Temporal Model-Based Federated Active Medical Image Classification

Yunlu Yan[1], Chun-Mei Feng[2], Yuexiang Li[3], Jinheng Xie[4], Jun Chen[5], Mohamed Elhoseiny[5], Ming Hu[6], Kaishun Wu[1], Lei Zhu[1,7] (✉)

[1] The Hong Kong University of Science and Technology (Guangzhou), China
[2] University College Dublin, Ireland
[3] Guangxi Medical University, China
[4] National University of Singapore, Singapore
[5] King Abdullah University of Science and Technology, Saudi Arabia
[6] Monash University, Australia
[7] The Hong Kong University of Science and Technology, China
leizhu@ust.hk
https://github.com/IAMJackYan/TM-FAL

Abstract. Traditional federated learning relies on fully labeled datasets in each medical institution, which is impractical in real-world clinical scenarios. Federated Active Learning (FAL) addresses this by selecting a few informative samples for labeling, but it faces challenges such as domain shift across institutions. Besides, existing FAL methods rely on single-round model knowledge to estimate prediction-level uncertainty, ignoring uncertainty from features and model evolution during training. In this work, we propose **TM-FAL**, a novel framework for federated active medical image classification under domain shift. TM-FAL proposes a new uncertainty by integrating feature differences and prediction confidence from temporal local and global models to capture both local-global differences and the inherent complexity of images. Additionally, we use the prediction of the global model as pseudo labels to group images to mitigate class imbalance caused by uncertainty-based selection. Experiments on two medical image classification datasets demonstrate that TM-FAL outperforms various state-of-the-art methods.

Keywords: Federated learning · Active learning · Medical image classification.

1 Introduction

Federated learning has shown success in a wide range of healthcare applications [8,24,29,6,30,27] by enabling multiple medical institutions to collaboratively train a global model using their datasets in a privacy-preserving manner. However, it typically requires all institutions to have a fully labeled dataset, which does not reflect real-world clinical scenarios. The high cost of labeling medical data makes it difficult for physicians to label all available data, leaving each institution with only a small amount of labeled data and a large amount

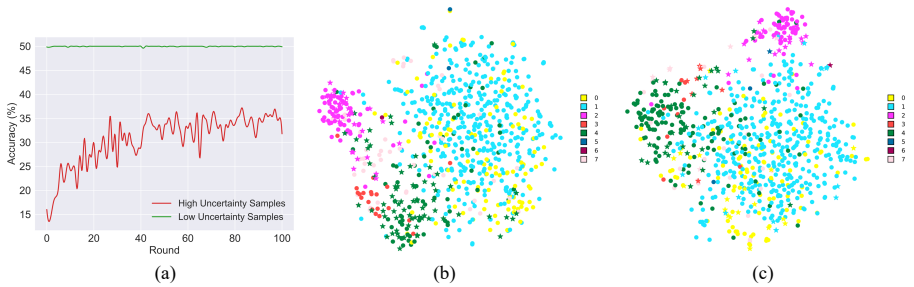

Fig. 1: (a) Accuracy (%) of low- and high-uncertainty samples versus communication rounds. The T-SNE [15] visualization of (b) TM-FAL without the pseudo-labeling-based grouping and (c) TM-FAL, where $\star$ denotes the selected samples, and different classes (0-7) are represented by different colors. Without the pseudo-labeling-based grouping, data selection bias towards two difficult-to-classify categories (2 and 4).

of unlabeled data. Active Learning [17,22,23] (AL) provides an effective solution to this challenge by letting the model, trained on the labeled dataset, select a small set of representative data from the unlabled dataset for labeling, thereby minimizing annotation costs while achieving performance comparable to that of full data labeling. Thus, the key lies in designing an effective query strategy to select the most informative data.

Unlike AL in centralized scenarios, Federated Active Learning [12,26,1] (FAL) poses new challenges due to involving multiple institutions. In federated medical scenarios, data from different medical institutions, such as medical images, are typically acquired using different devices, resulting in ***domain shift*** across datasets [14,31,11]. As a result, we cannot treat each institution as a subtask of FAL, as it would exacerbate the learning biases of local models. When selecting data, it is crucial to not only consider its contribution to local model optimization but also to focus on its potential to improve the generalization of the global model across diverse institutions.

To address this, recent state-of-the-art FAL methods [4,5] select data by simultaneously leveraging both local and global models to estimate uncertainty. However, we identified two key limitations in these methods. ***First***, they rely solely on the model's prediction (classifier output) to estimate uncertainty. While this captures the difficulty of the data, it does not effectively quantify the local and global differences. In FAL, it is crucial to select data where local and global models exhibit significant differences, as these samples are particularly valuable for improving the generalization of the global model across institutions. As shown in previous studies [14,31,28], domain shift leads to ***feature drift*** across different institutions. Consequently, the differences between local and global models are more pronounced in feature representation (encoder output). ***Second***, these methods only use knowledge from a single round of local and global models. Given the randomness of Stochastic Gradient Descent (SGD), relying on knowl-

edge from only one-time point (short-term) may fail to effectively capture the inherent complexity of data, which is a long-term attribute.

Based on the above insights, we propose Temporal Model-based Federated Active Learning, termed **TM-FAL**, for medical image classification. Specifically, we propose a novel local-global temporal uncertainty-based sampling strategy, inspired by the cognitive principle of the model: during the learning process, models typically classify simpler samples in the early stages and progressively handle more difficult samples in later stages [2]. As a result, the model's output varies across different learning phases, particularly for more challenging samples (see Fig. 1 (a)). By leveraging this, TM-FAL utilizes both the feature differences and prediction confidence of temporal local and global models to capture the local-global differences as well as the inherent complexity of the data. However, we observed that relying solely on this uncertainty-based strategy can lead to a bias in data selection towards more difficult-to-classify categories (see Fig. 1 (b)), resulting in class imbalance. To address this, we introduce a pseudo-labeling-based grouping strategy that maintains class diversity (see Fig. 1 (c)) by using the global model's output as pseudo labels. Experimental results show that TM-FAL significantly outperforms several state-of-the-art methods on two medical image classification datasets.

2 Preliminary

Problem Statement. Suppose there are K hospitals, communicating through a trusted central server. Each hospital $k \in [K]$ contains a small labeled dataset $\boldsymbol{L}_k = \{\boldsymbol{X}_i, \boldsymbol{Y}_i\}_{i=1}^{n_k^L}$ and a large unlabeled dataset $\boldsymbol{U}_k = \{\boldsymbol{X}_i\}_{i=1}^{n_k^U}$, where $\boldsymbol{X}$ is the image, $\boldsymbol{Y} \in [C]$ represents the corresponding label, and C is the total number of classes. Our goal is to train a classification network: $f = g \circ h$, where $g : \boldsymbol{X} \to \boldsymbol{Z}$ is a feature encoder that extracts the latent features $\boldsymbol{Z}$ from the image, $h : \boldsymbol{Z} \to \boldsymbol{Y}$ is a classifier that makes predictions, and $\boldsymbol{w} = \boldsymbol{w}_g \circ \boldsymbol{w}_h$ represents their parameters.

Federated Active Learning. FAL consists of two phases: the training phase and the data selection phase. First, we use the initial labeled datasets $\{\boldsymbol{L}_k^0\}_{k=1}^K$, where $\boldsymbol{L}_k^0 = \boldsymbol{L}_k$, to train the network, and the training phase follows the standard FL framework [16]. At each communication round $t \in [T]$, the local models are optimized by minimizing each local empirical risk ℓ_k, and the global model is updated by averaging the local updates as:

$$\ell = \gamma_k \ell_k(\boldsymbol{w}_k^t), \qquad \text{and} \quad \boldsymbol{w}_G^t = \sum_{k=1}^{K} \gamma_k \boldsymbol{w}_k^t, \tag{1}$$

where $\gamma_k = \frac{n_k^L}{\sum_{i=1}^K n_i^L}$, $\boldsymbol{w}_G^t$ and $\boldsymbol{w}_k^t$ represent the parameters of the global and local models, respectively. After the training phase, the well-trained models are used to select B_k images from the unlabeled dataset and annotate them, thus constructing a labeled query dataset $\boldsymbol{Q}_k^e = \{\boldsymbol{X}_i, \boldsymbol{Y}_i\}_{i=1}^{B_k}$, where $B_k \ll n_k^U$. The

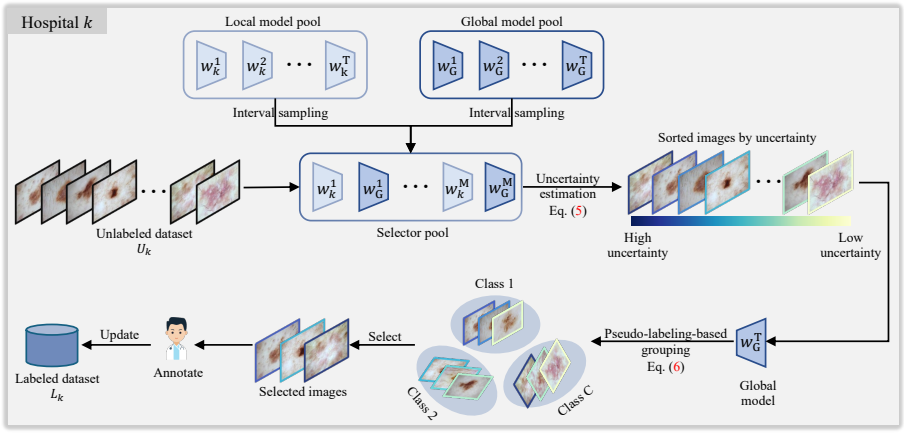

Fig. 2: **Illustration of TM-FAL**. For each hospital k, TM-FAL constructs a selector pool by selecting models from the local and global model pools using interval sampling, and estimates the uncertainty by leveraging the knowledge from all selectors. To address the class imbalance issue introduced by the uncertainty-based strategy, TM-FAL further groups the images using the predictions of the global model as pseudo-labels.

labeled dataset is then updated to $\boldsymbol{L}_k^e = \boldsymbol{L}_k^{e-1} \cup \boldsymbol{Q}_k^e$, and the unlabeled dataset is updated to $\boldsymbol{U}_k^e = \boldsymbol{U}_k^{e-1} \setminus \boldsymbol{Q}_k^e$. Next, the updated labeled datasets of all institutions are used to initiate a new training phase. The above FAL process can be repeated multiple epochs ($e \in [E]$), gradually improving the model's performance on $\boldsymbol{L}_k^e$, bringing it closer to the performance achieved with the fully labeled set $\boldsymbol{L}_k \cup \boldsymbol{U}_k$.

3 Methodology

To tackle the FAL challenge under domain shift, we propose a novel method, termed TM-FAL. Fig. 2 illustrates the data selection phase of our method, which consists of two key modules: i) local-global temporal uncertainty-based sampling and ii) pseudo-labeling-based grouping. We omit the training phase and assume it has already been completed on the initial labeled datasets $\{\boldsymbol{L}_k\}_{k=1}^K$. Each hospital k maintains a local model pool $\{\boldsymbol{w}_k^1, \boldsymbol{w}_k^2, \ldots, \boldsymbol{w}_k^T\}$ and a global model pool $\{\boldsymbol{w}_G^1, \boldsymbol{w}_G^2, \ldots, \boldsymbol{w}_G^T\}$, both trained using Eq. (1). For simplicity, we only show the process of a single FAL epoch, but this can be easily extended to multiple FAL epochs. We present the whole algorithm in Alg. 1.

3.1 Local-Global Temporal Uncertainty-Based Sampling

We introduce local-global temporal uncertainty as a novel metric for data selection. Unlike existing methods that rely solely on local and global models from

a single round, our approach leverages the knowledge from models at multiple different rounds. Besides, we estimate the uncertainty of images in both the feature encoding and prediction processes to simultaneously capture the local-global differences and inherent complexity. The detailed process is as follows.

Selector Pool. Since models from adjacent rounds tend to exhibit high similarity, directly utilizing local and global models from all rounds would diminish the temporal differences and increase computational overhead. We observed that temporal differences are more pronounced in the early rounds, as the models have not yet converged, and decrease in the later rounds as the models converge. In this way, for each hospital k, we select the first $2N$ local and global models at intervals of S training rounds from the local and global model pools to construct a selector pool as:

$$\boldsymbol{\Theta}_k = \{\boldsymbol{w}_k^1, \boldsymbol{w}_G^1, \boldsymbol{w}_k^{S+1}, \boldsymbol{w}_G^{S+1}, \ldots, \boldsymbol{w}_k^M, \boldsymbol{w}_G^M\}, \tag{2}$$

where $M = (N-1)S+1 \leq T$. The selector pool effectively reduces computational overhead and enhances temporal differences.

Uncertainty Estimation. Following, we can use the constructed selector pool to estimate the uncertainty of images. Given an image $\boldsymbol{X} \in \mathbb{R}^{r\times H\times W}$, with size $H \times W$ and r channels, we first estimate the uncertainty of the feature encoding process by computing the variance of features across all selectors, as follows:

$$\delta_g = \frac{1}{D}\sum_{d=1}^{D} Var(\{\boldsymbol{Z}^i\}_{i=1}^{2N})_d \in \mathbb{R}^1, \qquad \text{and} \quad \boldsymbol{Z}^i = g(\boldsymbol{\Theta}_{k,g}^i, \boldsymbol{X}) \in \mathbb{R}^{1\times D}, \tag{3}$$

where D is the dimension of the feature and $\boldsymbol{\Theta}_{k,g}^i$ is the parameters of i-th selector's feature encoder. δ_g effectively captures the local and global feature differences caused by domain shift. Next, we further estimate the uncertainty of the prediction process as:

$$\delta_h = \frac{1}{\rho}, \qquad \text{and} \quad \rho = \max(\frac{1}{2N}\sum_{i=1}^{2N} f(\boldsymbol{\Theta}_k^i, \boldsymbol{X})) \in [0, 1]. \tag{4}$$

where $\boldsymbol{\Theta}_k^i$ denotes the parameters of i-th selector, and ρ is the confidence score of the prediction by the ensemble of all selectors. For images with higher confidence, their prediction process exhibits lower uncertainty. δ_h reflects the long-term prediction uncertainty of both local and global models, effectively capturing the inherent complexity of images. Our proposed uncertainty is composed of the above two uncertainties as:

$$\delta = \delta_g\delta_h = \frac{\frac{1}{D}\sum_{d=1}^{D} Var(\{\boldsymbol{Z}^i\}_{i=1}^{2N})_d}{\max(\frac{1}{2N}\sum_{i=1}^{2N} f(\boldsymbol{\Theta}_k^i, \boldsymbol{X}))}. \tag{5}$$

Finally, we can estimate the uncertainty of all images $\{\delta_i\}_{i=1}^{n_k^U}$ based on Eq. (5) and select the top B_k images with the highest uncertainty for annotation as the query dataset.

Algorithm 1: Data selection of TM-FAL for hospital k

Input: Unlabeled dataset $\boldsymbol{U}_k$, local model pool $\{\boldsymbol{w}_k^1, \boldsymbol{w}_k^2, \ldots, \boldsymbol{w}_k^T\}$ and global model pool $\{\boldsymbol{w}_G^1, \boldsymbol{w}_G^2, \ldots, \boldsymbol{w}_G^T\}$ trained on initial labeled datasets $\{\boldsymbol{L}_k\}_{k=1}^K$, number of classes C, hyper-parameters S, N, and B_k

Output: Query dataset $\boldsymbol{Q}_k$

$M \leftarrow (N-1)S+1$

$\boldsymbol{\Theta}_k \leftarrow \{\boldsymbol{w}_k^1, \boldsymbol{w}_G^1, \boldsymbol{w}_k^{S+1}, \boldsymbol{w}_G^{S+1}, \ldots, \boldsymbol{w}_k^M, \boldsymbol{w}_G^M\}$

for *Image* $\boldsymbol{X_i} \in \boldsymbol{U}_k$ **do**

  $\delta_i \leftarrow$ Eq. (5) `// Estimate uncertainty for each image`

  $\hat{\boldsymbol{Y}}_i \leftarrow argmax(f(\boldsymbol{w}_G^T, \boldsymbol{X}_i))$

  Put $(\boldsymbol{X}_i, \delta_i)$ into group $\boldsymbol{\Gamma}_{\hat{\boldsymbol{Y}}_i}$

end

$\boldsymbol{Q}_k \leftarrow$ Select B_k images sequentially from $\{\boldsymbol{\Gamma}_c\}_{c=1}^C$ with higher uncertainty

Return $\boldsymbol{Q}_k$

3.2 Pseudo-Labeling-Based Grouping

Since images from difficult-to-classify classes typically exhibit higher inherent complexity than those from easier-to-classify classes, the above methods may introduce a bias towards these difficult classes, potentially exacerbating class imbalance. An effective selection strategy should consider both the uncertainty and diversity of images [13]. To address this, we further utilize the predictions of the global model as pseudo-labels to group images. Given that the global model has been trained on data from diverse hospitals, it demonstrates more generalized performance compared to the local model. The image group for each class $c \in [C]$ can be written as:

$$\boldsymbol{\Gamma}_c = \{(\boldsymbol{X}_i, \delta_i) | \hat{\boldsymbol{Y}}_i = c\}_{i=1}^{n_k^U}, \quad \hat{\boldsymbol{Y}}_i = argmax(f(\boldsymbol{w}_G^T, \boldsymbol{X}_i)). \tag{6}$$

Following, we sequentially sample from these C groups according to the uncertainty, until we have selected B_k images.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluated our approach on two real-world multi-institutional medical image classification benchmarks, *i.e.*, **Fed-ISIC** and **Fed-Camelyon**, both designed for FL scenarios. **Fed-ISIC** [5,21] is a skin lesion classification dataset that contains 8 categories, distributed across four different institutions with {12413, 3954, 3363, 2259} images. **Fed-Camelyon** [5,21,11] is a binary classification dataset for breast cancer histology, distributed across five independent medical institutions. The dataset contains {59436, 34904, 85054, 129838, 146722} image patches. For Fed-ISIC, we use EfficientNet-B0 [20] as the classification network and evaluate using balanced multi-class accuracy. For Fed-Camelyon, we use DenseNet-121 [9] as the classification network and employ accuracy as the evaluation metric. Both EfficientNet-B0 and DenseNet-121 are initialized with pre-trained weights from ImageNet [7].

Table 1: **Quantitative comparison** using balanced multi-class accuracy (%) on Fed-ISIC and accuracy (%) on Fed-Camelyon. We report the mean±std of the results from three independent trials across four FAL epochs (#**E2**→#**E5**). Best results are marked in red.

	Fed-ISIC				Fed-Camelyon			
Method	**#E2**	**#E3**	**#E4**	**#E5**	**#E2**	**#E3**	**#E4**	**#E5**
Random	$61.59_{\pm 1.45}$	$64.90_{\pm 1.53}$	$65.53_{\pm 1.31}$	$64.99_{\pm 1.43}$	$94.82_{\pm 0.30}$	$95.40_{\pm 0.24}$	$96.02_{\pm 0.12}$	$96.34_{\pm 0.07}$
Entropy	$63.21_{\pm 0.59}$	$64.86_{\pm 1.09}$	$66.35_{\pm 0.14}$	$65.57_{\pm 1.92}$	$95.03_{\pm 0.01}$	$96.08_{\pm 0.23}$	$96.52_{\pm 0.20}$	$96.88_{\pm 0.18}$
TOD	$58.10_{\pm 1.95}$	$66.56_{\pm 0.36}$	$66.26_{\pm 1.22}$	$65.51_{\pm 0.75}$	$93.17_{\pm 0.87}$	$95.27_{\pm 0.07}$	$96.07_{\pm 0.09}$	$96.50_{\pm 0.12}$
Gradnorm	$63.23_{\pm 1.25}$	$66.14_{\pm 1.51}$	$67.02_{\pm 1.00}$	$66.52_{\pm 0.75}$	$94.40_{\pm 0.06}$	$94.85_{\pm 0.21}$	$95.64_{\pm 0.04}$	$95.89_{\pm 0.13}$
CoreSet	$62.53_{\pm 1.50}$	$65.91_{\pm 0.78}$	$66.61_{\pm 0.20}$	$66.84_{\pm 0.21}$	$93.90_{\pm 0.25}$	$93.95_{\pm 0.31}$	$94.94_{\pm 0.09}$	$95.85_{\pm 0.13}$
BADGE	$59.45_{\pm 0.67}$	$64.27_{\pm 0.74}$	$66.73_{\pm 0.46}$	$64.71_{\pm 1.07}$	$94.97_{\pm 0.41}$	$95.62_{\pm 0.11}$	$96.25_{\pm 0.12}$	$96.37_{\pm 0.10}$
LoGo	$62.36_{\pm 2.30}$	$66.43_{\pm 0.69}$	$66.12_{\pm 2.64}$	$66.26_{\pm 0.50}$	$94.98_{\pm 0.07}$	$95.60_{\pm 0.15}$	$96.20_{\pm 0.26}$	$96.51_{\pm 0.05}$
KAFAL	$62.34_{\pm 0.30}$	$65.36_{\pm 1.15}$	$66.26_{\pm 1.12}$	$66.24_{\pm 1.31}$	$95.06_{\pm 0.17}$	$96.08_{\pm 0.07}$	$96.76_{\pm 0.11}$	$96.92_{\pm 0.04}$
FEAL	$65.18_{\pm 0.41}$	$67.77_{\pm 1.31}$	$68.41_{\pm 1.01}$	$68.46_{\pm 0.37}$	$95.79_{\pm 0.68}$	$96.54_{\pm 0.40}$	$97.04_{\pm 0.28}$	$97.29_{\pm 0.35}$
- $\mathcal{L}_{aux}$	$64.28_{\pm 1.64}$	$66.69_{\pm 0.95}$	$67.32_{\pm 1.16}$	$67.40_{\pm 0.22}$	$95.24_{\pm 0.03}$	$96.21_{\pm 0.04}$	$96.80_{\pm 0.07}$	$97.26_{\pm 0.06}$
TM-FAL	$67.80_{\pm 0.90}$	$69.57_{\pm 0.85}$	$70.10_{\pm 0.36}$	$70.46_{\pm 1.20}$	$96.02_{\pm 0.30}$	$97.13_{\pm 0.37}$	$97.89_{\pm 0.34}$	$98.13_{\pm 0.21}$

Task Setting. Following [5], for both Fed-ISIC and Fed-Camelyon, we randomly select 500 images from the training dataset of each institution as the initial labeled dataset, with the remaining images serving as the unlabeled dataset. We train the local and global models on the labeled datasets and use them to select images from the unlabeled datasets. The number of selected images, B_k, for each FAL epoch is set to 500. After each FAL epoch, we perform a new training phase using updated label datasets. Next, we evaluate the new global model on the test dataset of each institution and record the average metrics across all institutions. In this way, we can effectively evaluate the effectiveness of FAL.

Baselines. We evaluate our method against nine baselines, categorized into two groups: (1) Centralized AL methods adapted to FL: ❶ **Random** sampling, ❷ **Entropy** [19] ❸ **TOD** [10] ❹ **Gradnorm** [25] ❺ **CoreSet** [18] ❻ **BADGE** [3]; (2) State-of-the-art FAL methods: ❼ **LoGo** [12] ❽ **KAFAL** [4] ❾ **FEAL** [5], all of which leverage the knowledge from both local and global models. For a fair comparison, we enhance the first group (❷-❻) by employing an ensemble of local and global models as the selector, consistent with the FAL approaches. Since we adopted the exact same setup as FEAL [5], we directly report their published results of all baselines.

Implementation Details. All experiments are implemented using PyTorch and conducted on an NVIDIA RTX 4090 GPU with 24GB of memory. Following [5], we perform a total of $E = 5$ FAL epochs, with each training phase consisting of 100 communication rounds. The network is optimized using the Adam optimizer with a learning rate of 5e-4 and a batch size of 16. Weight decay is set to 5e-4 for Fed-ISIC and 1e-5 for Fed-Camelyon. We treat the process of randomly selecting images to form the initial labeled datasets as the first FAL epoch and report the performance in the subsequent four FAL epochs (#**E2** → #**E5**). Additionally, the total selected data of five FAL epochs does not exceed 85% of

the total training dataset. We conduct three independent trials with different random seeds and report the mean and standard deviation (std) of our results.

4.2 Comparison with State-of-the-Arts

Table 1 summarizes the quantitative comparison results across all evaluated methods on both Fed-ISIC and Fed-Camelyon datasets. Our proposed TM-FAL demonstrates consistent performance superiority over all baselines across all FAL epochs on both datasets. Specifically, on Fed-ISIC, TM-FAL yields a significant performance improvement over FEAL, the best baseline, from **65.18**% to **67.80**% (#**E2**). This strongly demonstrates TM-FAL's effectiveness in addressing FAL challenges under domain shift, as it simultaneously captures both local-global differences and inherent complexity of images. It should be noted that the original FEAL implementation incorporates an auxiliary loss function during training. To ensure a fair comparison of core algorithm capabilities, we conducted additional experiments with FEAL's auxiliary loss removed. The results clearly demonstrate a performance degradation in FEAL on both datasets when deprived of this auxiliary component, further validating the superiority of TM-FAL's data selection strategy.

4.3 Analytical Studies

Hyper-parameters Analysis. To deeply investigate the impact of two key hyper-parameters, we evaluate the average performance of TM-FAL over four FAL epochs for different values of $S\{1, 2, 3, 4\}$ and $N\{5, 10, 15, 20\}$, as shown in Fig. 3. By utilizing the sampling interval S, we can increase the temporal differences among the selected models, thereby reducing the total number of models N required. For instance, a configuration with $(S = 2, N = 10)$ achieves comparable performance to that of $(S = 1, N = 20)$. This demonstrates that a larger sampling interval effectively balances temporal differences and computational efficiency. The best performance of 69.48% is achieved when $(S = 3, N = 10)$.

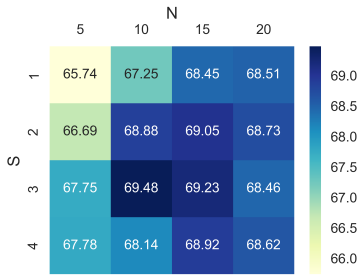

Fig. 3: Results of TM-FAL with different S and N on Fed-ISIC.

Ablation Studies. To gain a deeper understanding of our method, we further explore the two key modules, *i.e.*, Local-Global Temporal Uncertainty-based Sampling (LGTUS) and Pseudo-Labeling-based Grouping (PLG). By this, we built up two baselines: ❶ `M1`: TM-FAL without

Table 2: **Ablation studies** for different modules on Fed-ISIC.

Method	#E2	#E3	#E4	#E5
`M1`	$62.75_{\pm 1.32}$	$65.59_{\pm 0.61}$	$65.63_{\pm 1.20}$	$65.86_{\pm 0.79}$
`M2`	$66.43_{\pm 1.16}$	$68.66_{\pm 0.42}$	$69.21_{\pm 0.48}$	$69.63_{\pm 0.30}$
Ours	$67.80_{\pm 0.90}$	$69.57_{\pm 0.85}$	$70.10_{\pm 0.36}$	$70.46_{\pm 1.20}$

LGTUS module, and ❷ M2: TM-FAL without PLG module. Table 2 shows the ablation experiment results on Fed-ISIC dataset. Compared to TM-FAL, both M1 and M2 show significant performance degradation across different FAL epochs, demonstrating the effectiveness of the two key modules. Additionally, M2 outperforms M1, indicating that the LGTUS module plays a more crucial role in improving the method's performance, as it effectively measures the importance of the data.

5 Conclusion

In this work, we presented an effective solution, TM-FAL, for federated active medical image classification under domain shift by leveraging the knowledge from temporal local and global models to estimate image uncertainty. Additionally, we introduced a pseudo-labeling-based grouping strategy to mitigate class imbalance caused by uncertainty-based sampling. Experimental results on two medical image classification datasets demonstrate that TM-FAL significantly outperforms various state-of-the-art methods.

Acknowledgments. This work is supported by the Guangdong Science and Technology Department (No. 2024ZDZX2004), and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things(No.2023B1212010007).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahn, J.H., Ma, Y., Park, S., You, C.: Federated active learning (f-al): an efficient annotation strategy for federated learning. IEEE Access (2024) 2
2. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International conference on machine learning. pp. 233–242. PMLR (2017) 3
3. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671 (2019) 7
4. Cao, Y.T., Shi, Y., Yu, B., Wang, J., Tao, D.: Knowledge-aware federated active learning with non-iid data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22279–22289 (2023) 2, 7
5. Chen, J., Ma, B., Cui, H., Xia, Y.: Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11439–11449 (2024) 2, 6, 7
6. Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.S., et al.: Federated learning for predicting clinical outcomes in patients with covid-19. Nature medicine **27**(10), 1735–1743 (2021) 1

7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 6
8. Feng, C.M., Yan, Y., Wang, S., Xu, Y., Shao, L., Fu, H.: Specificity-preserving federated learning for mr image reconstruction. IEEE Transactions on Medical Imaging (2022) 1
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017) 6
10. Huang, S., Wang, T., Xiong, H., Huan, J., Dou, D.: Semi-supervised active learning with temporal output discrepancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3447–3456 (2021) 7
11. Jiang, M., Wang, Z., Dou, Q.: Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1087–1095 (2022) 2, 6
12. Kim, S., Bae, S., Song, H., Yun, S.Y.: Re-thinking federated active learning based on inter-class diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3944–3953 (2023) 2, 7
13. Li, J., Chen, P., Yu, S., Liu, S., Jia, J.: Bal: Balancing diversity and novelty for active learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 6
14. Li, X., JIANG, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. In: International Conference on Learning Representations (2021) 2
15. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. JMLR **9**(11) (2008) 2
16. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017) 3
17. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021) 2
18. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017) 7
19. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948) 7
20. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019) 6
21. Ogier du Terrail, J., Ayed, S.S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., et al.: Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. Advances in Neural Information Processing Systems **35**, 5315–5334 (2022) 6
22. Wang, H., Chen, J., Zhang, S., He, Y., Xu, J., Wu, M., He, J., Liao, W., Luo, X.: Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. IEEE Transactions on Medical Imaging (2024) 2
23. Wang, H., Luo, X., Chen, W., Tang, Q., Xin, M., Wang, Q., Zhu, L.: Advancing uwf-slo vessel segmentation with source-free active domain adaptation and a novel multi-center dataset. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 75–85. Springer (2024) 2

24. Wang, J., Jin, Y., Stoyanov, D., Wang, L.: Feddp: Dual personalization in federated medical image segmentation. IEEE Transactions on Medical Imaging (2023) 1
25. Wang, T., Li, X., Yang, P., Hu, G., Zeng, X., Huang, S., Xu, C.Z., Xu, M.: Boosting active learning via improving test performance. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8566–8574 (2022) 7
26. Wu, X., Pei, J., Chen, C., Zhu, Y., Wang, J., Qian, Q., Zhang, J., Sun, Q., Guo, Y.: Federated active learning for multicenter collaborative disease diagnosis. IEEE transactions on medical imaging **42**(7), 2068–2080 (2022) 2
27. Yan, Y., Feng, C.M., Li, Y., Li, P., Goh, R.S.M., Lei, B., Wang, W., Feng, D.D., Zhu, L.: Federated pseudo modality generation for incomplete multi-modal mri reconstruction. IEEE Journal of Biomedical and Health Informatics (2025) 1
28. Yan, Y., Fu, H., Li, Y., Xie, J., Ma, J., Yang, G., Zhu, L.: A simple data augmentation for feature distribution skewed federated learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25749–25758 (2025) 2
29. Yan, Y., Wang, H., Huang, Y., He, N., Zhu, L., Xu, Y., Li, Y., Zheng, Y.: Cross-modal vertical federated learning for mri reconstruction. IEEE Journal of Biomedical and Health Informatics **28**(11), 6384–6394 (2024) 1
30. Yan, Y., Zhu, L., Li, Y., Xu, X., Goh, R.S.M., Liu, Y., Khan, S., Feng, C.M.: A new perspective to boost performance fairness for medical federated learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–23. Springer (2024) 1
31. Zhou, T., Konukoglu, E.: FedFA: Federated feature augmentation. In: International Conference on Learning Representations (2023) 2