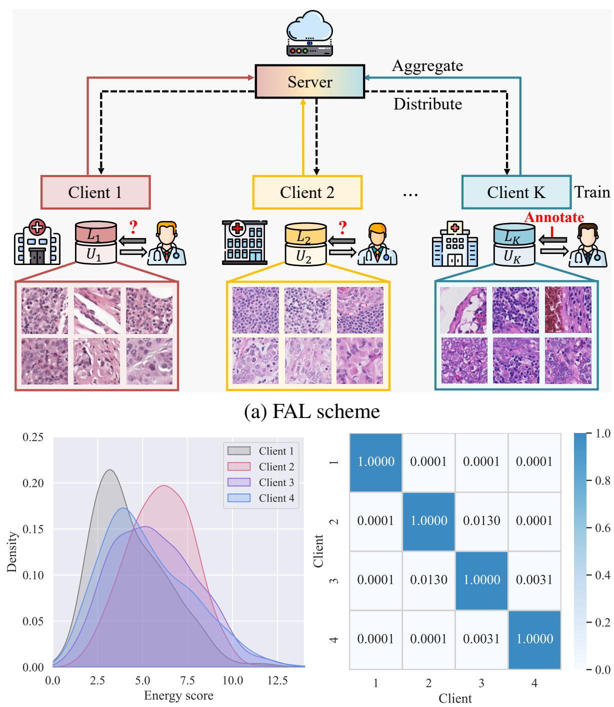
# Think Twice Before Selection: Federated Evidential Active Learning for Medical Image Analysis with Domain Shifts

Jiayi Chen[1*] Benteng Ma[2*] Hengfei Cui[1] Yong Xia[1,3†]
[1] School of Computer Science and Engineering, Northwestern Polytechnical University, China
[2] Hong Kong University of Science and Technology, Hong Kong SAR, China
[3] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, China
jiayichen@mail.nwpu.edu.cn, bentengma@ust.hk, hfcui@nwpu.edu.cn, yxia@nwpu.edu.cn

## Abstract

*Federated learning facilitates the collaborative learning of a global model across multiple distributed medical institutions without centralizing data. Nevertheless, the expensive cost of annotation on local clients remains an obstacle to effectively utilizing local data. To mitigate this issue, federated active learning methods suggest leveraging local and global model predictions to select a relatively small amount of informative local data for annotation. However, existing methods mainly focus on all local data sampled from the same domain, making them unreliable in realistic medical scenarios with domain shifts among different clients. In this paper, we make the first attempt to assess the informativeness of local data derived from diverse domains and propose a novel methodology termed **F**ederated **E**vidential **A**ctive **L**earning (FEAL) to calibrate the data evaluation under domain shift. Specifically, we introduce a Dirichlet prior distribution in both local and global models to treat the prediction as a distribution over the probability simplex and capture both aleatoric and epistemic uncertainties by using the Dirichlet-based evidential model. Then we employ the epistemic uncertainty to calibrate the aleatoric uncertainty. Afterward, we design a diversity relaxation strategy to reduce data redundancy and maintain data diversity. Extensive experiments and analysis on five real multi-center medical image datasets demonstrate the superiority of FEAL over the state-of-the-art active learning methods in federated scenarios with domain shifts. The code will be available at https://github.com/JiayiChen815/FEAL.*

*Equal contribution. †Yong Xia is the corresponding author. This work was supported in part by Shenzhen Science and Technology Program under Grants JCYJ20220530161616036, National Natural Science Foundation of China under Grants 62171377 and 62271405, Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project: 2022LYKFZD06), and Foshan HKUST Projects under Grants FSUST21-HKUST10E and FSUST21-HKUST11E.




(a) FAL scheme

(b) KDE of energy score

(c) $p$-value

Figure 1. **Illustration of federated active learning (FAL) in the presence of domain shift.** (a) FAL comprises model distribution, local training, model aggregation, and data annotation. (b) The KDE of energy scores depicts domain shifts across clients. (c) The low $p$-values in cross-client KDE of energy scores indicate the existence of significant domain shifts between all client pairs.

## 1. Introduction

Federated learning enables collaborative learning across multiple clinical institutions (*i.e.*, clients) to learn a unified model on the central server through model aggregation while preserving the data privacy at each client [21, 36, 57] (see Fig. 1 (a)). Unfortunately, such a learning pipeline requires each client to prepare its own labeled data, whose scale is constrained by the available expertise, time, and budget for data annotation.

One possible solution to alleviate the annotation cost is to select a part of highly informative data to annotate. Active learning (AL) has shown great potential in guiding the data selection process [1, 3, 20, 61], leading to the federated AL (FAL) framework. Such a pipeline [4, 14, 37, 41, 46, 65] allows each client to assess the informativeness of unlabeled data using either the local model at each client or the global model from the server, greatly alleviating the heavy annotation costs while retaining great performance. Nevertheless, when using a local model to select data, there is a bias toward prioritizing the data that improves the local updates while disregarding the overall generalizability of the global model. Client models trained on diverse domains may exhibit significant divergence within the parameter space, making the use of a global model aggregated from these models for data selection unreliable.

Recent advances in FAL, *e.g.*, LoGo [20] and KAFAL [3], tend to harness the knowledge of both local and global models to identify informative samples. Although this strategy has been proven to be more effective than employing a single model, these methods focus mainly on the class imbalance issue while assuming that the data at multiple clients is from the same domain. However, the domain shift across clients is commonly seen in real-world applications, which is evidenced by the extremely low $p$-values of the kernel density estimation (KDE) of energy scores [31] (see Fig. 1 (b) and (c)). The existence of domain shift renders two major challenges for FAL. **(1) Overconfidence;** Existing FAL methods evaluate data uncertainty based on the softmax prediction made by a deterministic model, which is essentially a point estimate and can be miscalibrated easily on data with domain shifts [31, 33, 42], resulting in unreliable uncertainty evaluation. **(2) Limited uncertainty representation.** Uncertainty can be divided into aleatoric uncertainty (or data uncertainty) and epistemic uncertainty (or knowledge uncertainty) [43]. The former reflects the inherent complexity of data, such as class overlap and instance noise [54]. The latter captures the restricted knowledge of a model caused by insufficient data or domain shifts. The softmax prediction can represent the aleatoric uncertainty but fails to capture the epistemic uncertainty, resulting in incomplete evaluations, which are particularly noticeable in the presence of domain shift.

To address both challenges, we propose the **Federated Evidential Active Learning (FEAL)** method. Built upon the Dirichlet-based evidential model [47, 62], FEAL treats the categorical prediction of a sample as following a Dirichlet distribution, thus allowing multiple potential predictions for a sample. FEAL comprises two key modules, *i.e.*, calibrated evidential sampling (CES) and evidential model learning (EML). CES is a novel FAL sampling strategy that incorporates both uncertainty and diversity measures. It utilizes the expected entropy of potential predictions to quantify aleatoric uncertainty and aggregates the aleatoric uncertainty in both global and local models. Further, CES employs the differential entropy of the Dirichlet distribution to characterize the epistemic uncertainty [51] and utilizes the epistemic uncertainty in the global model to calibrate the aggregated aleatoric uncertainty. To enhance data selection, diversity relaxation is also employed with the local model to reduce redundancy and maintain diversity among the selected samples. In addition to active sampling, we introduce evidence regularization in EML for accurate evidence representation and data assessment. The main contributions of this work are summarized as follows:

- We explore a rarely studied problem, FAL with domain shifts, which aims to attain a global model with a limited annotation budget for local clients amidst domain shifts.
- We propose the FEAL method, with a sampling strategy CES and a local training scheme EML, to tackle the challenges in FAL with domain shifts. CES is designed to select informative samples by leveraging aleatoric and epistemic uncertainty with both global and local models and retaining sample diversity. EML is developed to regularize the evidence for improved data evaluation.
- We conduct extensive experiments on five real multi-center medical image datasets, comprising two datasets for classification and three datasets for segmentation. The results suggest the superiority of our FEAL method over its AL and FAL counterparts.

## 2. Related Work

### 2.1. Federated Learning with Domain Shifts

Domain shift is a long-standing challenge for federated learning. Previous approaches can be divided into regularization-based, aggregation-based, and personalized ones. **Regularization-based methods** implemented regularization on model parameters [19, 27, 52] or feature embeddings [12, 15, 26, 60] to address the objective inconsistency induced by domain shift. **Aggregation-based methods** dynamically adjust aggregation weights based on data quality [32], estimated client contribution [17], generalization gap between global and local models [67], layer-wise divergence [44] or performance on proxy dataset [29]. **Personalized methods** aggregated domain-agnostic layers, while customizing domain-specific layers for local clients, including batch normalization (BN) [28], high-frequency convolution [6] and prediction layers [57]. Additionally, several methods enhanced data diversity [30, 68] to refine data distribution and mitigate statistical heterogeneity [64]. These approaches strive to mitigate the impact of domain shifts across clients in supervised scenarios with fully annotated training samples. Unfortunately, they ignore the substantial annotation costs for each client. In contrast, we further leverage active learning to reduce annotation costs by selecting the most informative data and propose a label-efficient method for federated learning with domain shifts.

### 2.2. AL Methods

Conventional AL methods can be categorized into uncertainty-based, diversity-based, and hybrid ones. **Uncertainty-based AL methods** aim to select the most ambiguous unlabeled samples for annotation. Classical approaches such as least confidence sampling [49], margin-based sampling [37], and entropy-based sampling [50] evaluate the data uncertainty based on categorical probabilities. Yoo *et al.* [65] and Huang *et al.* [14] estimated the loss for uncertainty assessment. Moreover, several approaches assess the data uncertainty by analyzing the prediction inconsistency among multiple augmented samples [11], standard and dropout inferences [9, 10], or original and disturbed features [41]. **Diversity-based AL methods** aim to identify a subset of samples that captures the distribution of the complete dataset. A variety of approaches have been proposed that exploit core-set techniques [4, 46] or clustering methods [23, 38, 55] in the latent feature space, incorporate a diversity constraint in the optimization process [8, 63], or model the distribution discrepancy between labeled and unlabeled samples [24] in order to identify a diverse collection of samples. **Hybrid AL methods** exploit both uncertainty and diversity in their sampling strategies. Ash *et al.* [2] clustered the gradient embeddings to guarantee both uncertainty and diversity. A two-stage sampling strategy has also been implemented [41, 56, 66]. However, these methods primarily focus on data selection driven by aleatoric uncertainty, often neglecting its sufficiency and reliability in practical scenarios. In this work, we developed a Dirichlet-based evidential model to capture both aleatoric and epistemic uncertainties. We further leveraged the epistemic uncertainty to calibrate uncertainty estimates, enhancing their reliability in the context of domain shifts.

### 2.3. FAL Methods

FAL aims to enhance the annotation efficacy of each local client in decentralized learning. In contrast to the centralized scenarios, there exist two potential query-selector models in FAL [20], including the global model and the local model. Both Wu *et al.* [61] and Ahn *et al.* [1] exclusively utilized a singular model for data evaluation. Specifically, Wu *et al.* [61] introduced a hybrid metric that considers both the locally predicted loss and the local feature distances between unlabeled and labeled samples. By contrast, Ahn *et al.* [1] argued that evaluating samples with the global model contributes to the objectives of federated learning and recommended applying sampling strategies solely with the global model. Nevertheless, as demonstrated in [20], the superiority of the two query-selector models depends on the global and local heterogeneous levels, and it is necessary to leverage the knowledge of both global and local models. Kim *et al.* [20] proposed a hybrid metric called LoGo, which applies $k$-means clustering technique [34] on the gradient space of the local model and subsequently conducts cluster-wise sampling using the global model. Cao *et al.* [3] proposed a knowledge-specialized sampling strategy, which leverages the discrepancy between the global model and local model to assess data uncertainty. However, these methods focus on the local data from a singular domain, which is less realistic. Though partial approaches [3, 20] account for heterogeneity caused by class imbalance, they often neglect another heterogeneous property known as domain shifts. In this work, we propose the uncertainty calibration method to achieve reliable uncertainty evaluation with domain shifts across multiple clients.

## 3. Methodology

### 3.1. Problem Formulation

The overview of our FEAL framework is displayed in Fig. 2 and Appendix A. Under this framework, we maintain $K$ local models $\{\boldsymbol{\theta}_k\}_{k=1}^K$ on clients and a global model $\boldsymbol{\theta}$ on the central server. The $k$-th local client contains a labeled set $L_k$ and an unlabeled set $U_k$. FAL comprises two iterative phases: federated model training and local data annotation. Federated model training involves model distribution, local training, and model aggregation. In the first round, the $k$-th client randomly selects $B_k$ unlabeled samples and annotates them to form the initial labeled set $L_k^1 = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{B_k}$, and the unlabeled set is updated to $U_k^1 = U_k \setminus L_k^1$. In the $r$-th FAL round, the $k$-th client constructs the query set $Q_k^r = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{j=1}^{B_k}$ for annotation using the sampling strategy and updates the labeled set to $L_k^r = L_k^{r-1} \cup Q_k^r$, whereas the unlabeled set is updated to $U_k^r = U_k^{r-1} \setminus Q_k^r$. Subsequent federated model training proceeds with the updated labeled set $L_k^r$. The FAL process is repeated for $R$ times as required.

### 3.2. Dirichlet-based Evidential Model in FAL

For federated active learning, we employ a Dirichlet-based evidential model to effectively capture aleatoric and epistemic uncertainties in both global and local models. In this section, we begin by presenting the foundational formulation of the Dirichlet-based evidential model.

We start with the general $C$-class classification task. Given an input sample $\boldsymbol{x}$ from the $k$-th client, a model $f$ parameterized with $\boldsymbol{\theta}$ projects $\boldsymbol{x}$ into a $C$-dimensional logits $f(\boldsymbol{x}, \boldsymbol{\theta})$. The classical CNN utilizes the softmax operator to transform the logits $f(\boldsymbol{x}, \boldsymbol{\theta})$ into the prediction of class probabilities $\boldsymbol{\rho}$. However, this approach essentially provides a single-point estimate of $\boldsymbol{\rho}$ and can be easily miscalibrated

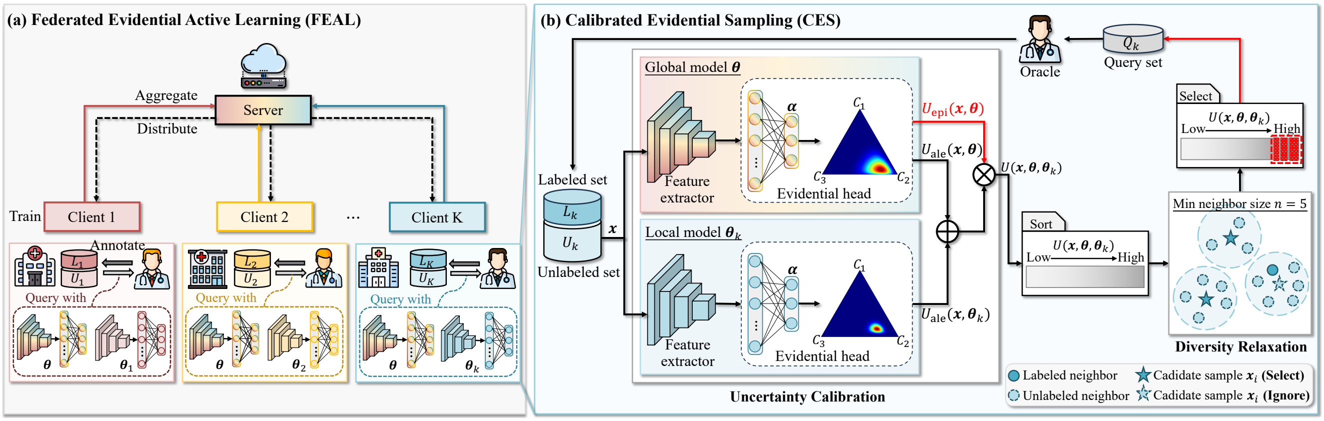



Figure 2. **Illustration of the proposed FEAL method.** (a) Overview of FEAL. (b) Illustration of CES module, including uncertainty calibration and diversity relaxation.

on local data from diverse domains. The Dirichlet-based evidential model, on the other hand, views the categorical prediction $\boldsymbol{\rho}$ as a random variable with a Dirichlet distribution $Dir(\boldsymbol{\rho}|\boldsymbol{\alpha})$. The probability density function of $\boldsymbol{\rho}$ [47, 62], given $\boldsymbol{x}$ and $\boldsymbol{\theta}$, is formulated as:

$$p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) = \begin{cases} \dfrac{\Gamma(\sum_{c=1}^{C} \alpha_c)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} \rho_c^{\alpha_c - 1}, (\boldsymbol{\rho} \in \Delta^C) \\ 0 \qquad\qquad\qquad\qquad\quad , (\text{otherwise}) \end{cases} \tag{1}$$

where $\boldsymbol{\alpha}$ denotes the parameters of the Dirichlet distribution for sample $\boldsymbol{x}$, $\Gamma(\cdot)$ is the Gamma function, and $\Delta^C = \{\boldsymbol{\rho}|\sum_{c=1}^{C} \rho_c{=}1 \text{ and } 0{<}\rho_c{<}1\}$ represents the $C$-dimensional unit simplex.

The posterior probability $P(y = c|\boldsymbol{x}, \boldsymbol{\theta})$ for class $c$, *a.k.a.*, the expected categorical prediction $\overline{\rho}_c$, is given by:

$$P(y = c|\boldsymbol{x}, \boldsymbol{\theta}) = \int p(y = c|\boldsymbol{\rho}) \cdot p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\rho} = \frac{\alpha_c}{S}, \tag{2}$$

where $S = \sum_{c=1}^{C} \alpha_c$ represents the Dirichlet strength. The derivation of Eq. 2 is provided in Appendix B.1.

Drawing on concepts from Dempster-Shafer theory [48] and subjective logic [18], the parameter $\boldsymbol{\alpha}$ is linked to the accumulated evidence $\boldsymbol{e}$ which quantifies the degree of support for the prediction on sample $\boldsymbol{x}$. The parameter $\boldsymbol{\alpha}$ is derived as $\boldsymbol{\alpha} = \boldsymbol{e} + 1 = \mathcal{A}(f(\boldsymbol{x}, \boldsymbol{\theta})) + 1$, where $\mathcal{A}(\cdot)$ is a non-negative activation function that transforms the logits $f(\boldsymbol{x}, \boldsymbol{\theta})$ into evidence $\boldsymbol{e}$.

In our study, all local models adopt the same Dirichlet-based evidential architecture with the global model to communicate between local clients and the central server.

### 3.3. Calibrated Evidential Sampling

In the context of FAL with domain shifts, we integrate both uncertainty and diversity measures to identify the most informative samples for annotation (see Fig. 2(b)). As for uncertainty evaluation, we leverage the epistemic uncertainty in the global model to calibrate the aleatoric uncertainty in both global and local models. We now delve into its details.

**Aleatoric uncertainty.** Dirichlet-based evidential models interpret the categorical prediction $\boldsymbol{\rho}$ as a distribution rather than a singular point estimate, which acknowledges a range of possible predictions. We use the expected entropy of all possible predictions to deliver the aleatoric uncertainty [62] to quantify the inherent complexity or ambiguity present in local data. Given a sample $\boldsymbol{x}$ and the global model $\boldsymbol{\theta}$, the aleatoric uncertainty of the sample $\boldsymbol{x}$ in the global model $\boldsymbol{\theta}$ is represented as:

$$\begin{aligned} U_{\text{ale}}(\boldsymbol{x}, \boldsymbol{\theta}) &= \mathbb{E}_{p(\boldsymbol{\rho}|\boldsymbol{x},\boldsymbol{\theta})}[\mathcal{H}[P(y|\boldsymbol{\rho})]] \\ &= -\sum_{c=1}^{C} \mathbb{E}_{p(\rho_c|\boldsymbol{x},\boldsymbol{\theta})}[\rho_c \cdot \log \rho_c] \\ &= \sum_{c=1}^{C} \frac{\alpha_c}{S} \cdot [\psi(S+1) - \psi(\alpha_c + 1)], \end{aligned} \tag{3}$$

where $\mathcal{H}(\cdot)$ denotes the Shannon entropy [50]. Similarly, the aleatoric uncertainty in the local model $k$ is $U_{\text{ale}}(\boldsymbol{x}, \boldsymbol{\theta}_k)$. The derivation of Eq. 3 is in Appendix B.2.

**Epistemic uncertainty.** In the Dirichlet distribution, the differential entropy quantifies how dispersed the probabilities are across different categories [35]. We employ the differential entropy of the Dirichlet distribution to quantify the epistemic uncertainty linked to domain shifts between the global model and local data. Specifically, given a sample $\boldsymbol{x}$ and the global model $\boldsymbol{\theta}$, the epistemic uncertainty of the sample $\boldsymbol{x}$ in the global model $\boldsymbol{\theta}$ is represented as:

$$\begin{aligned} U_{\text{epi}}(\boldsymbol{x}, \boldsymbol{\theta}) &= \mathcal{H}[p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta})] \\ &= -\int p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \cdot \log p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\rho} \\ &= \sum_{c=1}^{C} \log \frac{\Gamma(\alpha_c)}{\Gamma(S)} - (\alpha_c - 1) \cdot [\psi(\alpha_c) - \psi(S)]. \end{aligned} \tag{4}$$

The derivation of Eq. 4 is in Appendix B.2.

**Uncertainty calibration.** Given a sample $\boldsymbol{x}$, the global model $\boldsymbol{\theta}$, and the local model $\boldsymbol{\theta}_k$, we calculate the aleatoric uncertainty (Eq. 3) in both global and local models and subsequently calibrate the aleatoric uncertainty by incorporating the epistemic uncertainty (Eq. 4) from the global model. The overall calibrated uncertainty for sample $\boldsymbol{x}$ is

$$U(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\theta}_k) = [U_{\text{ale}}(\boldsymbol{x},\boldsymbol{\theta}) + U_{\text{ale}}(\boldsymbol{x},\boldsymbol{\theta}_k)] \cdot U_{\text{epi}}(\boldsymbol{x},\boldsymbol{\theta}). \quad (5)$$

**Diversity relaxation.** We adopt local constraints to ensure diversity among selected samples, contrasting with core-set techniques that impose global diversity constraints. As outlined in Alg. 1, we initially sort the unlabeled set $U_k^{r-1}$ by descending calibrated uncertainty $U(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\theta}_k)$, and then extract feature embeddings with local model $\boldsymbol{\theta}_k$. During the iteration over the unlabeled set $U_k^{r-1}$, we compute the cosine similarity $s(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for each candidate sample $\boldsymbol{x}_i$ against all other samples $\boldsymbol{x}_j \in U_k^{r-1} \setminus \boldsymbol{x}_i$ and form a neighbor set $N(\boldsymbol{x}_i)$ based on the similarity threshold $\tau$. A sample $\boldsymbol{x}_i$ is selected if its neighbor counts $|N(\boldsymbol{x}_i)|$ are less than the minimum neighbor size $n$ or if these neighbors remain unlabeled. Following this criterion, $B_k$ unlabeled samples are chosen to constitute the final set $Q_k^r$ for annotation, effectively balancing diversity and uncertainty in data selection.

**Algorithm 1** Diversity Relaxation for Local Client $k$

**Input:** unlabeled set $U_k^{r-1}$, local model $\boldsymbol{\theta}_k$, annotation budget $B_k$, similarity threshold $\tau$, minimum neighbor size $n$
**Output:** query set $Q_k^r$
1: Sort $U_k^{r-1}$ by descending calibrated uncertainty.
2: Initialize index $i = 1$ and query set $Q_k^r = \emptyset$.
3: **while** $|Q_k^r| < B_k$ **and** $i \leq |U_k^{r-1}|$ **do**
4: Select a candidate sample $\boldsymbol{x}_i$ from $U_k^{r-1}$.
5: Compute feature similarity $s(\boldsymbol{x}_i,\boldsymbol{x}_j)$ using $\boldsymbol{\theta}_k$, where $\boldsymbol{x}_j \in U_k^{r-1} \setminus \boldsymbol{x}_i$.
6: Form neighbor set $N(\boldsymbol{x}_i)$, including $\boldsymbol{x}_j$ with $s(\boldsymbol{x}_i,\boldsymbol{x}_j) \geq \tau$.
7: **if** $|N(\boldsymbol{x}_i)| < n$ **or** $N(\boldsymbol{x}_i) \cap Q_k^r = \emptyset$ **then**
8: Add $\boldsymbol{x}_i$ to $Q_k^r$.
9: **end if**
10: Increment $i$.
11: **end while**
12: **return** $Q_k^r$

### 3.4. Evidential Model Learning

Dirichlet-based evidential models treat the categorical prediction of a sample as a distribution, enabling multiple potential predictions to occur with specific probabilities. Considering all possible predictions, we adopt the Bayes risk of cross-entropy loss [47] as the task loss for classification tasks, formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{task}}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}) &= \int (\sum_{c=1}^{C} -y_c \log \rho_c) \cdot p(\boldsymbol{\rho}|\boldsymbol{x},\boldsymbol{\theta}_k)\, \mathrm{d}\boldsymbol{\rho} \\ &= \sum_{c=1}^{C} y_c \cdot [\psi(S) - \psi(\alpha_c)],\end{aligned} \quad (6)$$

where $\psi(\cdot)$ is the digamma function and $y_c$ is the label indicator for class $c$. Similarly, the Bayes risk of Dice loss [25] for segmentation tasks is:

$$\begin{aligned}\mathcal{L}_{\text{task}}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}) &= \int (1 - \frac{2}{C}\sum_{c=1}^{C} \frac{\|\boldsymbol{y}_c \circ \boldsymbol{\rho}_c\|_1}{\|\boldsymbol{y}_c^2\|_1 + \|\boldsymbol{\rho}_c^2\|_1}) \cdot p(\boldsymbol{\rho}|\boldsymbol{x},\boldsymbol{\theta}_k)\, \mathrm{d}\boldsymbol{\rho} \\ &= 1 - \frac{2}{C}\sum_{c=1}^{C} \frac{\|\boldsymbol{y}_c \circ \overline{\boldsymbol{\rho}}_c\|_1}{\|\boldsymbol{y}_c^2\|_1 + \|\overline{\boldsymbol{\rho}}_c^2\|_1 + \|\frac{\overline{\boldsymbol{\rho}}_c \circ (\mathbf{1} - \overline{\boldsymbol{\rho}}_c)}{\boldsymbol{S}+\mathbf{1}}\|_1},\end{aligned} \quad (7)$$

where $\circ$ is the Hadamard product and the expected categorical probability of $\boldsymbol{x}$ is $\overline{\boldsymbol{\rho}}_c = \frac{\boldsymbol{\alpha}_c}{\boldsymbol{S}}$. The derivation of Eq. 6 and Eq. 7 are in Appendix B.3.

We incorporate evidence regularization to further reduce incorrect evidence [47] and improve correct evidence [40].

$$\mathcal{L}_{\text{reg}}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}) = KL[Dir(\boldsymbol{\rho}|\tilde{\boldsymbol{\alpha}}) \| Dir(\boldsymbol{\rho}|\mathbf{1})] - \frac{C}{S} \cdot f(\boldsymbol{x},\boldsymbol{\theta}_k), \quad (8)$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (\mathbf{1} - \boldsymbol{y}) \odot \boldsymbol{\alpha}$ and $KL(\cdot)$ denotes the Kullback-Leibler divergence [22]. Notably, we calculate the average pixel-wise $\mathcal{L}_{\text{reg}}$ in segmentation.

The overall training objective, combining task loss $\mathcal{L}_{\text{task}}$ and evidence regularization $\mathcal{L}_{\text{reg}}$, is formulated as:

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}) = \mathcal{L}_{\text{task}}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}) + \lambda \cdot \mathcal{L}_{\text{reg}}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}), \quad (9)$$

where $\lambda$ is the trade-off weight. between the task loss and the regularization term.

## 4. Experiments

### 4.1. Experimental Settings

**Datasets.** We evaluated FEAL on five real multi-center medical image datasets, comprising two classification and three segmentation datasets. The classification datasets included

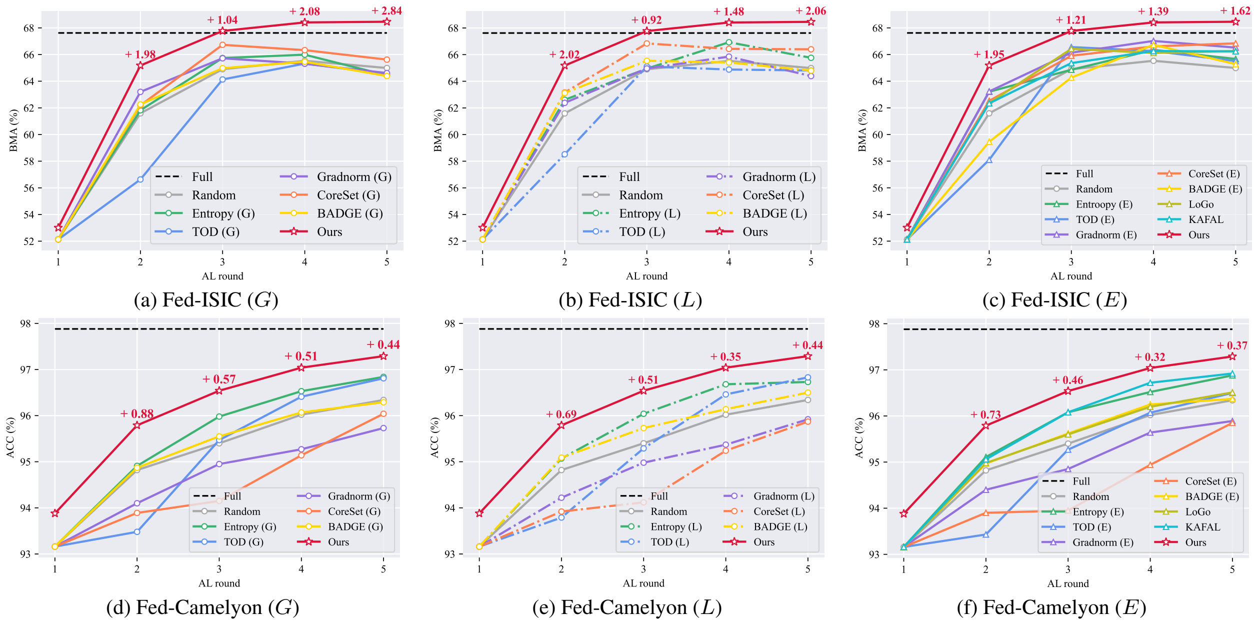
- **Fed-ISIC**: A skin lesion dataset from 4 data sources [39] containing {12413, 3954, 3363, 2259} images.
- **Fed-Camelyon**: A breast cancer histology dataset from 5 centers [16] comprising {59436, 34904, 85054, 129838, 146722} patches.

The segmentation datasets included

- **Fed-Polyp**: A endoscopic polyp dataset from 4 centers [57] with {1000, 196, 379, 612} samples.
- **Fed-Prostate**: A prostate MRI dataset from 6 data sources [30] with {261, 384, 158, 468, 421, 175} slices.
- **Fed-Fundus**: A retinal fundus dataset from 4 centers [30] with {101, 159, 400, 400} samples.

In our study, each dataset was divided using an 8:2 train-to-test split ratio at the patient level. Details of these datasets are provided in Appendix C.1.

**Evaluation metrics.** For classification, we utilized the Balanced Multi-class Accuracy (BMA) for skin lesion classification [5] and measured accuracy (ACC) for breast cancer histology classification. In the context of segmentation,




(a) Fed-ISIC ($G$) (b) Fed-ISIC ($L$) (c) Fed-ISIC ($E$)

(d) Fed-Camelyon ($G$) (e) Fed-Camelyon ($L$) (f) Fed-Camelyon ($E$)

Figure 3. **Comparison of FAL methods in medical image classification.** (a)-(c) and (d)-(f) depict the results of the Fed-ISIC and Fed-Camelyon datasets, respectively. Performance enhancements over the second-best method in each FAL round are emphasized in red text.

we used the Dice score and the 95% Hausdorff Distance (HD95) to assess segmentation results.

**Implementation details.** We conducted $R = 5$ rounds of FAL involving federated model training and data annotation. The annotation budget $B_k$ is 500 for Fed-ISIC and Fed-Camelyon, 50 for Fed-Polyp, and 20 for Fed-Prostate and Fed-Fundus. During model training, we followed the previous work [16, 57, 59] to utilize EfficientNet-B0 [53] for Fed-ISIC, DenseNet-121 [13] for Fed-Camelyon, and U-Net [30, 45] for segmentation datasets. Notably, both EfficientNet-B0 and DenseNet-121 were pre-trained on ImageNet [7]. Each experiment was conducted three times using different random seeds, and the average results were reported. More details are in Appendix C.1.

**Comparison methods.** We compared FEAL with eight FAL methods, including random sampling (Random), entropy-based sampling (Entropy) [50], TOD [14], Gradnorm [58], CoreSet [46], BADGE [2], LoGo [20], and KAFAL [3]. The first six strategies are primarily developed for standard active learning, whereas LoGo and KAFAL are specifically tailored for decentralized scenarios. To incorporate these standard AL strategies into the FAL framework, we implemented them in three distinct manners: using only the global model (referred to as $G$), depending solely on the local model ($L$), or employing a simple ensemble method with both models ($E$). It guarantees a comprehensive evaluation of these strategies in FAL. Details of comparison methods are summarized in Appendix C.1.

## 4.2. Results

**Image classification.** The comparative analysis of image classification results in Fig. 3 indicates that FEAL achieves superior results on both Fed-ISIC and Fed-Camelyon datasets. As depicted in Fig. 3, the performance of all methods exhibits a general trend of improvement with the incremental inclusion of labeled samples. However, an exception to this trend is observed in the Fed-ISIC dataset as shown in Fig. 3(a). As observed, the exclusive use of Entropy, Gradnorm, and CoreSet with a single model, whether it is a global (see Fig. 3(a)) or local model (see Fig. 3(b)), results in suboptimal performance, leading to a notable decrease in effectiveness beginning from the third round. The global model delivers unreliable uncertainty evaluations, which may result in suboptimal data selection and adversely affect the ability of the model to generalize effectively. Moreover, selecting data based on evaluations from the local model can cause overfitting to its specific client, negatively impacting the performance. Conversely, methods like Gradnorm ($E$) and TOD ($E$) that combine both global and local models often outperform those relying solely on the global model, benefiting from the additional domain-specific knowledge of the local model. However, it is important to note that without proper calibration of the global model, the combined use of both models does not always guarantee better performance than solely using the local model.

Remarkably, FEAL consistently outperforms state-of-the-art FAL methods on Fed-ISIC, as shown in Fig. 3(a)-(c). This superiority is especially noticeable in the fifth FAL round, where FEAL achieves a substantial performance

Table 1. **Comparison of FAL methods in medical image segmentation.** Dice scores for three segmentation datasets are reported. For Fed-Fundus, Dice scores for both optic disc and optic cup segmentation and their average are presented. $G$ and $L$ stand for sampling solely with the global or local model, while $E$ represents sampling with both models. Red and blue highlight the Top-1 and Top-2 results.

| Model | Method | Fed-Polyp (%) | | | | Fed-Prostate (%) | | | | Fed-Fundus (%) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R2 | R3 | R4 | R5 | R2 | R3 | R4 | R5 | R2 | R3 | R4 | R5 |
| - | Full | 78.18 | | | | 88.02 | | | | 94.32 / 85.70 (90.01) | | | |
| - | Random | 67.70 | 72.16 | 75.58 | 76.32 | 80.29 | 82.70 | 83.94 | 84.77 | 92.30 / 81.41 (86.85) | 93.33 / 84.45 (88.89) | 94.29 / 84.80 (89.54) | 94.46 / 85.05 (89.76) |
| $G$ | Entropy [50] | 67.45 | 74.65 | 75.30 | 76.69 | 82.17 | 82.53 | 84.05 | 86.10 | 93.19 / 82.61 (87.90) | 93.84 / 84.35 (89.10) | 94.27 / 85.34 (89.80) | 94.47 / 85.29 (89.88) |
| | TOD [14] | 64.99 | 74.61 | 76.24 | 78.26 | 80.75 | 83.48 | 84.31 | 85.82 | 92.70 / 82.49 (87.60) | 93.95 / 85.01 (89.48) | 94.27 / 85.63 (89.95) | 94.71 / 85.58 (90.14) |
| | Gradnorm [58] | 69.14 | 74.58 | 75.79 | 78.51 | 82.10 | 83.01 | 84.85 | 86.02 | 93.20 / 82.01 (87.60) | 94.12 / 84.71 (89.41) | 94.33 / 85.38 (89.85) | 94.56 / 85.43 (89.99) |
| | CoreSet [46] | 69.50 | 73.37 | 76.71 | 78.18 | 82.11 | 83.68 | 84.56 | 85.86 | 93.00 / 83.07 (88.03) | 93.90 / 84.75 (89.32) | 94.16 / 85.35 (89.75) | 94.51 / 85.63 (90.07) |
| | BADGE [2] | 70.09 | 74.11 | 76.38 | 76.55 | 82.78 | 83.91 | 85.39 | 85.97 | 93.17 / 82.54 (87.85) | 94.07 / 84.46 (89.26) | 94.40 / 85.37 (89.89) | 94.58 / 85.19 (89.89) |
| $L$ | Entropy | 67.48 | 73.41 | 75.07 | 78.63 | 81.08 | 82.22 | 84.36 | 85.19 | 93.19 / 83.22 (88.21) | 93.83 / 84.49 (89.16) | 94.36 / 84.97 (89.66) | 94.63 / 85.68 (90.15) |
| | TOD [14] | 65.95 | 72.92 | 75.19 | 77.97 | 79.59 | 83.74 | 85.50 | 86.03 | 92.82 / 82.34 (87.58) | 93.98 / 85.00 (89.49) | 94.37 / 85.28 (89.83) | 94.65 / 85.56 (90.10) |
| | Gradnorm [58] | 70.06 | 74.69 | 77.25 | 78.84 | 80.52 | 83.43 | 84.94 | 86.04 | 93.29 / 83.04 (88.16) | 94.13 / 84.69 (89.41) | 94.33 / 85.60 (89.97) | 94.42 / 85.53 (89.98) |
| | CoreSet [46] | 68.92 | 74.06 | 75.59 | 77.75 | 81.49 | 83.49 | 84.65 | 86.19 | 92.80 / 83.20 (88.00) | 93.87 / 84.70 (89.28) | 94.28 / 85.42 (89.85) | 94.47 / 85.54 (90.00) |
| | BADGE [2] | 70.28 | 73.96 | 76.21 | 77.63 | 82.07 | 83.54 | 85.30 | 86.06 | 93.06 / 82.65 (87.85) | 93.95 / 84.44 (89.19) | 94.34 / 85.02 (89.68) | 94.54 / 85.52 (90.03) |
| $E$ | Entropy [50] | 67.85 | 75.10 | 76.80 | 77.20 | 80.95 | 83.66 | 84.81 | 85.42 | 93.26 / 82.77 (88.01) | 94.04 / 84.69 (89.36) | 94.33 / 85.31 (89.82) | 94.38 / 85.10 (89.74) |
| | TOD [14] | 67.25 | 70.43 | 74.84 | 77.53 | 81.45 | 84.46 | 84.51 | 85.65 | 93.13 / 82.70 (87.92) | 93.63 / 84.64 (89.14) | 94.31 / 85.30 (89.81) | 94.54 / 85.82 (90.18) |
| | Gradnorm [58] | 68.01 | 75.75 | 77.73 | 75.67 | 81.21 | 83.43 | 85.30 | 85.13 | 93.36 / 83.09 (88.23) | 93.83 / 84.91 (89.37) | 94.33 / 85.59 (89.96) | 94.65 / 85.52 (90.08) |
| | CoreSet [46] | 67.77 | 74.28 | 77.69 | 75.87 | 81.30 | 84.52 | 84.75 | 86.50 | 93.24 / 82.55 (87.89) | 93.63 / 84.86 (89.24) | 94.20 / 85.50 (89.85) | 94.62 / 85.89 (90.25) |
| | BADGE [2] | 69.12 | 75.45 | 77.37 | 76.24 | 81.31 | 84.34 | 85.92 | 85.55 | 93.37 / 82.95 (88.16) | 93.99 / 85.00 (89.50) | 94.50 / 85.22 (89.86) | 94.62 / 85.44 (90.03) |
| | LoGo [20] | 69.07 | 75.76 | 74.63 | 77.24 | 82.35 | 84.56 | 85.53 | 85.97 | 93.14 / 83.01 (88.08) | 93.93 / 84.55 (89.24) | 94.18 / 85.68 (89.93) | 94.61 / 85.64 (90.12) |
| | KAFAL [3] | 69.69 | 73.83 | 75.38 | 77.97 | 82.65 | 83.49 | 85.58 | 85.96 | 93.11 / 82.75 (87.93) | 94.01 / 84.12 (89.06) | 94.37 / 85.16 (89.77) | 94.46 / 85.02 (89.74) |
| | **FEAL (Ours)** | 72.06 | 76.39 | 78.62 | 80.18 | 82.94 | 85.29 | 86.77 | 87.42 | 93.53 / 83.72 (88.63) | 94.25 / 85.19 (89.72) | 94.60 / 85.96 (90.28) | 94.89 / 86.27 (90.58) |

gain of $1.62\%$ over the second-best method, CoreSet ($E$), as demonstrated in Fig. 3(c). Additionally, it is noteworthy that FEAL achieves a performance comparable to training with the fully annotated dataset in the third round and even exceeds the fully supervised performance by $0.84\%$ in the fifth round. These advancements are primarily attributable to the effective uncertainty calibration and demonstrate the efficacy of FEAL. It is noteworthy that the baseline methods KAFAL and LoGo, designed for FAL underperform in real-world federated scenarios. Despite showing impressive results in simulated federated datasets, they fail to replicate this success in actual multi-center federated scenarios. This is mainly due to the inherent domain shift characteristics of multi-center medical data. As depicted in Fig. 3(d)-(f), FEAL also achieves superior performance on the large-scale dataset Fed-Camelyon, where each local client contains tens of thousands of patches. By employing a low-data regime, where merely about $3.43\%$ of the total training samples are annotated in the active learning process, FEAL attains $99.40\%$ of fully supervised performance after five rounds of FAL. This achievement represents a significant improvement compared to the second-best method, KAFAL, which reaches $98.93\%$ of the fully supervised performance, demonstrating the effectiveness of uncertainty calibration in FEAL. Additional results with different annotation budgets/ratios are available in Appendix C.2.

**Image segmentation.** To further evaluate the effectiveness of FEAL in segmentation tasks, we conducted experiments on three real multi-center datasets: Fed-Polyp, Fed-Prostate, and Fed-Fundus, with the results summarized in Tab. 1. As can be seen, FEAL exhibits superior performance on three multi-center segmentation datasets, as evidenced by its higher Dice scores. Specifically, for Fed-Polyp, FEAL yields a Dice score of $80.18\%$ in the fifth round, outperforming the second-best method Gradnorm ($L$) by $1.34\%$ and surpassing fully-supervised training by $2.00\%$. For Fed-Prostate, FEAL demonstrates improvements of $0.85\%$ and $0.92\%$ over the second-best method in the fourth and fifth FAL rounds, respectively. For Fed-Fundus, FEAL not only surpasses other methods in segmenting both the optic disc and optic cup but also outperforms fully supervised training in the fourth and fifth rounds of FAL. Complete results including HD95 and standard deviation are available in Appendix C.2.

### 4.3. Discussion

**Effect of uncertainty calibration.** We conducted experiments on Fed-ISIC to evaluate the effects of different uncertainty combinations: $U^G_{\text{epi}}$, $U^G_{\text{ale}}$, and $U^L_{\text{ale}}$. As summarized in Tab. 2, combining aleatoric uncertainty from both global and local models proves more effective than relying on just one model. The best results are obtained with $U^G_{\text{epi}}$, $U^G_{\text{ale}}$, and $U^L_{\text{ale}}$, showcasing the effectiveness of the proposed uncertainty calibration. The ablation results on Fed-Polyp are in Appendix C.3. Moreover, we visualize the aleatoric uncertainty in both models on Fed-Polyp in Fig. 4. It is noticeable that $U^G_{\text{ale}}$ and $U^L_{\text{ale}}$ highlight different regions, underscoring the importance of combining aleatoric uncertainty in both models for a more comprehensive assessment.

Table 2. Ablation study of uncertainty calibration on Fed-ISIC.

| $U^G_{\text{epi}}$ | $U^G_{\text{ale}}$ | $U^L_{\text{ale}}$ | Round 2 | Round 3 | Round 4 | Round 5 |
|---|---|---|---|---|---|---|
| - | ✓ | - | $60.61_{\pm 1.57}$ | $66.60_{\pm 0.33}$ | $67.09_{\pm 1.02}$ | $66.57_{\pm 1.21}$ |
| - | - | ✓ | $62.20_{\pm 3.56}$ | $66.84_{\pm 1.99}$ | $66.13_{\pm 1.52}$ | $67.45_{\pm 0.69}$ |
| - | ✓ | ✓ | $63.43_{\pm 1.11}$ | $67.18_{\pm 0.55}$ | $66.58_{\pm 1.02}$ | $66.70_{\pm 0.28}$ |
| ✓ | - | - | $61.97_{\pm 1.25}$ | $65.87_{\pm 0.59}$ | $67.09_{\pm 1.24}$ | $66.41_{\pm 1.10}$ |
| ✓ | ✓ | - | $61.95_{\pm 2.12}$ | $66.08_{\pm 0.40}$ | $67.19_{\pm 1.02}$ | $66.85_{\pm 0.84}$ |
| ✓ | - | ✓ | $61.07_{\pm 1.24}$ | $65.17_{\pm 1.58}$ | $67.16_{\pm 0.73}$ | $65.92_{\pm 1.95}$ |
| ✓ | ✓ | ✓ | $65.18_{\pm 0.41}$ | $67.77_{\pm 1.31}$ | $68.41_{\pm 1.01}$ | $68.46_{\pm 0.37}$ |

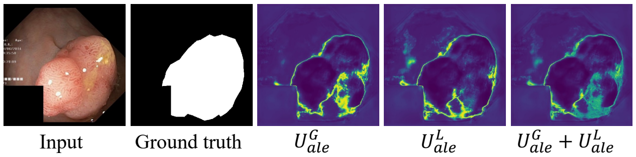
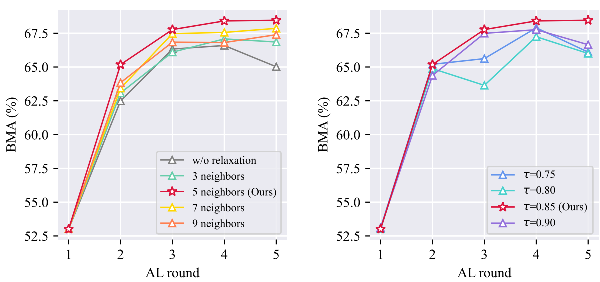



Figure 4. Visualization of aleatoric uncertainty on Fed-Polyp. $U^G_{\text{ale}}$ and $U^L_{\text{ale}}$ denote the aleatoric uncertainty in the global and local models, respectively.

**Effect of diversity relaxation.** We analyzed the impact of hyperparameters, *i.e.* minimum neighbor size $n$ and similarity threshold $\tau$, on Fed-ISIC. As depicted in Fig. 5(a), eliminating diversity relaxation ('w/o relaxation') results in a notable reduction in BMA in the fifth round, and the best performance is achieved with $n$=5 and $\tau$=0.85. The ablation results on Fed-Polyp are reported in Appendix C.3.




(a) Minimum neighbor size $n$ (b) Similarity threshold $\tau$

Figure 5. Ablation study of diversity relaxation on Fed-ISIC.

**Effect of evidential model training.** We performed experiments to compare the evidential loss ($\mathcal{L}$ in Eq. 9) against cross-entropy loss (CE) on Fed-ISIC and against dice loss (Dice) on Fed-Polyp. The results are detailed in Tab. 3. As can be seen, training with evidential loss results in an average performance gain of 1.03% on Fed-ISIC and 1.16% on Fed-Polyp, respectively. This improvement can be primarily attributed to evidence regularization, demonstrating the efficacy of evidential model training. The ablation results on the other three datasets are available in Appendix C.3.

Table 3. Ablation study of loss function.

| Dataset | Loss | Round 2 | Round 3 | Round 4 | Round 5 |
|---|---|---|---|---|---|
| Fed-ISIC | CE | $64.28_{\pm 1.64}$ | $66.69_{\pm 0.95}$ | $67.32_{\pm 1.16}$ | $67.40_{\pm 0.22}$ |
| | $\mathcal{L}$ | $65.18_{\pm 0.41}$ | $67.77_{\pm 1.31}$ | $68.41_{\pm 1.01}$ | $68.46_{\pm 0.37}$ |
| Fed-Polyp | Dice | $70.14_{\pm 0.10}$ | $75.77_{\pm 0.67}$ | $77.23_{\pm 0.21}$ | $79.48_{\pm 0.62}$ |
| | $\mathcal{L}$ | $72.06_{\pm 0.72}$ | $76.39_{\pm 0.66}$ | $78.62_{\pm 1.44}$ | $80.18_{\pm 0.10}$ |

**Effect of trade-off weight $\lambda$.** We further determined the optimal setting for the hyperparameter $\lambda$ on Fed-ISIC, choosing from the candidate set $\{1e{-}3, 5e{-}3, 1e{-}2, 5e{-}2, 1e{-}1\}$, the results are detailed in Tab. 4. As can be seen, the best performance is achieved when $\lambda = 1e{-}2$. The ablation results on Fed-Polyp are reported in Appendix C.3.

**Analysis of Dirichlet simplex.** We analyze the Dirichlet simplex on a subset of Fed-ISIC encompassing three classes. As illustrated in Fig. 6, when selecting samples with FEAL, the Dirichlet distribution becomes more concentrated at the simplex's corner for unlabeled samples, indicating reduced epistemic uncertainty in the global model. This trend verifies the effectiveness of CES in addressing domain shifts. Additionally, starting with an identical set of labeled samples, we tracked the selection of samples in the second FAL round utilizing various FAL methods. The Dirichlet simplexes of different methods are visualized in Fig. 7. As can be seen, the Dirichlet distribution of samples selected by FEAL showcases a broader spread across the simplex, indicating that FEAL effectively models the global model's knowledge of local data and prioritizes selecting samples characterized by high epistemic uncertainty. More details and results are available in Appendix C.3.

Table 4. Ablation study of trade-off weight $\lambda$ on Fed-ISIC.

| $\lambda$ | Round 2 | Round 3 | Round 4 | Round 5 |
|---|---|---|---|---|
| $1e{-}3$ | $63.49_{\pm 3.00}$ | $64.57_{\pm 2.70}$ | $66.25_{\pm 1.17}$ | $65.45_{\pm 1.06}$ |
| $5e{-}3$ | $63.10_{\pm 2.07}$ | $65.79_{\pm 2.57}$ | $66.00_{\pm 2.09}$ | $66.48_{\pm 0.86}$ |
| $1e{-}2$ | $65.18_{\pm 0.41}$ | $67.77_{\pm 1.31}$ | $68.41_{\pm 1.01}$ | $68.46_{\pm 0.37}$ |
| $5e{-}2$ | $62.12_{\pm 0.99}$ | $67.21_{\pm 1.42}$ | $66.92_{\pm 0.70}$ | $66.90_{\pm 0.93}$ |
| $1e{-}1$ | $63.53_{\pm 2.03}$ | $66.21_{\pm 0.35}$ | $66.03_{\pm 2.34}$ | $67.78_{\pm 1.17}$ |

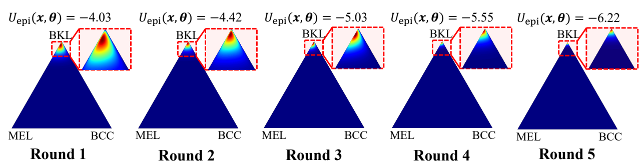



Figure 6. Visualization of the Dirichlet simplex for unlabeled samples across five FAL rounds using FEAL.

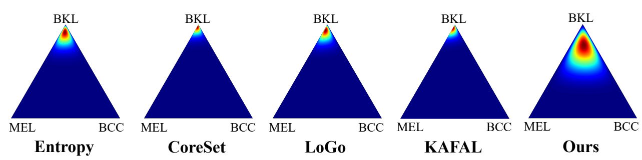



Figure 7. Visualization of the Dirichlet simplex for samples selected in the second FAL round using various sampling strategies.

## 5. Conclusion and Social Impact

To address the challenge of unreliable data assessment using the global model under domain shifts, we proposed a method FEAL, which places a Dirichlet prior over categorical probabilities to treat the prediction as a distribution over the probability simplex and leverages both aleatoric uncertainty and epistemic uncertainty to calibrate the uncertainty evaluation, enhancing the reliability of data assessment and incorporating diversity relaxation to maintain sample diversity. Extensive results verify the effectiveness. This work holds the potential to advance healthcare by preserving data privacy and facilitating collaborative research, ultimately leading to more accessible and effective patient care.

## References


[1] Jin-Hyun Ahn, Kyungsang Kim, Jeongwan Koh, and Quanzheng Li. Federated active learning (f-al): an efficient annotation strategy for federated learning. *arXiv preprint arXiv:2202.00195*, 2022. 2, 3

[2] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020. 3, 6, 7

[3] Yu-Tong Cao, Ye Shi, Baosheng Yu, Jingya Wang, and Dacheng Tao. Knowledge-aware federated active learning with non-iid data. In *ICCV*, pages 22279–22289, 2023. 2, 3, 6, 7

[4] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *CVPR*, pages 9583–9592, 2021. 2, 3

[5] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Med. Image. Anal.*, 75:102305, 2022. 5

[6] Zhen Chen, Meilu Zhu, Chen Yang, and Yixuan Yuan. Personalized retrogress-resilient framework for real-world medical federated learning. In *MICCAI*, pages 347–356, 2021. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 6

[8] Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and S Shankar Sasrty. A convex optimization framework for active learning. In *ICCV*, pages 209–216, 2013. 3

[9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *ICML*, 2016. 3

[10] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *ICML*, 2017. 3

[11] Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *ECCV*, pages 510–526, 2020. 3

[12] Yongxin Guo, Xiaoying Tang, and Tao Lin. Fedbr: improving federated learning on heterogeneous data via local learning bias reduction. In *ICML*, pages 12034–12054, 2023. 2

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 6

[14] Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. In *ICCV*, pages 3447–3456, 2021. 2, 3, 6, 7

[15] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pages 16312–16322. IEEE, 2023. 2

[16] Meirui Jiang, Zirui Wang, and Qi Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *AAAI*, pages 1087–1095, 2022. 5, 6

[17] Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *CVPR*, pages 16302–16311, 2023. 2

[18] Audun Jøsang. *Subjective logic*. 2016. 4

[19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, pages 5132–5143, 2020. 2

[20] SangMook Kim, Sangmin Bae, Hwanjun Song, and Se-Young Yun. Re-thinking federated active learning based on inter-class diversity. In *CVPR*, pages 3944–3953, 2023. 2, 3, 6, 7

[21] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016. 1

[22] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951. 5

[23] Natsumaro Kutsuna, Takumi Higaki, Sachihiro Matsunaga, Tomoshi Otsuki, Masayuki Yamaguchi, Hirofumi Fujii, and Seiichiro Hasezawa. Active learning framework with iterative clustering for bioimage classification. *Nat. Commun.*, 3 (1):1032, 2012. 3

[24] Haohan Li and Zhaozheng Yin. Attention, suggestion and annotation: a deep active learning framework for biomedical image segmentation. In *MICCAI*, pages 3–13, 2020. 3

[25] Hao Li, Yang Nan, Javier Del Ser, and Guang Yang. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. *Neural Comput. Appl.*, pages 1–15, 2022. 5

[26] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *CVPR*, pages 10713–10722, 2021. 2

[27] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *MLSys*, 2:429–450, 2020. 2

[28] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021. 2

[29] Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. *arXiv preprint arXiv:2302.10911*, 2023. 2

[30] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021. 2, 5, 6

[31] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 33: 21464–21475, 2020. 2

[32] Benteng Ma, Yu Feng, Geng Chen, Changyang Li, and Yong Xia. Federated adaptive reweighting for medical image classification. *Pattern Recogn.*, 144:109880, 2023. 2

[33] Benteng Ma, Jing Zhang, Yong Xia, and Dacheng Tao. Vnas: Variational neural architecture search. *Int. J. Comput. Vis.*, 2024. 2

[34] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297, 1967. 3

[35] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *NeurIPS*, 31, 2018. 4

[36] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pages 1273–1282, 2017. 1

[37] Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021. 2, 3

[38] Hieu T Nguyen and Arnold Smeulders. Active learning using pre-clustering. In *ICML*, page 79, 2004. 3

[39] Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *NeurIPS*, 35:5315–5334, 2022. 5

[40] Deep Shankar Pandey and Qi Yu. Learn to accumulate evidence from all training samples: Theory and practice. In *ICML*, pages 26963–26989, 2023. 5

[41] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Gholamreza Reza Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing. In *CVPR*, pages 12237–12246, 2022. 2, 3

[42] Tim Pearce. *Uncertainty in neural networks; bayesian ensembles, priors & prediction intervals*. PhD thesis, University of Cambridge, 2020. 2

[43] Tim Pearce, Alexandra Brintrup, and Jun Zhu. Understanding softmax confidence and uncertainty. *arXiv preprint arXiv:2106.04972*, 2021. 2

[44] Yasar Abbas Ur Rehman, Yan Gao, Pedro Porto Buarque De Gusmão, Mina Alibeigi, Jiajun Shen, and Nicholas D Lane. L-dawa: Layer-wise divergence aware weight aggregation in federated self-supervised visual representation learning. In *ICCV*, pages 16464–16473, 2023. 2

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 6

[46] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 2, 3, 6, 7

[47] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *NeurIPS*, 31, 2018. 2, 4, 5

[48] Kari Sentz and Scott Ferson. Combination of evidence in dempster-shafer theory. 2002. 4

[49] Burr Settles. Active learning literature survey. 2009. 3

[50] Claude Elwood Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. 3, 4, 6, 7

[51] Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. Post-hoc uncertainty learning using a dirichlet meta-model. In *AAAI*, pages 9772–9781, 2023. 2

[52] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. *arXiv preprint arXiv:1910.07796*, 2019. 2

[53] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 6

[54] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *TMLR*, 2023. 2

[55] Ruth Urner, Sharon Wulff, and Shai Ben-David. Plal: Cluster-based active learning. In *COLT*, pages 376–397, 2013. 3

[56] Fan Wang, Zhongyi Han, Zhiyan Zhang, Rundong He, and Yilong Yin. Mhpl: Minimum happy points learning for active source free domain adaptation. In *CVPR*, pages 20008–20018, 2023. 3

[57] Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *ECCV*, pages 456–472, 2022. 1, 2, 5, 6

[58] Tianyang Wang, Xingjian Li, Pengkun Yang, Guosheng Hu, Xiangrui Zeng, Siyu Huang, Cheng-Zhong Xu, and Min Xu. Boosting active learning via improving test performance. In *AAAI*, pages 8566–8574, 2022. 6, 7

[59] Jeffry Wicaksana, Zengqiang Yan, and Kwang-Ting Cheng. Fca: Taming long-tailed federated medical image classification by classifier anchoring. *arXiv preprint arXiv:2305.00738*, 2023. 6

[60] Nannan Wu, Li Yu, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. Fediic: Towards robust federated learning for class-imbalanced medical image classification. In *MICCAI*, pages 692–702, 2023. 2

[61] Xing Wu, Jie Pei, Cheng Chen, Yimin Zhu, Jianjia Wang, Quan Qian, Jian Zhang, Qun Sun, and Yike Guo. Federated active learning for multicenter collaborative disease diagnosis. *IEEE Trans. Med. Imaging.*, 2022. 2, 3

[62] Mixue Xie, Shuang Li, Rui Zhang, and Chi Harold Liu. Dirichlet-based uncertainty calibration for active domain adaptation. *arXiv preprint arXiv:2302.13824*, 2023. 2, 4

[63] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *Int. J. Comput. Vis.*, 113:113–127, 2015. 3

[64] Mang Ye, Xiuwen Fang, Bo Du, Pong C Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput. Surv.*, 56(3):1–44, 2023. 2

[65] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *CVPR*, 2019. 2, 3

[66] Jiakang Yuan, Bo Zhang, Xiangchao Yan, Tao Chen, Botian Shi, Yikang Li, and Yu Qiao. Bi3d: Bi-domain active learning for cross-domain 3d object detection. In *CVPR*, pages 15599–15608, 2023. 3

[67] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *CVPR*, pages 3954–3963, 2023. 2

[68] Tianfei Zhou and Ender Konukoglu. Fedfa: Federated feature augmentation. *arXiv preprint arXiv:2301.12995*, 2023. 2