of centralized training models. In response, federated semi-supervised learning has garnered substantial attention from medical institutions. However, it faces challenges arising from knowledge discrepancies among local clients and class imbalance in non-independent and identically distributed data. Existing methods like class balance adaptation for addressing class imbalance often overlook low-confidence yet valuable rare samples in unlabeled data and may compromise client privacy. To address these issues, we propose a novel framework with class awareness balance and dual teacher distillation called FedCD. FedCD introduces a global-local framework to balance and purify global and local knowledge. Additionally, we introduce a novel class awareness balance module to effectively explore potential rare classes and encourage balanced learning in unlabeled clients. Importantly, our approach prioritizes privacy protection by only exchanging network parameters during communication. Experimental results on two medical datasets under various settings demonstrate the effectiveness of FedCD. The code is available at https://github.com/YunzZ-Liu/FedCD.

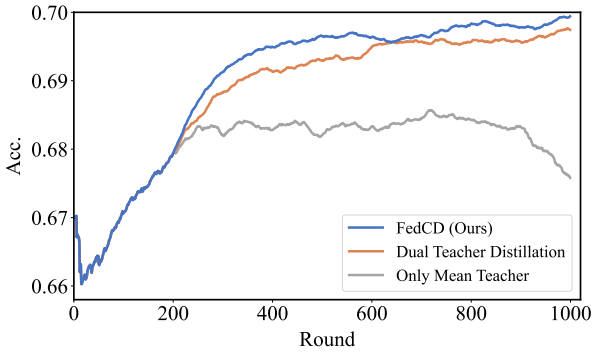
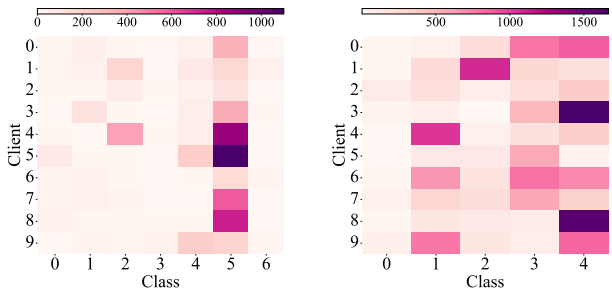


Figure 1: Comparisons of the test accuracy curves showed that our proposed FedCD method with dual teacher distillation outperformed the variant with only mean teacher distillation. The performance of the model relying solely on the mean teacher declined due to the inherent limitations of local knowledge. However, after incorporating the dual teacher distillation and class awareness balance modules, the issue of localized knowledge limitation was substantially mitigated, resulting in remarkable performance improvements.

Introduction

Federated Learning (FL) is a decentralized machine learning framework, allowing multiple entities to collectively refine a model while upholding data confidentiality by not divulging raw data (Li et al. 2020a; Mammen 2021; Huang et al. 2023). Particularly within medical image diagnostics, FL emerges as a transformative force in healthcare (Chen et al. 2022b; Zhu and Luo 2022). By harnessing the combined expertise of disparate healthcare providers while meticulously safeguarding data privacy, FL provides a pathway for more precise diagnoses. It concurrently fulfills the necessity for collaborative learning in healthcare, while ensuring stringent patient confidentiality and data security (Dong and Voiculescu 2021; Antunes et al. 2022). Nonetheless, annotating medical image data presents formidable challenges, primarily driven by its considerable expenses, time-intensive demands, and the potential risk to privacy and security when centralized (Liu et al. 2020; Huynh, Nibali, and He 2022). In light of these difficulties, the advent of federated semi-supervised learning (FSSL) has offered a promising solution. The fundamental objective of FSSL is to facilitate collaborative model training within a distributed setting. This is achieved by harnessing a finite pool of labeled data in conjunction with a more copious supply of unlabeled data (Kassem et al. 2022; Lin et al. 2021; Long et al. 2020). This innovative paradigm synergistically combines the principles of semi-supervised learning and federated learning. It thereby enables numerous clients, each in possession of a restricted private cache of labeled data, to contribute to the incremental refinement of a shared model while upholding stringent data privacy safeguards. Existing approaches to FSSL can be categorized into three primary types based on the distribution of labeled data: centralized on servers (Jiang et al. 2022; Wang et al. 2023), scattered across all clients (Yang et al. 2022) or siloed on individual clients (Liang et al. 2022). In this paper, we focus on the third type, utilizing a small number of labeled clients and a large number of unla-

*Corresponding author. Email: hswu@szu.edu.cn.



(a) HAM10000 Dataset (b) RSNA ICH Dataset

Figure 2: The heatmap of sample distribution across clients. Each rectangle represents the number of data samples for a specific class in each client. The non-independent and identically distributed setting exacerbates the skewed class distribution within the imbalanced dataset

.

beled clients to train a high-performing server model.

Although extensive research has been conducted on FSSL, its performance is still limited by two major challenges: local knowledge shift and class imbalance within and among clients, as depicted in Figure 1 and Figure 2. RSCFed (Liang et al. 2022) proposed the use of sub-consensus models to eliminate noise generated during client aggregation. However, the mean teacher model in unlabeled clients may still lead to a knowledge shift. CBAFed (Li, Li, and Wang 2023) introduced the utilization of global prior knowledge to fix pseudo-labels and identify tail categories but might overlook many low-confidence yet informative rare class samples.

Hence, we propose a FSSL framework with **c**lass awareness balance and **d**ual teacher distillation named FedCD. In order to balance global and local knowledge, we introduce a dual-teacher framework to guide client learning. However, both global and local teachers may impart erroneous knowledge. Therefore, we further purify the teacher knowledge by regulating the quality of the output knowledge, which can reduce overconfidence in potentially incorrect labels. Additionally, we propose a novel class awareness balance module aimed at uncovering rare samples belonging to underrepresented classes hidden within the unlabeled clients. Specifically, we first identify proficiently learned classes for which the model is highly confident. Subsequently, we discern unreliable instances as potential rare-class samples, characterized by cases where the confident classes appear with a lower rank in the probability distribution for unlabeled samples. Ultimately, we recalibrate the loss function to allocate greater significance to these rare class samples during training. By correcting the imbalance, we enable more balanced federated learning. Overall, our main contributions can be summarized as follows:

- We propose a novel federated learning approach called FedCD, designed to tackle the issues of knowledge drift and class imbalance in local clients. Unlike existing methods, our approach fully utilizes the data from unlabeled clients while ensuring privacy preservation between clients.
- In FedCD, dual teacher distillation provides more reliable pseudo labeling foundations, while class awareness balance excavates rare class samples to increase model attention on class imbalance. This allows local clients to attain more diverse and balanced knowledge during federated learning.
- Experiment on two medical datasets: HAM10000 and RSNA ICH. Our proposed method achieved significant improvements over the state-of-the-art FSSL methods in various experimental settings.

Related work

Federated Learning

Federated learning (FL) stands out as a robust approach to preserving data privacy through its decentralized framework (Dou et al. 2021; Kaissis et al. 2020; Rieke et al. 2020). However, this decentralization leads to data heterogeneity among clients. FedAvg (McMahan et al. 2017) firstly introduces the concept of averaging local models to derive a global model, serving as a baseline and the seminal contribution in this field. Subsequently, an increasing number of scholars have proposed solutions to address the data heterogeneity in FL. These solutions can be categorized into two main approaches: improve client model aggregation (Tan et al. 2022; Chen and Chao 2020) and local training (Li et al. 2020b; Liu et al. 2021a; Andreux et al. 2020; Li, He, and Song 2021).

Semi-supervised Leaning

Semi-supervised learning aims to train an optimal model by leveraging a small quantity of labeled data in conjunction with a substantial amount of unlabeled data. Common paradigms include consistency regularization (Bachman, Alsharif, and Precup 2014; Xu et al. 2022), entropy minimization (Grandvalet and Bengio 2004; Chen et al. 2021); and self-training (Du et al. 2022; Chen et al. 2022a). Additionally, data augmentation (Kim et al. 2022; Olsson et al. 2021) can enhance the generalization capacity of semi-supervised learning. In recent years, various high-performing approaches for semi-supervised learning have emerged (Dou et al. 2021; Kaissis et al. 2020; Rieke et al. 2020). However, these approaches are not directly applicable to FSSL, as each client may exclusively possess either labeled or unlabeled data, which precludes them from adequately capturing the intrinsic structure and properties of holistic data distribution.

Federated Semi-supervised Leaning

The objective of FSSL is to optimize a model under the non-independent and identically distributed (Non-IID) data sets. Unlike traditional semi-supervised learning, FSSL faces the challenges of distributed data and privacy protection (Zhang et al. 2021a). Existing FSSL methods can be categorized into

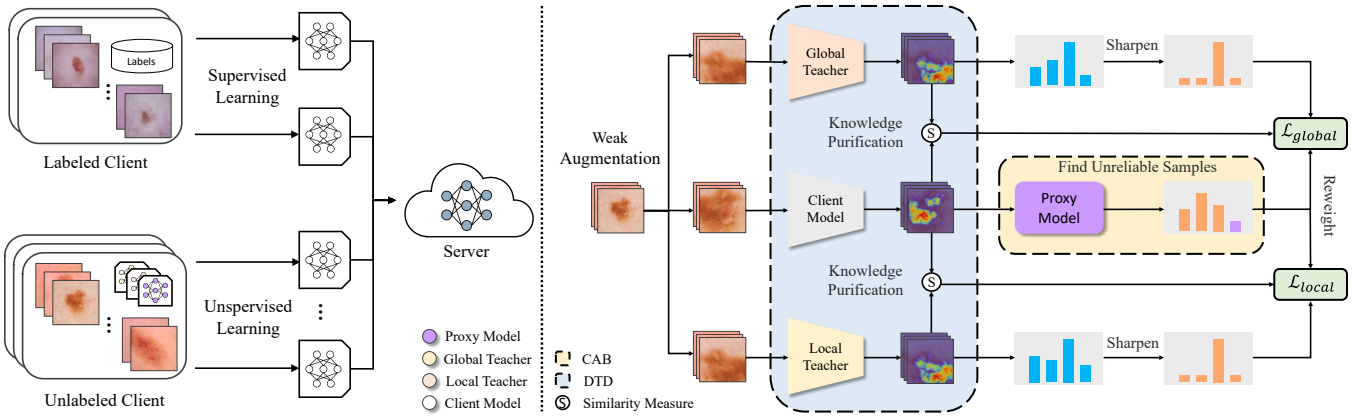


Figure 3: An overview of our proposed FedCD framework. The left side shows our overall architecture, where we introduce a proxy model and global-local teacher framework to assist unlabeled clients. The right side depicts the training process within unlabeled clients. We propose the dual teacher distillation and class awareness balance module for effective balanced learning.

three types: (a) where a small amount of labeled data is available on the server (Jeong et al. 2020; Long et al. 2021; Zhang et al. 2021b), (b) where each client has a small amount of labeled data (Che et al. 2021; Shi, Chen, and Zhang 2022; Itahara et al. 2021), and (c) where only a few clients possess labeled data (Guo et al. 2022). RSCFed (Liang et al. 2022) proposed random client sampling for consensus modeling over direct aggregation, aiming for greater model robustness. However, reliance on mean-teacher models in unlabeled clients persists as a limitation, owing to constrained local knowledge which hampers global performance. CBAFed (Li, Li, and Wang 2023) proposes utilizing fixed pseudo labels and exploring tail classes. However, this approach risks overlooking many low-confidence but information-rich rare class samples. Moreover, exchanging the empirical distribution of local data may raise privacy concerns.

Methodology

In this section, we introduce our innovative federated learning framework, depicted in Figure 3. For labeled clients, we utilize an entropy loss function for supervised learning. For unlabeled clients, we propose a dual teacher mechanism with global and local teachers to mitigate challenges from local knowledge bias and forgetting global knowledge. Moreover, we employ local proxy models to identify confident classes and extract potential rare class samples from among unreliable instances. The details of each component are elaborated in the following sections.

Problem Settings

In federated semi-supervised learning, we have m labeled clients C^l and n unlabeled clients C^u. Besides, there are N_l samples $\mathcal{S}^l = \left\{\left(x_i^l, y_i^l\right)\right\}_{i=1}^{N_l}$ at labeled clients and N_u samples $\mathcal{S}^u = \{(x_i^u)\}_{i=1}^{N_u}$ at unlabeled clients. The aim of FSSL is to leverage both labeled and unlabeled client data to learn a server federated model θ_s, which could be represented as follows:

$$\begin{aligned}\arg\min_{\theta_s} \mathcal{L}\left(\theta_s\right) \quad &= \sum_{i=1}^{m} \frac{\left|\mathcal{S}_i^l\right|}{|\mathcal{S}|} \mathcal{L}_{ce}\left(\theta_c\right) + \\ &\sum_{i=1}^{n} \frac{\left|\mathcal{S}_i^u\right|}{|\mathcal{S}|} \mathcal{L}_u\left(\theta_c\right)\end{aligned} \tag{1}$$

where $\mathcal{L}_{ce}$ is the cross entropy loss for supervised learning, $\mathcal{L}_u$ is the loss for unsupervised learning. θ_c denotes the network parameters for the local client. We utilize FedAvg (McMahan et al. 2017) to update the server model, where $|\mathcal{S}|$ represents the total sum of the data across all clients, *i.e.*, $|\mathcal{S}| = \sum_{i=1}^{m} \left|\mathcal{S}_i^l\right| + \sum_{i=1}^{n} \left|\mathcal{S}_i^u\right|$.

Dual Teacher Distillation

In unlabeled clients, relying on mean teacher-based consistency regularization frameworks may lead to a client bias towards local knowledge while neglecting global knowledge, which is undesirable in the context of FSSL. To address this issue, we propose a dual teacher distillation module that aims to refine knowledge and mitigate client bias by utilizing both global and local teachers.

Specifically, each client maintains a global teacher model θ_g in addition to its local teacher θ_l. The global teacher provides a unified representation across all clients, while the local teacher adapts to the unique data distribution of each client. Their role is to transfer distilled knowledge to guide the learning of the client model θ_c, which acts as a student.

In this context, we establish pairs of data augmentation samples to improve representation learning by enforcing consistency between differently perturbed versions of the same input. Upon traversing the network, we obtain the feature projections F_c, F_g, and F_l along with the softmax predictions P_c, P_g, and P_l from the student and teacher models respectively. Subsequently, we adopt a sharpening (Berthelot et al. 2019) operation on the prediction of teachers to distill global and local knowledge, *i.e.*, $\hat{P} = P_i^{\frac{1}{\tau}} / \sum_j P_j^{\frac{1}{\tau}}$, where τ is the temperature parameter. Finally, we use the mean-square-error loss to ensure alignment in the represen-

Algorithm 1: The pipeline of unlabeled client

Input: θ_s^t: the server model of $t-1^{th}$ round
Output: θ_c^{t+1}: the unlabeled client model of t^{th} round
Unlabeled Client(θ_s^t):

for each unlabeled client **do**
  for each local epoch **do**
    $w \leftarrow$ **ProxyModel** (θ_s^t)
    $\mathcal{L}_{local} \leftarrow$ **ComputeLoss** by Eq.7.
    $\mathcal{L}_{global} \leftarrow$ **ComputeLoss** by Eq.8.
    $\mathcal{L}_u \leftarrow$ **ReWeightLoss**$(w, \mathcal{L}_{local}, \mathcal{L}_{global})$ Eq.12.
    Update θ_c^t using $\mathcal{L}_u$
  end for
end for
return θ_c^{t+1}

Proxy Model(θ_s^t):

Find confident classes k by Eq. 10
Find unreliable samples D_u by Eq.11
Compute imbalance weight factor w by Eq.13
return w

tations between the teacher and student models:

$$\mathcal{L}_{mse-local} = \left\| \hat{P}_l - P_c \right\| \tag{2}$$

$$\mathcal{L}_{mse-global} = \left\| \hat{P}_g - P_c \right\| \tag{3}$$

where $\hat{P}_g$ and $\hat{P}_l$ come from sharpening of their predicted. Therefore, the loss function for the unlabeled clients can be expressed as follows:

$$\mathcal{L}_u = \lambda_1 \mathcal{L}_{mse-global} + \lambda_2 \mathcal{L}_{mse-local} \tag{4}$$

λ_1 and λ_2 are the hyperparameters that balance the global and local losses. In each local iteration, the local teacher retains local knowledge by receiving θ_c through the exponential moving average. Meanwhile, the global teacher receives θ_s to propagate global knowledge.

While the direction from both global and local teachers assists in alleviating knowledge shifts among local clients, incorrect teacher knowledge can worsen this condition. To address this challenge, our emphasis is on enhancing the accuracy of knowledge generated by teachers. This is evaluated through the calculation of variance between predictions of the student and teacher models, which can be obtained from the KL divergence:

$$V_{local} = \mathrm{KL}\left(F_c \,||\, F_l\right) \tag{5}$$

$$V_{global} = \mathrm{KL}\left(F_c \,||\, F_g\right) \tag{6}$$

A high computed variance suggests that the knowledge distilled from the global or local teacher may be inaccurate. Consequently, we adjust equations 2 and 3 to refine the knowledge transferred by the teachers:

$$\mathcal{L}_{local} = \mathrm{e}^{-V_{local}} * \mathcal{L}_{mse-local} \tag{7}$$

$$\mathcal{L}_{global} = \mathrm{e}^{-V_{global}} * \mathcal{L}_{mse-global} \tag{8}$$

By explicitly modeling the alignment between student and teacher outputs, we refine the knowledge distillation process

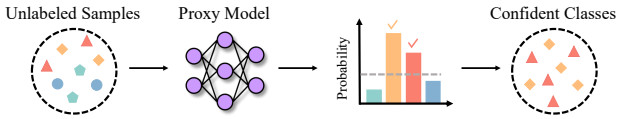


Figure 4: The pipeline of exploiting confident classes. We leverage global information to mine the confidence classes of each unlabeled client.

to filter out and minimize the impact of unreliable pseudo-labels. In summary, the loss function for unlabeled clients can be reformulated as:

$$\mathcal{L}_u = \lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{local} \tag{9}$$

Class Awareness Balance

The concurrent existence of intra-client and inter-client class imbalance significantly undermines the efficacy of federated learning. Furthermore, accurately identifying rare classes is a daunting task for unlabeled clients. In response, we introduce an innovative class awareness balance module that extracts implicit insights from both labeled and unlabeled clients. Specifically, we achieve this by exploiting confident classes, identifying unreliable samples, and recalibrating loss function.

Exploit Confident Classes. Typically, the cumulative predicted probabilities of confident classes surpass those of non-confident classes due to the model's better understanding of the former (Lin et al. 2022). Nonetheless, predicted pseudo-labels often lack reliability, and confident classes in unlabeled clients might not necessarily be trustworthy. To mitigate this issue, we can utilize labeled client data for pre-training, enabling the central server model to acquire the capacity to effectively differentiate between all classes. During the subsequent collaborative training phase, labeled clients employ their labeled data to sustain the global model's competence in identifying confident classes. Meanwhile, unlabeled clients employ local proxy servers θ_p to investigate local confident classes through the reception of model parameters θ_s, as illustrated in Figure 4. The cumulative predicted probabilities are calculated as outlined below:

$$s_k = \sum_{i}^{n_u} \mathrm{G}_k\left(\theta_p\left(x_i\right)\right) \tag{10}$$

where $\mathrm{G}_k\left(\cdot\right)$ represents the softmax output for the k^{th} class corresponding to sample x. In order to make cumulative probabilities more interpretable, we normalize s_k to [0, 1], *i.e.*, $s_k = \frac{s_k - \min(s)}{\max(s) - \min(s)}$, where $s = [s_0, s_1, \ldots, s_{k-1}]$ represents the cumulative probability vector for k classes. We identify class k as a confident class if the value of s_k exceeds the threshold β, denoted as $s_k > \beta$.

Identify Unreliable Samples. In unlabeled clients, accurately assigning an exact class label to a sample through pseudo-labeling presents a formidable challenge. However, distinguishing samples by identifying the class they do not pertain to is comparatively less complex. When confident

classes manifest themselves toward the lower spectrum of the softmax probability distribution, it implies that the associated samples are likely to be unreliable. In other words, these samples possess a substantial probability of not belonging to the confident classes but rather align with the rare categories. Consequently, the set of unreliable samples D_u can be expressed as follows:

$$D_u = \{x \mid t_l < T(k) < t_h\} \tag{11}$$

where $T(\cdot)$ is the order operation $\mathrm{argsort}(P_l)$ for softmax output of sample.t_l and t_h are the low and high rank threshold, respectively. k is the confident class where $s_k > \beta$.

Recalibrate Loss Function. To achieve a more balanced learning capability for the model, we introduced a weighting factor w in front of the loss function. Unreliable instances get a higher weight while reliable samples get a lower weight. The loss function of the unlabeled client in equation 9 can be updated as:

$$\mathcal{L}_u = \alpha w * (\lambda_1 \mathcal{L}_{global} + \lambda_2 \mathcal{L}_{local}) \tag{12}$$

where

$$w = \begin{cases} \frac{1}{N_w}, & \text{Unreliable} \\ \frac{1}{N_u - N_w}, & \text{others} \end{cases} \tag{13}$$

where N_w and N_u represent the number of unreliable samples and the number of samples from the unlabeled client. α is the warming-up factor. As the communication rounds progress, the reliability of unreliable samples gradually increases. We utilize a linear function to achieve it:

$$\alpha_t = \alpha_0 + (\alpha_n - \alpha_0)\frac{t}{\mathrm{Rounds}} \tag{14}$$

where a_t,a_0 and a_n are the warming-up weights for the first round, the t round, and the final round, respectively.Rounds is the total number of synchronization round. By reweighting the loss function for potential rare classes, we enable more even class distribution learning on each client before federated aggregation. The whole algorithm in the unlabeled client is presented in Algorithm 1.

Experiments

Experimental Setup

Datasets. We evaluate our method on two medical image classification tasks, $i.e.$, skin lesion diagnosis for dermoscopy images and intracranial hemorrhage (ICH) diagnosis for brain CT slices. We perform the HAM10000 dataset (Tschandl, Rosendahl, and Kittler 2018) for skin lesion classification, which contains 10015 images and 7 classes. For ICH diagnosis, we follow the setup in FedIRM (Liu et al. 2021b) that randomly selects 25,000 images from the RSNA ICH dataset (Flanders et al. 2020), which consists of 5 subtypes. For both benchmark datasets, we employ 70% for training, 10% for validation, and 20% for testing. We apply the same preprocessing to both datasets that we resize the images into 240×240, randomly crop a 224×224 region, and normalize before input to the network.

Federated Learning Setting. We follow the exits method (Li, Li, and Wang 2023) using Dirichlet distribution to generate a Non-IID data partition among 1 labeled client and 9 unlabeled clients, where $Dir(\gamma) = 0.8$.

Implementation Details. We implement our method in PyTorch with the SGD optimizer. We utilize ResNet18 pretrained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) as the backbone network followed by two MLP layers and a fully connected layer for classification. The same classification network is employed across all compared methods for fair comparison. The learning rates for labeled and unlabeled clients are 0.02 and 0.01 respectively. The batch size is 12 for the HAM10000 dataset and 24 for the RSNA ICH dataset. We set 1 local epoch for all clients and train for 1000 rounds (200 warm-ups). The loss function parameters λ_1 and λ_2 are both set to 0.02. We enpirically set temperature parameter $\tau = 0.5$, confident class threshold $\beta = 0.4$, warming-up weights $\alpha_0 = 0.01$ and $\alpha_n = 0.1$, low-rank threshold $t_l = 5$ and high-rank threshold $t_h = 6$.

Comparisons with State-of-the-arts

Compare Method. We compare our approach with the state-of-the-art methods including FedIRM (Liu et al. 2021b), Fed-Consist (Yang et al. 2021), RSCFed (Liang et al. 2022) and CBAFed (Li, Li, and Wang 2023). Furthermore, we conduct a comparative analysis of our approach against FedAvg (McMahan et al. 2017), which serves as the upper bound training by 10 labeled clients and lower bound training by 1 labeled client.

Implementation Details. Following the settings in state-of-the-art FSSL methods (Li, Li, and Wang 2023), we assign a weight of 50% to the labeled client, while the nine unlabeled clients share the remaining 50% weight in each FSSL synchronization round.

Quantitative Comparisons. Table 1 presents outstanding experimental results comparing our approach with other methods on two medical datasets. Our approach achieves the best performance, significantly surpassing the lower bound. This demonstrates the capability of our approach to balance knowledge and master rare classes. Furthermore, FedIRM and Fed-Conist exhibit lower performance than the lower bound, indicating their methods are ineffective in the Non-IID setting.

Evaluation on Two Datasets. For skin lesion diagnosis, our approach achieves the best accuracy of 70.99% (1.2% improvement), AUC of 83.64% (0.58% improvement), precision of 42.22% (4.23% improvement), and recall of 35.63% (2.91% improvement). For intracranial hemorrhage diagnosis, we also attain the highest accuracy of 63.10% (3.76% improvement), AUC of 79.55% (1.38% improvement), precision of 47.77% (0.21% improvement), and recall of 46.93% (3.92% improvement). Our approach outperforms the state-of-the-art methods in terms of all four metrics on both datasets. The superior performance primarily arises from our framework's ability to mitigate global-local discrepancies and account for potential rare samples among unlabeled clients.

Two Labeled Clients. In order to demonstrate the superior performance of FedCD, we conducted a comparison that set the number of labeled clients to 2 and unlabeled clients to 8 with other FSSL methods on the HAM10000 dataset. As Table 2 shows, our method achieved the best performance. The increased labeled data enabled more informed global

Labeling Strategy	Method	Client Num.		Metrics			
		labeled	unlabeled	Acc. (%)	AUC (%)	Precision (%)	Recall (%)
		Task 1: Skin Lesion Diagnosis					
Fully supervised	FedAvg (upper-bound)	10	0	80.42	93.47	71.57	54.39
	FedAvg(lower-bound)	1	0	68.07	79.02	34.86	31.37
Semi supervised	Fed-Consist	1	9	67.84	81.25	37.49	29.08
	FedIRM	1	9	68.39	81.6	37.49	31.81
	RSCFed	1	9	69.09	82.59	37.94	32.59
	CBAFed	1	9	69.79	83.06	37.99	32.75
	Ours	1	9	**70.99**	**83.64**	**42.22**	**35.63**
		Task 2: Intracranial Hemorrhage Diagnosis					
Fully supervised	FedAvg(upper-bound)	10	0	72.03	88.19	62.85	59.86
	FedAvg(lower-bound)	1	0	59.27	77.45	46.49	42.27
Semi supervised	Fed-Consist	1	9	58.96	75.86	46.07	42.04
	FedIRM	1	9	58.98	74.79	45.37	42.88
	RSCFed	1	9	59.32	77.51	47.53	43.04
	CBAFed	1	9	59.34	78.17	47.56	43.01
	Ours	1	9	**63.10**	**79.55**	**47.77**	**46.93**

Table 1: Resluts on the HAM10000 and RSNA ICH datasets under heterogeneous data partition. We employ four commonly used metrics for method comparison, including Accuracy(Acc.), Area under the ROC Curve (AUC), Precision, and Recall. The best results are in bold. It reports that our method achieves the best performance among all methods.

Method	Client Num.		Metrics	
	labeled	unlabeled	Acc.(%)	AUC(%)
FedAvg$^+$	10	0	80.42	93.47
FedAvg$^-$	2	0	68.97	85.96
Fed-Consist	2	8	67.54	85.55
FedIRM	2	8	68.24	85.06
RSCFed	2	8	69.19	86.73
CBAFed	2	8	69.34	88.07
Ours	2	8	**71.09**	**89.15**

Table 2: Comparison of our method against Fed-Consist, FedIRM, RSCFed and CBAFed with the number of labeled and unlabeled clients set to 2 and 8. Superscript $^+$ and $^-$ denote the upper and lower bound, respectively.

Method	Client Num.		Metrics	
	labeled	unlabeled	Acc.(%)	AUC(%)
FedAvg$^+$	10	0	85.07	95.67
FedAvg$^-$	1	0	70.33	82.92
CBAFed	1	9	70.59	86.23
Ours	1	9	**71.69**	**87.37**

Table 3: Comparison of our method against state-of-the-art methods CBAFed with ViT-Tiny backbone.

teacher knowledge, aiding unlabeled clients in identifying confident classes and unreliable samples more effectively.

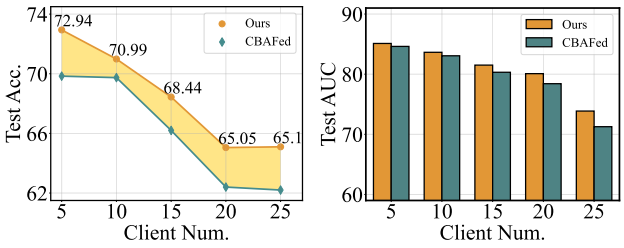


Figure 5: The accuracy score and AUC vary with the change in the number of unlabeled clients. Noted that the number of labeled clients remains at 1.

Unlabeled Client Ratio. We also evaluated our method with 1 labeled client and varying unlabeled clients on the HAM10000 dataset. As Figure 5 shows, our method consistently outperforms CBAFed as unlabeled clients increase from 5 to 25. As the number of clients increases, individual data diminishes, which potentially compromises balance. However, our proposed method exhibits improved performance, showcasing its ability for balanced learning even in the presence of reduced local data.

ViT Backbone. Recently, vision transformers (ViT) have been widely utilized in federated learning due to their remarkable robustness in handling heterogeneous data (Li, Li, and Wang 2023; Qu et al. 2022). Hence, we employ ViT-Tiny (Dosovitskiy et al. 2020) as the backbone for experiments on the HAM10000 dataset. As shown in Table 3, our method exhibits superior performance surpassing the state-

	CAB	DTD	Acc.(%)	AUC(%)
Basic	×	×	68.07	79.02
Basic+CAB	✓	×	70.29	82.41
Basic+DTD	×	✓	70.44	83.53
Ours	✓	✓	**70.99**	**83.64**

Table 4: Ablation studies on the effectiveness of dual teacher distillation and class awareness balance.

Local Teacher	Global Teacher	Knowledge Purification	Metrics	
			Acc.(%)	AUC(%)
✓	×	×	68.07	79.02
✓	×	✓	70.24	82.63
✓	✓	×	69.94	82.79
✓	✓	✓	**70.44**	**83.53**

Table 5: Ablation studies on the effectiveness of dual teachers and knowledge purification.

of-the-art approach. These consistent improvements demonstrate the efficacy of the proposed dual teacher distillation and class awareness balance techniques, which translate to gains regardless of the underlying feature extractor used.

Efficiency Analysis. CBAFed requires 11 local epochs due to residual connections, whereas our method achieves higher performance with fewer local epochs (set to 1). This substantial difference highlights the improved efficiency and faster convergence of our approach.

Ablation Studies

In this section, we aim to validate our core insights by adding or removing the proposed components, $i.e.$, dual teacher distillation (DTD) and class awareness balance module (CAB). We further discuss the reasons behind the outstanding performance of each component. All the experiments in this section are based on the HAM10000 dataset.

Effectiveness of DTD and CAB. Table 4 show the result ablated each component. It can be observed that after incorporating CAB, the accuracy score increased by 2.22%, and the AUC increased by 3.39%. The addition of DTD resulted in an improvement of 2.37% in accuracy score and an increase of 4.51% in AUC. With both components, the model achieved significant improvements compared to the baseline, where the accuracy score increased by 2.92% and the AUC improved by 4.62%, which validates the effectiveness of the proposed method.

Effectiveness of Subcomponents in DTD. In our DTD method, we propose two subcomponents: dual teachers (DT) and knowledge purification (KP). As Table 5 shows, DT and KP can effectively balance and optimize global and local knowledge, resulting in the improvement of Acc. (1.82%) and AUC(4.51%). These improvements underscore the significance of the purified distillation knowledge from global and local teachers.

Effectiveness of Identifying Unreliable Samples. The intention behind the class awareness balance (CAB) module is to employ confident classes to detect potential samples belonging to rare classes. As depicted in Figure 6, our approach exhibits significant accuracy in accurately recognizing rare class samples in a majority of clients. Despite some clients potentially misclassifying rare samples, these samples can be regarded as intricate instances that challenge the model and ultimately contribute to the improvement of the overall model performance.

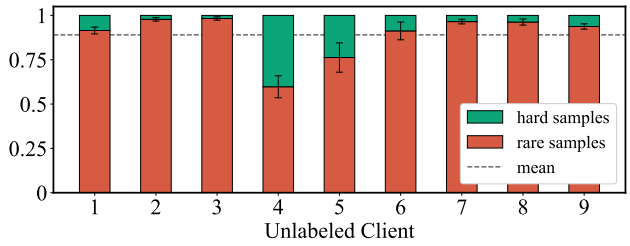


Figure 6: The proportion of correctly identified rare samples from all mined samples in the unlabeled clients during the training process. Noted our method achieves higher average accuracy in mining rare class samples.

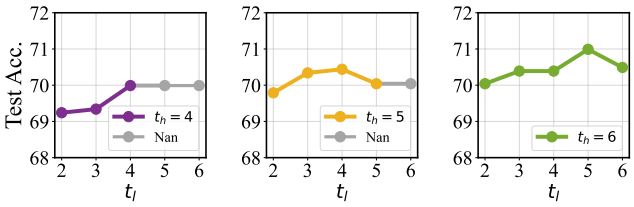


Figure 7: The accuracy score varies with the change in t_l and t_h.The gray curve represents illegal values.

Hyper-parameters. Note that our method involves two crucial hyperparameters, they are the low-rank threshold t_l and the high-rank threshold t_h in the CAB module. As Table 7 shows, the occurrence of confidence classes in lower intervals indicates that the samples are more likely to belong to rare classes.

Limitations

Both DTD and CAB are limited by the performance of the warm-up model obtained from labeled data. In the presence of limited labeled data availability, the performance of the server model may be impeded.

Conclusion

In this paper, we endeavored to mitigate two salient challenges in FSSL, the limited knowledge acquired by unlabeled clients and how to improve the performance of FSSL on imbalanced datasets. To overcome these challenges, we proposed the dual teacher distillation to sublimate both global and local knowledge, as well as the class awareness balance module to mine local rare classes for more balanced learning. Our method achieved the best performance across different medical tasks and different experiment settings, demonstrating its effectiveness and superiority.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (Nos. 62273241 and 61973221), Natural Science Foundation of Guangdong Province, China (No. 2019A1515011165), and the Hong Kong RGC Theme-based Research Scheme (project no.T45-401/22-N).

References

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

Antunes, R. S.; André da Costa, C.; Küderle, A.; Yari, I. A.; and Eskofier, B. 2022. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology*, 13(4): 1–23.

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32.

Che, L.; Long, Z.; Wang, J.; Wang, Y.; Xiao, H.; and Ma, F. 2021. Fedtrinet: A pseudo labeling method with three players for federated semi-supervised learning. In *IEEE International Conference on Big Data*, 715–724. IEEE.

Chen, B.; Jiang, J.; Wang, X.; Wan, P.; Wang, J.; and Long, M. 2022a. Debiased self-training for semi-supervised learning. *Advances in Neural Information Processing Systems*, 35: 32424–32437.

Chen, H.; Jin, Y.; Jin, G.; Zhu, C.; and Chen, E. 2021. Semisupervised semantic segmentation by improving prediction confidence. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9): 4991–5003.

Chen, H.-Y.; and Chao, W.-L. 2020. Fedbe: Making bayesian model ensemble applicable to federated learning. *ArXiv:2009.01974*.

Chen, Z.; Yang, C.; Zhu, M.; Peng, Z.; and Yuan, Y. 2022b. Personalized Retrogress-Resilient Federated Learning Toward Imbalanced Medical Data. *IEEE Transactions on Medical Imaging*, 41(12): 3663–3674.

Dong, N.; and Voiculescu, I. 2021. Federated contrastive learning for decentralized unlabeled medical images. In *Medical Image Computing and Computer Assisted Intervention*, 378–387. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dou, Q.; So, T. Y.; Jiang, M.; Liu, Q.; Vardhanabhuti, V.; Kaissis, G.; Li, Z.; Si, W.; Lee, H. H.; Yu, K.; et al. 2021. Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study. *NPJ digital medicine*, 4(1): 60.

Du, Y.; Shen, Y.; Wang, H.; Fei, J.; Li, W.; Wu, L.; Zhao, R.; Fu, Z.; and Liu, Q. 2022. Learning from future: A novel self-training framework for semantic segmentation. *Advances in Neural Information Processing Systems*, 35: 4749–4761.

Flanders, A. E.; Prevedello, L. M.; Shih, G.; Halabi, S. S.; Kalpathy-Cramer, J.; Ball, R.; Mongan, J. T.; Stein, A.; Kitamura, F. C.; Lungren, M. P.; et al. 2020. Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. *Radiology: Artificial Intelligence*, 2(3): e190211.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17.

Guo, Q.; Qi, Y.; Qi, S.; and Wu, D. 2022. Dual Class-Aware Contrastive Federated Semi-Supervised Learning. *arXiv:2211.08914*.

Huang, W.; Ye, M.; Shi, Z.; Li, H.; and Du, B. 2023. Rethinking federated learning with domain shift: A prototype view. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16312–16322. IEEE.

Huynh, T.; Nibali, A.; and He, Z. 2022. Semi-supervised learning for medical image classification using imbalanced training data. *Computer Methods and Programs in Biomedicine*, 216: 106628.

Itahara, S.; Nishio, T.; Koda, Y.; Morikura, M.; and Yamamoto, K. 2021. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. *IEEE Transactions on Mobile Computing*, 22(1): 191–205.

Jeong, W.; Yoon, J.; Yang, E.; and Hwang, S. J. 2020. Federated semi-supervised learning with inter-client consistency & disjoint learning. *arXiv:2006.12097*.

Jiang, M.; Yang, H.; Li, X.; Liu, Q.; Heng, P.-A.; and Dou, Q. 2022. Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance. In *Medical Image Computing and Computer-Assisted Intervention*, 196–206. Springer.

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

Kassem, H.; Alapatt, D.; Mascagni, P.; AI4SafeChole, C.; Karargyris, A.; and Padoy, N. 2022. Federated cycling (FedCy): Semi-supervised Federated Learning of surgical phases. *IEEE Transactions on Medical Imaging*.

Kim, J.; Jang, J.; Seo, S.; Jeong, J.; Na, J.; and Kwak, N. 2022. Mum: Mix image tiles and unmix feature tiles for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14512–14521.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, M.; Li, Q.; and Wang, Y. 2023. Class Balanced Adaptive Pseudo Labeling for Federated Semi-Supervised Learning.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16292–16301.

Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 10713–10722.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020a. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020b. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2: 429–450.

Liang, X.; Lin, Y.; Fu, H.; Zhu, L.; and Li, X. 2022. RSCFed: random sampling consensus federated semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10154–10163.

Lin, H.; Lou, J.; Xiong, L.; and Shahabi, C. 2021. Semifed: Semi-supervised federated learning with consistency and pseudo-labeling. *arXiv preprint arXiv:2108.09412*.

Lin, H.; Zhang, Y.; Qiu, Z.; Niu, S.; Gan, C.; Liu, Y.; and Tan, M. 2022. Prototype-guided continual adaptation for class-incremental unsupervised domain adaptation. In *European Conference on Computer Vision*, 351–368. Springer.

Liu, Q.; Chen, C.; Qin, J.; Dou, Q.; and Heng, P.-A. 2021a. 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*, 1013–1023.

Liu, Q.; Yang, H.; Dou, Q.; and Heng, P.-A. 2021b. Federated semi-supervised medical image classification via inter-client relation matching. In *Medical Image Computing and Computer Assisted Intervention*, 325–335. Springer.

Liu, Q.; Yu, L.; Luo, L.; Dou, Q.; and Heng, P. A. 2020. Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging*, 39(11): 3429–3440.

Long, Z.; Che, L.; Wang, Y.; Ye, M.; Luo, J.; Wu, J.; Xiao, H.; and Ma, F. 2020. FedSiam: Towards adaptive federated semi-supervised learning. *arXiv:2012.03292*.

Long, Z.; Wang, J.; Wang, Y.; Xiao, H.; and Ma, F. 2021. FedCon: A contrastive framework for federated semi-supervised learning. *arXiv:2109.04533*.

Mammen, P. M. 2021. Federated learning: Opportunities and challenges. *arXiv:2101.05428*.

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

Olsson, V.; Tranheden, W.; Pinto, J.; and Svensson, L. 2021. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1369–1378.

Qu, L.; Zhou, Y.; Liang, P. P.; Xia, Y.; Wang, F.; Adeli, E.; Fei-Fei, L.; and Rubin, D. 2022. Rethinking architecture design for tackling data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10061–10071.

Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H. R.; Albarqouni, S.; Bakas, S.; Galtier, M. N.; Landman, B. A.; Maier-Hein, K.; et al. 2020. The future of digital health with federated learning. *NPJ digital medicine*, 3(1): 119.

Shi, Y.; Chen, S.; and Zhang, H. 2022. Uncertainty minimization for personalized federated semi-supervised learning. *IEEE Transactions on Network Science and Engineering*, 10(2): 1060–1073.

Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8432–8440.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1): 1–9.

Wang, J.; Zeng, S.; Long, Z.; Wang, Y.; Xiao, H.; and Ma, F. 2023. Knowledge-Enhanced Semi-Supervised Federated Learning for Aggregating Heterogeneous Lightweight Clients in IoT. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 496–504. SIAM.

Xu, H.; Liu, L.; Bian, Q.; and Yang, Z. 2022. Semi-supervised semantic segmentation with prototype-based consistency regularization. *Advances in Neural Information Processing Systems*, 35: 26007–26020.

Yang, D.; Xu, Z.; Li, W.; Myronenko, A.; Roth, H. R.; Harmon, S.; Xu, S.; Turkbey, B.; Turkbey, E.; Wang, X.; et al. 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. *Medical Image Analysis*, 70: 101992.

Yang, Q.; Liu, X.; Chen, Z.; Ibragimov, B.; and Yuan, Y. 2022. Semi-supervised Medical Image Classification with Temporal Knowledge-Aware Regularization. In *Medical Image Computing and Computer Assisted Intervention*, 119–129. Springer.

Zhang, Z.; Ma, S.; Nie, J.; Wu, Y.; Yan, Q.; Xu, X.; and Niyato, D. 2021a. Semi-supervised federated learning with non-iid data: Algorithm and system design. In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, 157–164. IEEE.

Zhang, Z.; Yang, Y.; Yao, Z.; Yan, Y.; Gonzalez, J. E.; Ramchandran, K.; and Mahoney, M. W. 2021b. Improving semi-supervised federated learning by reducing the gradient diversity of models. In *IEEE International Conference on Big Data*, 1214–1225. IEEE.

Zhu, W.; and Luo, J. 2022. Federated medical image analysis with virtual sample synthesis. In *Medical Image Computing and Computer-Assisted Intervention*, 728–738. Springer.