Class-Aware Active Annotation in Federated Semi-Supervised Learning for Medical Image Classification

Meiting Xue[1], Miaoqi Li[2], Yukun Shi[1], Yan Zeng[2], Jilin Zhang[1*] Jing Ma[3]
[1]School of Cyberspace, Hangzhou Dianzi University, Hangzhou, 310018, China
[2]School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China
[3]Hangzhou TCM Hospital Affiliated to Zhejiang Chinese Medical University, Hangzhou, 310007
munuan@hdu.edu.cn,mql@hdu.edu.cn,yukun.shi@hdu.edu.cn,yz@hdu.edu.cn,jilin.zhang@hdu.edu.cn,
2021b085@zcmu.edu.cn

Abstract

In medical image classification, data privacy constraints and the high cost of expert annotations pose significant challenges to building generalizable models. Federated semi-supervised learning (FSSL), which combines the privacy-preserving nature of federated learning with the label efficiency of semi-supervised learning, offers a promising direction. However, in real-world deployments, client data often exhibits highly non-independent and identically distributed (Non-IID) characteristics. This distributional heterogeneity undermines the reliability of pseudo-labels generated by global models, ultimately limiting model generalization. A key limitation of existing FSSL approaches lies in their reliance on a static labeled set fixed prior to training. Such strategies lack the ability to adaptively correct pseudo-label noise or address class imbalance throughout training, particularly under Non-IID settings. To address this, we propose FSSAL, a novel framework that introduces an active learning component into the FSSL pipeline. By continuously identifying informative and representative samples during training, our method adaptively refines the labeled set and enhances the model's robustness to distribution shifts. FSSAL employs client-private models for pseudo-label generation to reduce global bias, applies a class-aware dynamic thresholding mechanism to ensure more reliable and balanced label selection, and incorporates a sample selection strategy guided by both feature diversity and model uncertainty. Extensive experiments on four public medical image classification datasets demonstrate that FSSAL consistently outperforms competitive FSSL methods in accuracy and F1-score, especially under highly Non-IID conditions, highlighting its robustness and practical potential.

Introduction

Deep learning has achieved remarkable success in medical image classification, powering applications such as diabetic retinopathy screening, skin lesion diagnosis, and intracranial hemorrhage detection (Litjens et al. 2017). However, two persistent challenges continue to hinder the deployment of robust and generalizable models in real-world medical scenarios: data privacy constraints (Sheller et al. 2020) and limited availability of labeled samples. Due to the sensitive nature of medical data and regulatory restrictions, data cannot be directly shared across institutions. Moreover, expert annotation is costly and time-consuming, making large-scale supervised learning impractical in many clinical contexts.

Federated semi-supervised learning (FSSL), which combines the privacy-preserving nature of federated learning (FL) with the label efficiency of semi-supervised learning (SSL), offers a promising paradigm for collaborative model training without requiring centralized data collection. In FSSL, a global model is trained across decentralized clients by leveraging both labeled and unlabeled data locally. While this approach mitigates some limitations of data isolation and annotation scarcity, it remains highly vulnerable to the non-independent and identically distributed (Non-IID) nature of medical data across clients (Li et al. 2019). Distributional heterogeneity, such as class imbalance and demographic differences between hospitals, often leads to unreliable pseudo-labels generated by the global model. These noisy pseudo-labels can propagate errors during training, especially when minority classes are systematically underrepresented.

We observe that many existing FSSL approaches rely on a fixed set of labeled samples established at the start of training. Once initialized, the model passively learns from pseudo-labels on the remaining unlabeled data, without revisiting the decision of which samples to label. This static, pseudo-label-driven learning paradigm fails to adapt to the evolving model state and dynamic error patterns—particularly in Non-IID settings. For instance, RSCFed (Liang et al. 2022) employs random sampling consensus and distance-reweighted aggregation to mitigate heterogeneity but overlooks class imbalance in pseudo-label generation. FedCD (Liu et al. 2024) uses a dual-teacher architecture to enhance minority class recognition yet faces pseudo-label instability and increased costs in data-scarce scenarios. Similarly, SAGE (Liu et al. 2025) dynamically adjusts pseudo-labels via confidence discrepancies but falters under extreme imbalance or noisy initial estimates in heterogeneous environments.

We argue that this lack of dynamic supervision is a key bottleneck in current FSSL systems. To address this, we explore the integration of active learning (AL) into FSSL to selectively and iteratively expand the labeled set during training. AL has been effective in centralized settings for acquiring high-value labels under budget constraints, yet remains

*Corresponding author

underexplored in federated semi-supervised regimes, particularly under Non-IID conditions. By identifying informative and representative samples based on model uncertainty and feature structure, active sample selection enables targeted correction of pseudo-label noise and improved class coverage.

In this paper, we propose **FSSAL**, a Federated Semi-Supervised Active Learning framework that introduces an active annotation mechanism into the FSSL pipeline. Rather than relying solely on initial supervision, FSSAL periodically selects high-value samples to be labeled during training, based on both local and global signals. Our framework integrates three key components: (1) private model-specific pseudo-labeling to mitigate early-stage global bias; (2) a class-aware dynamic thresholding strategy that adapts to global class distribution shifts; and (3) a dual-view active sample selection mechanism that fuses local feature clustering and global entropy-based uncertainty estimation.

To validate the effectiveness of our approach, we conduct extensive experiments on four publicly available medical image classification datasets. Each dataset is partitioned under varying Non-IID conditions to emulate realistic institutional disparities. Compared to competitive FSSL baselines, FSSAL consistently achieves superior performance in terms of accuracy, F1-score, and minority-class coverage, particularly under strongly skewed client distributions. Further ablation studies confirm the independent contributions of pseudo-labeling, threshold control, and active sample selection to the observed performance gains.

Our main contributions are summarized as follows:

- We propose **FSSAL**, a novel framework that integrates active learning into federated semi-supervised learning to dynamically improve annotation efficiency and pseudo-label robustness under Non-IID conditions.
- We introduce a class-aware thresholding strategy and a dual-model active sample selection algorithm that jointly enhance pseudo-label precision and class-level balance.
- We perform extensive evaluations on four medical datasets, demonstrating that FSSAL outperforms state-of-the-art FSSL methods, especially in highly heterogeneous data environments.

Related Work

Federated Learning

Federated Learning (FL) is a distributed machine learning paradigm that enables collaborative model training across multiple decentralized clients without sharing raw data, thereby preserving data privacy. Due to its privacy-preserving nature, FL has garnered increasing attention in the field of medical imaging. A significant line of FL research addresses client-side data heterogeneity, which often manifests as class imbalance and domain shifts. For instance, FedIIC (Wu et al. 2022) tackles class imbalance by optimizing feature representations, while FedAWA (Yue et al. 2023) introduces adaptive aggregation and dynamic feature fusion to enhance classification performance. FedMoP (Du et al. 2024) mitigates catastrophic forgetting and aggregation issues through model projection on both client and server sides. Nevertheless, most FL approaches assume sufficient labeled data, which is often unrealistic in medical scenarios. The scarcity of labeled data has motivated growing interest in leveraging abundant unlabeled data for federated training.

Semi-Supervised Leaning

Semi-Supervised Learning (SSL) leverages a small amount of labeled data along with a large pool of unlabeled data to improve model generalization. Popular SSL strategies include consistency regularization, pseudo-labeling, and hybrid approaches. Consistency-based methods, such as UDA (Xie et al. 2020), enforce invariance between weakly and strongly augmented views of unlabeled inputs. Pseudo-labeling methods generate soft or hard labels using model predictions; for example, FreeMatch (Wang et al. 2023) filters pseudo-labels based on confidence thresholds to reduce label noise. Hybrid methods like FlexMatch (Zhang et al. 2021a) combine consistency regularization and curriculum-based pseudo-labeling for robust training. However, directly applying these methods in FL settings poses two primary challenges: (1) Pseudo-label reliability may degrade due to local distribution biases (Itahara et al. 2021; Zhang et al. 2021b); and (2) Medical datasets are often highly imbalanced, making it difficult to ensure robust learning across all classes (Che et al. 2021).

Federated Semi-Supervised Leaning

Federated Semi-Supervised Learning (FSSL) combines the strengths of FL and SSL to utilize unlabeled data while maintaining data privacy. Recent works have explored FSSL for medical image classification to address the dual constraints of limited annotations and distributional heterogeneity. For instance, FedIRM (Liu et al. 2021) leverages feature clustering for local model updates, although its reliance on global disease relationship modeling may limit adaptability. RSCFed (Liang et al. 2022) uses random sampling consensus and distance-reweighted aggregation to mitigate heterogeneity but does not perform pseudo-label filtering. FedLoKe (Zhang et al. 2023) improves global accuracy by incorporating local knowledge, though it struggles under severely imbalanced data. CBAFed (Li, Li, and Wang 2023) introduces class-wise adaptive pseudo-label weighting but remains sensitive to pseudo-label quality. FedCD (Liu et al. 2024) utilizes dual-teacher models to enhance minority class learning, yet suffers from increased computation and label instability in low-data regimes. SAGE (Liu et al. 2025) dynamically adjusts pseudo-labels using confidence discrepancies, but its efficacy is diminished under noisy or imbalanced conditions.

While these approaches enhance FSSL from a modeling perspective, most overlook the pivotal role of data quality. Improper utilization of unlabeled data—especially under highly Non-IID conditions—can introduce noisy pseudo-labels that propagate and amplify over training rounds. Moreover, not all unlabeled samples contribute equally; low-quality samples may even hinder model robustness. Therefore, an open problem in FSSL lies in effectively filtering

and utilizing informative unlabeled data to improve learning efficiency and generalization.

Active Learning

Active Learning (AL) is a label-efficient strategy that reduces annotation costs by selectively querying the most informative samples for labeling (Settles 2009). In medical imaging, AL has been widely applied to minimize expert annotation burden. Ren et al. (Ren, Zhang, and Liu 2021) combined uncertainty sampling with deep convolutional networks to improve classification accuracy while reducing the volume of labeled data. Zhang et al. (Zhang, Li, and Li 2022) proposed a GAN-based AL framework, leveraging generative models for pseudo-labeling and employing AL to select representative samples. While AL has proven successful in centralized SSL, its integration with FSSL remains underexplored, particularly under Non-IID distributions. Incorporating AL into FSSL holds promise for adaptively improving pseudo-label quality, correcting class imbalance, and enhancing annotation efficiency.

Methodology

Figure 1 illustrates the overall architecture of our proposed Federated Semi-Supervised Active Learning (FSSAL) framework. Each client possesses a small labeled dataset and a large unlabeled dataset locally. To ensure reliable pseudo-label generation under data heterogeneity, we introduce a novel dual-model design. In contrast to conventional federated semi-supervised learning (FSSL) approaches that rely on a single model, each client in our framework maintains both a local model for global aggregation and a private model trained solely on its local labeled data to generate pseudo-labels. Furthermore, we propose a class-aware dynamic thresholding mechanism to select high-confidence unlabeled samples for semi-supervised training. For the remaining low-confidence samples, we integrate an active learning strategy to select informative and class-diverse samples for annotation. The components of our framework are detailed in the following sections.

Problem Setup

We consider a federated semi-supervised learning scenario with K medical clients coordinated by a central server. Each client $k \in \{1, \ldots, K\}$ has a local dataset $D_k = D_k^l \cup D_k^u$, where $D_k^l = \{(x_i, y_i)\}_{i=1}^{n_l}$ is a small labeled subset and $D_k^u = \{x_i\}_{i=1}^{n_u}$ is a larger unlabeled subset with $n_l \ll n_u$. The data distribution across clients is assumed to be non-IID and class-imbalanced. Each client trains a local model θ_k using its local data, and the goal is to collaboratively learn a global model θ_g without sharing raw data. The overall training objective is defined as:

$$\min_{\theta} \mathcal{L} = \frac{1}{|D^l|} \sum_{(x,y)\in D^l} \mathcal{L}_{\text{sup}}((x,y),\theta) + \lambda \cdot \frac{1}{|D^u|} \sum_{x\in D^u} \mathbf{1}(p(x) > \tau_c) \cdot \mathcal{L}_{\text{unsup}}((x,\hat{y}),\theta) \tag{1}$$

where $\hat{y} = \arg\max p(x)$ denotes the predicted pseudo-label, τ_c is the class-specific confidence threshold, and $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{unsup}}$ represent the supervised and unsupervised loss terms, respectively.

Decoupled Dual-Model Pseudo-Labeling Framework

Traditional FSSL approaches typically generate pseudo-labels using the global model, but due to data heterogeneity, such pseudo-labels are often biased toward majority classes and may degrade model performance. To address this, we introduce a decoupled dual-model design on each client: a private model ω for pseudo-labeling and a local model θ for semi-supervised learning and global aggregation. This structure aims to enhance pseudo-label accuracy under non-IID conditions.

Specifically, the private model ω_k is trained only on the labeled set D_k^l of client k, capturing local label distribution. The training objective for ω_k at round t is:

$$\mathcal{L}_k^{\text{private}} = \frac{1}{|D_k^l|} \sum_{(x,y)\in D_k^l} \text{CE}\left(y, f(\mathcal{A}_w(x); \omega_k^t)\right) \tag{2}$$

where CE denotes the cross-entropy loss and $\mathcal{A}_w$ is a weak augmentation. The model is updated via stochastic gradient descent (SGD) to obtain ω_k^{t+1}.

To mitigate overfitting due to the small labeled set, we adopt an Exponential Moving Average (EMA) fusion of the local private model and the global model, inspired by FedLoKe (Zhang et al. 2023):

$$\omega_k^{t+1} = \alpha_{\text{ema}} \cdot \omega_k^{t+1} + (1 - \alpha_{\text{ema}}) \cdot \theta_g^{t+1} \tag{3}$$

where α_{ema} controls the fusion ratio.

The updated private model is then used to generate soft pseudo-labels for D_k^u:

$$p = \text{softmax}(f(\mathcal{A}_w(x); \omega_k)), \quad \hat{y} = \arg\max(p) \tag{4}$$

These pseudo-labels are considered as candidates and will be filtered based on class-aware thresholds.

Meanwhile, the local model θ_k is initialized from the global model θ_g and trained on both D_k^l and the pseudo-labeled set D_k^p. The local training loss is:

$$\mathcal{L}_k^{\text{local}} = \mathcal{L}_k^{\text{sup}} + \lambda \cdot \mathcal{L}_k^{\text{unsup}} \tag{5}$$

$$\mathcal{L}_k^{\text{sup}} = \frac{1}{|D_k^l|} \sum_{(x,y)\in D_k^l} \text{CE}(y, f(\mathcal{A}_w(x); \theta_k)) \tag{6}$$

$$\mathcal{L}_k^{\text{unsup}} = \frac{1}{|D_k^p|} \sum_{(x,\hat{y})\in D_k^p} \text{CE}(\hat{y}, f(\mathcal{A}_s(x); \theta_k)) \tag{7}$$

where $\mathcal{A}_s$ is a strong augmentation. After local optimization, θ_k is uploaded to the server and aggregated using FedAvg. The decoupling ensures the private model focuses on robust pseudo-labeling, while the local model contributes to federated optimization.

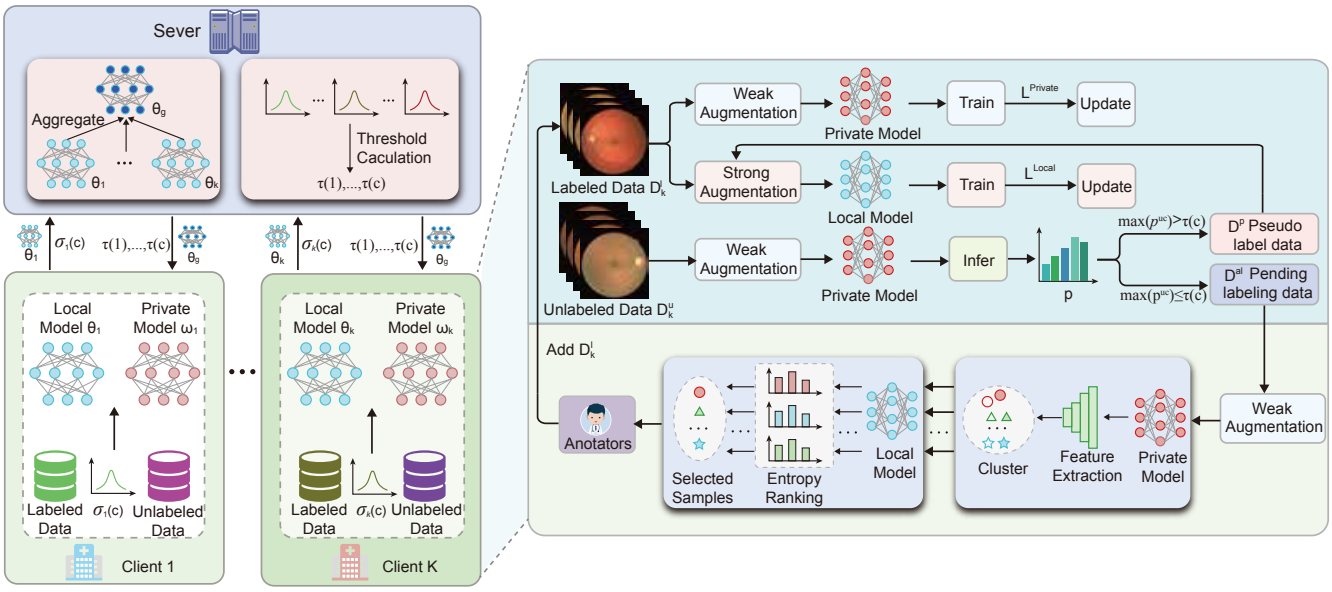

Figure 1: An overview of our proposed FSSAL framework.

Labeled Data-Aware Class Threshold

As mentioned, the private model provides candidate pseudo-labels for unlabeled data. A critical step is to set appropriate confidence thresholds to filter reliable labels. Instead of fixed thresholds, we propose class-specific dynamic thresholds derived from the labeled data distribution, enabling adaptation to class imbalance and distribution shifts.

Each client reports the number of labeled samples per class to the server. For client k and class c, the count is:

$$\sigma_k(c) = \sum_{m=1}^{|D_k^l|} \mathbf{1}(y_m^l = c) \tag{8}$$

The server aggregates global counts:

$$\sigma_s(c) = \sum_{k=1}^{K} \sigma_k(c) \tag{9}$$

The empirical class distribution is:

$$\beta_s(c) = \frac{\sigma_s(c)}{\sum_{c=1}^{C} \sigma_s(c)} \tag{10}$$

Following CBAFed (Li, Li, and Wang 2023), we introduce a standard deviation term to quantify class imbalance:

$$\text{std}(\beta_s) = \sqrt{\frac{1}{C-1}\sum_{c=1}^{C}(\beta_s(c) - \beta_s')^2}, \quad \text{where } \beta_s' = \frac{1}{C}\sum_{c=1}^{C}\beta_s(c) \tag{11}$$

This metric allows the threshold to adapt: lowering the threshold for rare classes and maintaining higher ones for common classes. The class-wise threshold is thus computed as:

$$\tau(c) = \beta_s(c) + \tau - \text{std}(\beta_s) \tag{12}$$

Here, τ is the base threshold, a global constant that adjusts the overall confidence level required for pseudo-labeling. For each unlabeled sample $x_i^u \in D_k^u$, if its predicted probability for class c exceeds $\tau(c)$, it is added to D_k^p as a pseudo-labeled sample; otherwise, it is excluded.

Active Class-Balanced Sample Selection

Low-confidence samples filtered out in the previous step may still hold valuable information. Discarding them risks reducing generalization. Thus, we adopt an active learning strategy to select the most informative and class-diverse samples from the uncertain set $D_k^{al} = D_k^u - D_k^p$ for annotation.

The active learning process is conducted locally. First, the private model ω_k is used to extract diversity features from the penultimate layer:

$$F_i = \varphi(x_i^{al}; \omega_k) \in \mathbb{R}^d \tag{13}$$

where $\varphi(\cdot)$ denotes the feature extractor and d is the feature dimension.

Then, KMeans++ clustering is applied to F to form B clusters (each representing one sample in the labeling budget):

$$\{C_1, \ldots, C_B\} = \text{KMeans++}(F), \quad \forall b, |C_b| \geq 1 \tag{14}$$

Next, uncertainty sampling is applied using the current global model θ_g, selecting the most uncertain sample (with highest entropy) in each cluster:

$$x_b^s = \arg\max_{x \in C_b} \mathcal{H}(p(y|x; \theta_g)) \tag{15}$$

Algorithm 1: FSSAL

Require: Private models $\{\omega_k\}_{k=1}^K$; Local models $\{\theta_k\}_{k=1}^K$; Global model θ_g; Labeled data $\{D_k^l\}_{k=1}^K$; Unlabeled data $\{D_k^u\}_{k=1}^K$; Active learning budget B per round and total budget TB; Class threshold base τ; EMA weight α_{ema}; Active learning rounds T_{al}.
Ensure: Converged global model θ_g.

Active Learning Cycle:
while $B \leq TB$ **do**
  for active learning round $t = 1$ to T_{al} **do**
    Federated Semi-Supervised Training
  end for
  for each client k in parallel do **do**
    Perform Algorithm 2 with budget B
  end for
  $B \leftarrow TB - B$
end while
while θ_g not converged **do**
  Federated Semi-Supervised Training
end while
Federated Semi-Supervised Training:
for each client k in parallel **do**
  Compute labeled data loss $\mathcal{L}_k^{\text{private}}$ on ω_k by Eq.2 and $\mathcal{L}_k^{\text{sup}}$ onθ_k by Eq.6.
  Generate pseudo-labels by Eq.4
  $D_k^p \leftarrow \{(x_i^u, y_i^u) | \max(p_i^u) > \tau(y_i^u)\}$
  Compute unlabeled data loss L_k^{unsup} on θ_k by Eq.7 and total loss L_k^{local} by Eq.5
  Update ω_k with $\mathcal{L}_k^{\text{private}}$
  Update θ_k with L_k^{local}
end for
Server aggregates $\theta_g \leftarrow \sum_{k=1}^K \frac{|D_k^l|}{|D^l|}\theta_k$
Update $\theta_k \leftarrow \theta_g$
EMA update ω_k by Eq.3

Algorithm 2: Client k Active Learning Procedure

Require: Private model ω_k, global model θ_g, unlabeled data D_k^u, pseudo-labeled data D_k^p, annotation budget B
Ensure: Updated labeled data D_k^l, unlabeled data D_k^u

$D_k^{al} \leftarrow D_k^u - D_k^p$ {Unreliable unlabeled samples}
for each $x_i^{al} \in D_k^{al}$ **do**
  $\mathcal{F}_i \leftarrow$ **Diversity Extraction** by Eq.13.
end for
$\{C_1, \ldots, C_B\} \leftarrow$ **Cluster Formation**$(\{\mathcal{F}_i\}_{i=1}^{|D_k^{al}|})$ by Eq.13.
for each cluster C_b **do**
  $x_b^s \leftarrow$ **Uncertainty Sample Selection** by Eq.15.
  $D_k^l \leftarrow D_k^l \cup \{x_b^s\}$ {Add to labeled data}
  $D_k^u \leftarrow D_k^u - \{x_b^s\}$ {Remove from unlabeled data}
end for
Reset: $D_k^{al} \leftarrow \emptyset$
Recalculate class thresholds due to distribution shift
return Updated D_k^l, D_k^u

Selected samples are annotated and added to D_k^l, while removed from D_k^u. D_k^{al} is reset. After active learning, thresholds are recalculated to reflect the updated class distribution.

Our method improves pseudo-label reliability via private modeling, introduces class-aware confidence control, and enhances sample utility through active learning. The full pipeline is outlined in Algorithm 1.

Experiments

Experimental Setup

Datsets. We evaluate our method on four medical image classification tasks, i.e., skin lesion diagnosis for dermoscopy images using HAM10000 dataset (Tschandl, Rosendahl, and Kittler 2018), intracranial hemorrhage (ICH) diagnosis for brain CT slice using RSNA ICH dataset (RSNA 2019), dabetic retinopathy classification using Diabetic Retinopathy dataset (Dane and Karthik 2019; Porwal et al. 2018; Decencière et al. 2014) and glaucoma fundus image classification using Fundus dataset (EyePACS 2015). We follow the preprocessing of datasets in FedIRM (Liu et al. 2021), that resizes the images into 240×240, randomly crops a 224×224 region, and normalizes before input to the network. We also employ 70% for training, 10% for validation, and 20% for testing. We use the Dirichlet distribution to simulate Non-IID data partition in federated systems. Table 1 presents statistical information of datasets, where D_T denotes total training samples, D_t denotes total testing samples, C denotes the number of categories in the datasets, K denotes clients in the federated system.

Dataset	HAM10000	RSNA ICH	Diabetic	Fundus
D_T	7010	17500	4150	4200
D_t	2003	5000	1180	1200
C	7	5	5	2
K	10	10	5	5

Table 1: Statistical information of datasets.

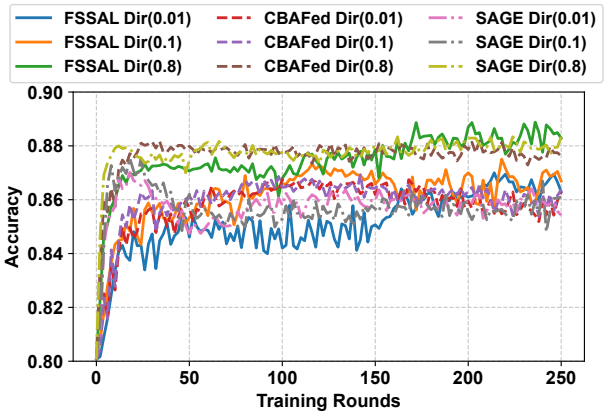

Figure 2: The Accuracy of different methods under varying Non-IID degrees on the Diabetic dataset.

Method	Skin Lesion			RSNA ICH			Diabetic			Fundus		
	Acc(%)	Recall(%)	F1(%)	Acc(%)	Recall(%)	F1(%)	Acc(%)	Recall(%)	F1(%)	Acc(%)	Recall(%)	F1(%)
FedAvg	93.03	42.32	44.93	83.33	42.91	43.26	86.16	48.10	48.29	83.95	72.13	72.20
RSCFed	93.69	54.11	57.59	83.90	43.06	44.28	87.25	49.22	50.05	84.45	73.01	72.16
FedLoKe	93.65	52.45	55.40	84.42	48.69	47.86	87.55	49.61	50.89	85.20	72.06	71.56
CBAFed	93.42	40.48	44.29	84.56	46.95	46.48	87.89	50.63	51.64	85.45	73.88	74.05
SAGE	93.37	45.36	51.58	84.20	43.60	43.48	88.11	49.55	51.35	85.95	71.18	73.23
Ours	**94.47**	**64.05**	**61.91**	**86.49**	**56.19**	**55.38**	**88.61**	**55.13**	**54.62**	**87.00**	**76.58**	**77.67**

Table 2: Performance comparison under Dir(0.8) and $\gamma = 0.2$ across four medical datasets.

Method	Skin Lesion			RSNA ICH			Diabetic			Fundus		
	Acc(%)	Recall(%)	F1(%)	Acc(%)	Recall(%)	F1(%)	Acc(%)	Recall(%)	F1(%)	Acc(%)	Recall(%)	F1(%)
FedAvg	92.36	41.54	44.03	82.67	41.21	42.64	85.65	46.35	46.87	82.75	71.44	71.20
RSCFed	93.08	52.94	54.72	82.79	41.65	43.28	86.73	46.67	47.32	83.79	71.23	70.88
FedLoKe	93.37	48.20	50.32	83.91	47.68	45.96	87.16	47.43	48.90	84.87	70.95	71.06
CBAFed	93.02	40.08	43.57	83.74	45.95	44.98	86.91	48.60	49.75	84.17	72.34	72.15
SAGE	92.75	42.28	46.46	82.90	40.66	41.48	87.37	46.98	48.12	85.33	70.11	71.35
Ours	**94.08**	**60.87**	**59.54**	**86.11**	**54.11**	**54.08**	**88.01**	**52.81**	**52.34**	**86.20**	**75.29**	**76.39**

Table 3: Performance comparison under Dir(0.8) and $\gamma = 0.15$ across four medical datasets.

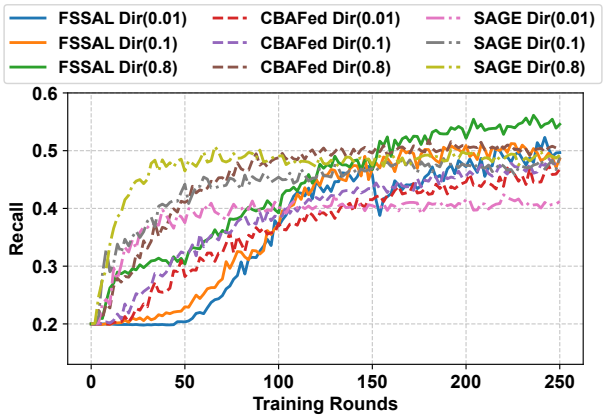

Figure 3: The Recall of different methods under varying Non-IID degrees on the Diabetic dataset.

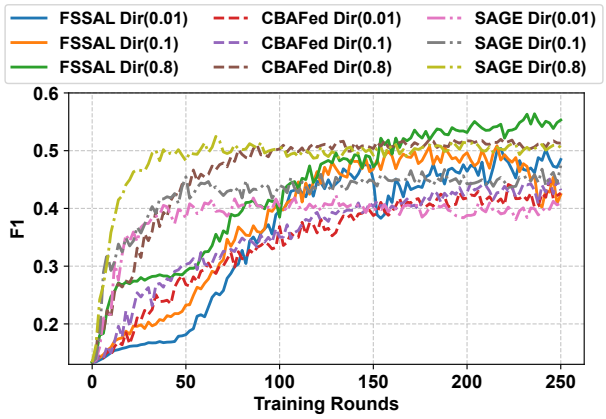

Figure 4: The F1 Score of different methods under varying Non-IID degrees on the Diabetic dataset.

Implementation Details. We implemented FSSAL using PyTorch 1.11.0 and trained it on RTX 3090 GPU. We employ pre-trained ResNet-50 (He et al. 2016) as the backbone network for medical image classification tasks. In all experiments, the standard Stochastic Gradient Descent (SGD) (Bottou 2010) optimizer is used to train the model. The batch size is 12 for the HAM10000 dataset, RSNA ICH dataset and 16 for the Diabetic dataset, Fundus dataset. The learning rate is 0.001 and $Dir(\alpha) = 0.8$. We enpirically set threshold base $\tau = 0.85$, model fusion ratio $\alpha_{ema} = 0.95$.

Comparisons with State-of-the-Arts

Compare Method. We compare our approach with the state-of-the-art FSSL methods including RSCFed (Liang et al. 2022), FedLoKe (Zhang et al. 2023), CBAFed (Li, Li, and Wang 2023) and SAGE (Liu et al. 2025). For all methods, we let the proportion of labeled data $\gamma = 0.2$. We also conduct a comparative analysis against FedAvg (McMahan et al. 2017), which serves as the lower bound training only by the labeled data.

Evaluation Metrics. We use Accuracy (ACC), Recall, and F1-score to evaluate model performance. ACC measures overall prediction correctness. Recall quantifies sensitivity

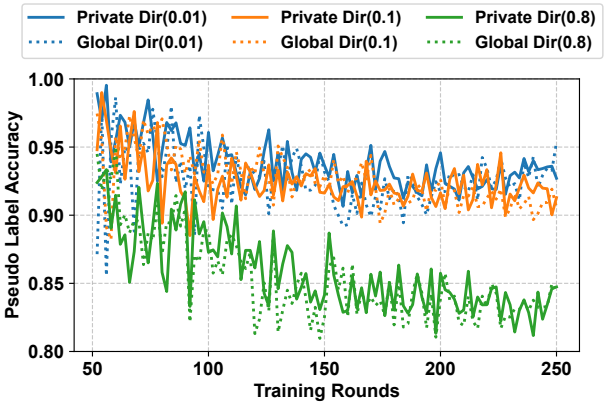

Figure 5: Pseudo label accuracy of Private Model and Global Model under varying Non-IID degrees on the Diabetic dataset.

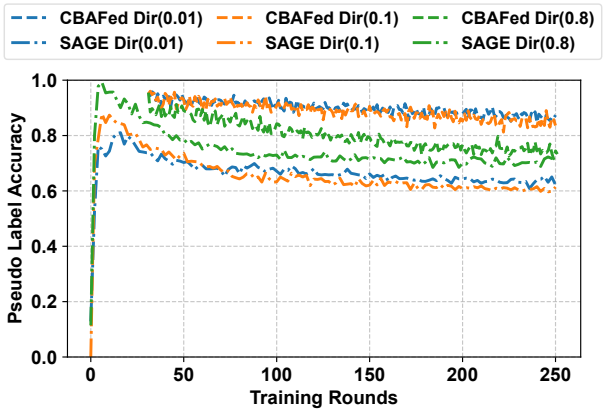

Figure 6: Pseudo label accuracy of CBAFed Model and SAGE Model under varying Non-IID degrees on the Diabetic dataset.

to positive cases, crucial for detecting medical anomalies in imbalanced datasets. F1-score balances precision and recall, offering a robust metric for class-imbalanced scenarios.

Quantitative Comparisons. Table 2 presents outstanding experimental results comparing our approach with other methods on four medical datasets. For the FSSAL method proposed in this paper, at the beginning of training, the proportion of labeled data $\gamma = 0.1$. Then, two rounds of active learning are performed, each time selecting 5% from the training set to add to the labeled data. Finally, the model is trained in a federated semi-supervised manner with labeled data proportion $\gamma = 0.2$, and the performance of the global model at convergence is measured. Our method outperforms the baseline on all three metrics, particularly most notably on Recall and F1.

Performance Evaluation

Influence of the Proportion of Labeled Data. To compare the effects under different labeled data ratio experiments, the

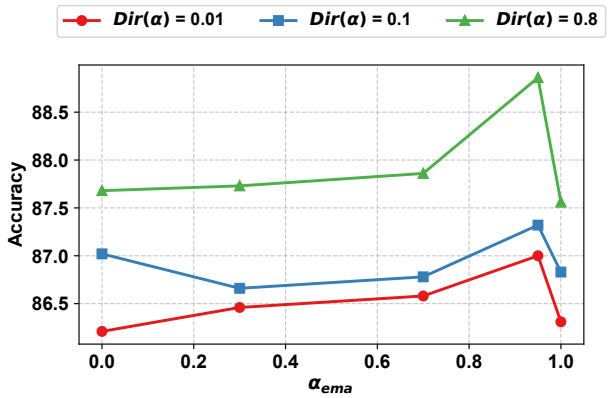

Figure 7: The Accuracy of FSSAL with different EMA parameters α_{ema} on Diabetic under different non-IID degrees.

proportion of labeled data in Table 3 is set to 0.15, with only one round of active learning performed, while keeping other conditions unchanged. When the amount of labeled data decreases, our method still maintains strong performance compared to other approaches, making it more suitable for scenarios with limited annotated data.

Non-IID Degree. We evaluate methods under three different Non-IID distributions, i.e., $\alpha \in 0.01, 0.1, 0.8$, which all have label ratios $\gamma = 0.2$. To delve deeper into the impact of Non-IID degrees on different methods, we conducted comparisons along two dimensions: model performance and pseudo-label accuracy. As shown in Figures 2, 3, and 4, our method exhibits significant improvements across all three metrics after 150 rounds, when active learning completes with a 0.2 labeled data ratio, demonstrating consistent advantages thereafter. Figures 5 and 6 show that the model trained with our method achieves the highest pseudo-label accuracy, which also explains why our method performs so well.

EMA Weight. We also explore how the EMA weight α_{ema} influences the performance of FSSAL. These results are shown in Figures 7 and 8. The initial rise followed by a decline demonstrates the effectiveness of model fusion.

Active Learning Selector. Our approach employs both private and global models during active learning. The private model performs clustering, while the global model conducts query selection based on clustering outcomes, promoting globally balanced sample selection. To validate this, we compared FSSAL's performance at convergence against variants using only the global or private model for class-balanced selection (Dir(α) = 0.01, Diabetic dataset). As shown in Table 4, our dual-model selector achieves superior performance across all metrics, validating its effectiveness in active learning sample selection.

Ablation Studies

We present a comprehensive ablation analysis in Table 5, combining both single-component and pairwise-component evaluations. The results highlight the individual and joint

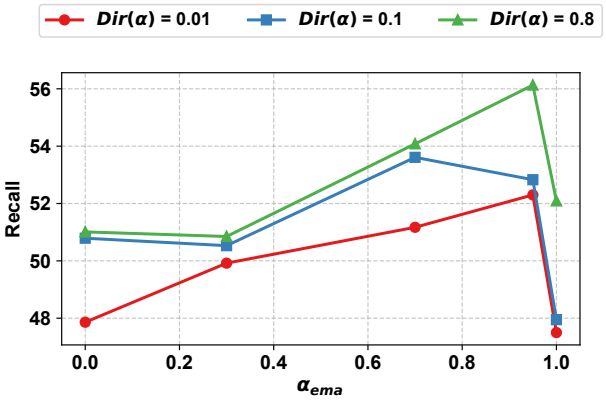

Figure 8: The Recall of FSSAL with different EMA parameters α_{ema} on Diabetic under different non-IID degrees.

Selector	Metrics		
	Acc (%)	Recall (%)	F1 (%)
Private Model	86.62	51.40	49.20
Global Model	86.72	48.67	46.28
Both	**87.00**	**52.30**	**50.40**

Table 4: Performance comparison of different active learning selectors on the Diabetic dataset.

contributions of the three core modules in FSSAL: Private Model Pseudo-label Prediction (PMPP), Labeled Data-Aware Thresholding (LAT), and Active Class-Balanced Sampling (ABS).

When evaluated independently, LAT demonstrates the most substantial standalone impact, while PMPP and ABS also contribute positively—PMPP delivering strong supervised guidance and ABS providing diverse label selection.

Pairwise combinations further amplify performance. Notably, LAT combined with PMPP yields stronger gains than either alone, confirming LAT's role as a foundational module. The full FSSAL framework, integrating all three components, consistently outperforms all partial variants, with improvements ranging from 0.37% to 9.93% across key metrics, including Accuracy, Recall, and F1-score.

Conclusion

This paper presents FSSAL, a federated semi-supervised learning framework enhanced by active learning to address the challenges of limited labeled data, data heterogeneity, and pseudo-label noise in medical image classification. Unlike existing methods that rely on static supervision, FSSAL dynamically updates the labeled set through uncertainty- and diversity-driven sample selection, while leveraging client-specific pseudo-labeling and class-aware thresholding to ensure label reliability and class balance. Extensive experiments on four public medical datasets under various Non-IID settings show that FSSAL consistently outperforms representative FSSL baselines in terms of accuracy, recall, and F1-score. Ablation studies further confirm the effectiveness of each proposed component. Overall, FSSAL offers a practical and scalable solution for privacy-preserving collaborative learning in real-world medical environments. In future work, we plan to extend the framework to more complex tasks such as medical image segmentation and multi-modal federated learning, as well as explore communication-efficient variants to support deployment in resource-constrained settings.

PMPP	LAT	ABS	Acc(%)	Recall(%)	F1(%)
✓	×	×	93.40	58.39	54.80
×	✓	×	93.80	57.89	57.28
×	×	✓	93.71	54.12	53.96
✓	✓	×	94.10	61.55	59.53
✓	×	✓	93.50	58.46	57.64
×	✓	✓	93.85	62.12	60.17
✓	✓	✓	**94.47**	**64.05**	**61.91**

Table 5: Ablation studies on the effectiveness of one component.

Acknowledgments

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No.LTGG24F020007; the National Natural Science Foundation of China (NSFC) under Grant No.62302133; the Key Research and Development Program of Zhejiang Province under Grant (2024C01026, 2025C01115).

References

Bottou, L. 2010. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*, 177–186. Springer.

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.

Dane, S.; and Karthik, M. 2019. APTOS 2019 Blindness Detection. Kaggle Competition.

Decencière, E.; Zhang, X.; Cazuguel, G.; Lay, B.; Cochener, B.; Trone, C.; Gain, P.; Ordóñez-Varela, J.-R.; Massin, P.; Erginay, A.; et al. 2014. Feedback on a Publicly Distributed Image Database: The Messidor Database. *Image Analysis & Stereology*, 33: 231–234.

Du, J.; Li, W.; Liu, P.; Vong, C.-M.; You, Y.; Lei, B.; and Wang, T. 2024. Federated Learning Using Model Projection for Multi-Center Disease Diagnosis with Non-IID Data. *Neural Networks*, 171: 1–13.

EyePACS. 2015. EyePACS Diabetic Retinopathy Dataset. Kaggle Dataset.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

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.

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 (CVPR)*, 16292–16301.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2019. Federated learning: challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 36(3): 50–60.

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 (CVPR)*, 10144–10153. IEEE.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J. A.; van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88.

Liu, Q.; Yang, H.; Dou, Q.; and Heng, P.-A. 2021. Federated Semi-Supervised Medical Image Classification via Inter-Client Relation Matching. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 325–335.

Liu, Y.; Shang, X.; Zhang, Y.; Lu, Y.; Gong, C.; Xue, J.-H.; and Wang, H. 2025. Mind the Gap: Confidence Discrepancy Can Guide Federated Semi-Supervised Learning Across Pseudo-Mismatch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ArXiv:2503.13227.

Liu, Y.; et al. 2024. FedCD: Federated Semi-Supervised Learning with Class Awareness Balance via Dual Teachers. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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.

Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; and Meriaudeau, F. 2018. Indian Diabetic Retinopathy Image Dataset (IDRiD). *IEEE Dataport*.

Ren, M.; Zhang, H.; and Liu, W. 2021. Active Learning for Medical Image Classification: Combining Uncertainty Sampling with Deep Convolutional Networks. *IEEE Transactions on Medical Imaging*, 40(5): 1213–1224.

RSNA. 2019. RSNA Intracranial Hemorrhage Detection Challenge. https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/. Accessed: 2023-01-01.

Settles, B. 2009. Active Learning Literature Survey. Technical Report 1648, University of Wisconsin-Madison.

Sheller, M. J.; Edwards, B.; Reina, G. A.; Martin, J.; Pati, S.; Kotrotsou, A.; et al. 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1): 1–12.

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: 180161.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; et al. 2023. FreeMatch: Self-Adaptive Thresholding for Semi-Supervised Learning. In *International Conference on Learning Representations (ICLR)*. OpenReview.net.

Wu, N.; et al. 2022. FedIIC: Towards Robust Federated Learning for Class-Imbalanced Medical Image Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; and Le, Q. V. 2020. Unsupervised Data Augmentation for Consistency Training. In *Advances in Neural Information Processing Systems*, volume 33, 6256–6268. Curran Associates, Inc.

Yue, G.; Wei, P.; Zhou, T.; Song, Y.; Zhao, C.; Wang, T.; and Lei, B. 2023. Specificity-Aware Federated Learning with Dynamic Feature Fusion Network for Imbalanced Medical Image Classification. *IEEE Journal of Biomedical and Health Informatics*.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; Shinozaki, T.; and Liu, K. 2021a. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, C.; Wu, F.; Yi, J.; Xu, D.; Yu, Y.; Wang, J.; Wang, Y.; Xu, T.; Xie, X.; and Chen, E. 2023. Non-IID Always Bad? Semi-Supervised Heterogeneous Federated Learning with Local Knowledge Enhancement. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*.

Zhang, X.; Li, Z.; and Li, X. 2022. Generative Adversarial Network-Based Active Learning for Medical Image Classification. *Journal of Medical Imaging*, 9(1): 013503.

Zhang, Z.; Ma, S.; Nie, J.; Wu, Y.; Yan, Q.; Xu, X.; and Niyato, D. 2021b. Semi-Supervised Federated Learning with Non-IID Data: Algorithm and System Design. In *IEEE 23rd International Conference on High Performance Computing and Communications*, 157–164.