Supplementary Material of "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

A. Workflow of FEAL

The detailed workflow of FEAL is summarized in Alg. 2.

Algorithm 2 Workflow of FEAL

Input: global model $\boldsymbol{\theta}$, local models $\{\boldsymbol{\theta}_k\}_{k=1}^K$, unlabeled sets $\{U_k\}_{k=1}^K$, annotation budget $\{B_k\}_{k=1}^K$, active learning rounds R, aggregation weights $\{\alpha_k\}_{k=1}^K$, communication rounds T
Output: trained global model $\boldsymbol{\theta}^*$

/* 1st AL round */
for $k = 1$ **to** K **do**
  Randomly annotate B_k samples from U_k to construct $L_k^1 = \{(\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{B_k}$ and update $U_k^1 = U_k \setminus L_k^1$.
  $\{\alpha_k\}_{k=1}^K \leftarrow \frac{\{|L_k^1|\}_{k=1}^K}{\sum_{k=1}^K |L_k^1|}$
end for
$\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}$
for $t = 1$ **to** T **do**
  for $k = 1$ **to** K **do**
    $\boldsymbol{\theta}_k^1 \leftarrow \boldsymbol{\theta}^*$
    $\boldsymbol{\theta}_k^1 \leftarrow \text{LocalTraining}(\boldsymbol{\theta}_k^1, L_k^1)$
  end for
  $\boldsymbol{\theta}^* \leftarrow \text{FedAvg}(\{\boldsymbol{\theta}_k^1\}_{k=1}^K, \{\alpha_k\}_{k=1}^K)$
end for
/* 2nd $\sim$ R-th AL round */
for $r = 2$ **to** R **do**
  for $k = 1$ **to** K **do**
    $Q_k^r = \text{FEAL}(\boldsymbol{\theta}^*, \boldsymbol{\theta}_k^{r-1}, U_k^{r-1})$ ▷ Local data annotation
    $L_k^r = L_k^{r-1} \cup Q_k^r,\ U_k^r = U_k^{r-1} \setminus Q_k^r$
    $\{\alpha_k\}_{k=1}^K \leftarrow \frac{\{|L_k^r|\}_{k=1}^K}{\sum_{k=1}^K |L_k^r|}$
  end for
  for $t = 1$ **to** T **do**
    for $k = 1$ **to** K **do**
      $\boldsymbol{\theta}_k^r \leftarrow \boldsymbol{\theta}^*$ ▷ Model distribution
      $\boldsymbol{\theta}_k^r \leftarrow \text{LocalTraining}(\boldsymbol{\theta}_k^r, L_k^r)$ ▷ Local training
    end for
    $\boldsymbol{\theta}^* \leftarrow \text{FedAvg}(\{\boldsymbol{\theta}_k^r\}_{k=1}^K, \{\alpha_k\}_{k=1}^K)$ ▷ Model aggregation
  end for
end for
return $\boldsymbol{\theta}^*$

The sampling strategy employed by FEAL and the aggregation method utilized in FedAvg are detailed in Alg. 3 and Alg. 4, respectively.

Algorithm 3 Sampling strategy of FEAL

Input: global model $\boldsymbol{\theta}^*$, local model $\boldsymbol{\theta}_k^{r-1}$, unlabeled set U_k^{r-1}, annotation budget B_k
Output: query set Q_k^r

for all $\boldsymbol{x} \in U_k^{r-1}$ **do**
  Compute $U_{\text{ale}}(\boldsymbol{x}, \boldsymbol{\theta}_k^{r-1})$ and $U_{\text{ale}}(\boldsymbol{x}, \boldsymbol{\theta}^*)$ using Eq. 3.
  Compute $U_{\text{epi}}(\boldsymbol{x}, \boldsymbol{\theta}^*)$ using Eq. 4.
  Compute $U(\boldsymbol{x}, \boldsymbol{\theta}^*, \boldsymbol{\theta}_k^{r-1})$ by Eq. 5.
end for
Determine the query set Q_k^r according to Alg. 1.
return Q_k^r

Algorithm 4 Aggregation method of FedAvg

Input: local models $\{\boldsymbol{\theta}_k^r\}_{k=1}^K$, aggregation weights $\{\alpha_k\}_{k=1}^K$
Output: global model $\boldsymbol{\theta}^*$

$\boldsymbol{\theta}^* \leftarrow \sum_{k=1}^K \alpha_k \cdot \boldsymbol{\theta}_k^r$
return $\boldsymbol{\theta}^*$

B. Derivations

B.1. Dirichlet-based Evidential Model in FAL

In Dirichlet-based evidential models, given a sample $\boldsymbol{x}$ and a model $\boldsymbol{\theta}$, the categorical prediction $\boldsymbol{\rho}$ follows a Dirichlet distribution, denoted as $p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \sim Dir(\boldsymbol{\rho}|\boldsymbol{\alpha})$. The probability density function of $\boldsymbol{\rho}$ [28, 37], conditioned on $\boldsymbol{x}$ and $\boldsymbol{\theta}$, is formulated as:

$$p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(\sum_{c=1}^C \alpha_c)}{\prod_{c=1}^C \Gamma(\alpha_c)} \prod_{c=1}^C \boldsymbol{\rho}_c^{\alpha_c - 1}, (\boldsymbol{\rho} \in \Delta^C) \\ 0 \qquad\qquad\qquad\quad , (\text{otherwise}) \end{cases} \tag{10}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \cdots, \alpha_C\}$ denotes the parameters of the Dirichlet distribution for sample $\boldsymbol{x}$, $\Gamma(\cdot)$ is the Gamma

function, and $\Delta^C = \{\boldsymbol{\rho} | \sum_{c=1}^{C} \rho_c = 1 \text{ and } 0 < \rho_c < 1\}$ is the C-dimensional unit simplex.

As stated in [22], the marginal distributions of the Dirichlet distribution follow Beta distributions. Consequently, given $p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \sim Dir(\boldsymbol{\rho}|\boldsymbol{\alpha})$, we can express $p(\rho_c|\boldsymbol{x}, \boldsymbol{\theta}) \sim Beta(\rho_c|\alpha_c, S - \alpha_c)$, where $S = \sum_{c=1}^{C} \alpha_c$ represents the Dirichlet strength. The probability density function of ρ_c, given $\boldsymbol{x}$ and $\boldsymbol{\theta}$, is formulated as:

$$p(\rho_c|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{\mathcal{B}(\alpha_c, S - \alpha_c)} \rho_c^{\alpha_c - 1} (1 - \rho_c)^{S - \alpha_c - 1}, \quad (11)$$

where $\mathcal{B}(\cdot, \cdot)$ is the Beta function and $\mathcal{B}(\alpha_c, S - \alpha_c) = \frac{\Gamma(\alpha_c) \cdot \Gamma(S - \alpha_c)}{\Gamma(\alpha_c + S - \alpha_c)} = \frac{\Gamma(\alpha_c) \cdot \Gamma(S - \alpha_c)}{\Gamma(S)}$.

Combining Eq. 10 and Eq. 11, the posterior probability for class c, given $\boldsymbol{x}$ and $\boldsymbol{\theta}$, can be obtained as:

$$\begin{aligned} P(y = c|\boldsymbol{x}, \boldsymbol{\theta}) &= \int p(y = c|\boldsymbol{\rho}) \cdot p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\rho} \\ &= \int \rho_c \cdot p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\rho} \\ &= \int \rho_c \cdot p(\rho_c|\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\rho_c \\ &= \frac{\mathcal{B}(\alpha_c + 1, S - \alpha_c)}{\mathcal{B}(\alpha_c, S - \alpha_c)} \int \frac{\rho_c^{\alpha_c} (1 - \rho_c)^{S - \alpha_c - 1}}{\mathcal{B}(\alpha_c + 1, S - \alpha_c)} \, \mathrm{d}\rho_c \\ &= \frac{\mathcal{B}(\alpha_c + 1, S - \alpha_c)}{\mathcal{B}(\alpha_c, S - \alpha_c)} \\ &= \frac{\Gamma(\alpha_c + 1) \cdot \Gamma(S)}{\Gamma(S + 1) \cdot \Gamma(\alpha_c)} \\ &= \frac{\alpha_c}{S}. \end{aligned} \quad (12)$$

The Dirichlet distribution parameter $\boldsymbol{\alpha}$ is closely linked to the evidence $\boldsymbol{e}$ which reflects the support for the model prediction on the given sample $\boldsymbol{x}$ [28]. The parameter $\boldsymbol{\alpha}$ is formulated as:

$$\boldsymbol{\alpha} = \boldsymbol{e} + 1 = \mathcal{A}(f(\boldsymbol{x}, \boldsymbol{\theta})) + 1, \quad (13)$$

where $f(\boldsymbol{x}, \boldsymbol{\theta})$ denotes the output logits of model $\boldsymbol{\theta}$ for sample $\boldsymbol{x}$ and $\mathcal{A}(\cdot)$ is a non-negative activation function to transform the logits $f(\boldsymbol{x}, \boldsymbol{\theta})$ into evidence $\boldsymbol{e}$. There are several common activation functions $\mathcal{A}(\cdot)$ [25], including: $ReLU(\cdot) = \max(0, \cdot)$, $SoftPlus(\cdot) = \log(1 + \exp(\cdot))$, and $\exp(\cdot)$. In our study, $ReLU(\cdot)$ was employed as the non-negative activation function.

B.2. Calibrated Evidential Sampling

Aleatoric uncertainty. Given a sample $\boldsymbol{x}$ and the global model $\boldsymbol{\theta}$, the expected entropy of all possible predictions is utilized to depict the aleatoric uncertainty, quantifying the inherent complexity or ambiguity present in sample $\boldsymbol{x}$. The aleatoric uncertainty of the sample $\boldsymbol{x}$ in the gloabl model $\boldsymbol{\theta}$ is formulated as:

$$\begin{aligned} U_{\text{ale}}(\boldsymbol{x}, \boldsymbol{\theta}) &= \mathbb{E}_{p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta})}[\mathcal{H}[P(y|\boldsymbol{\rho})]] \\ &= -\sum_{c=1}^{C} \mathbb{E}_{p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta})}[\rho_c \cdot \log \rho_c] \\ &= -\sum_{c=1}^{C} \mathbb{E}_{p(\rho_c|\boldsymbol{x}, \boldsymbol{\theta})}[\rho_c \cdot \log \rho_c] \\ &= -\sum_{c=1}^{C} \int \log(\rho_c) \cdot \frac{\rho_c^{\alpha_c} (1 - \rho_c)^{S - \alpha_c - 1}}{\mathcal{B}(\alpha_c, S - \alpha_c)} \, \mathrm{d}\rho_c \\ &= -\sum_{c=1}^{C} \frac{\mathcal{B}(\alpha_c + 1, S - \alpha_c)}{\mathcal{B}(\alpha_c, S - \alpha_c)} \int \log(\rho_c) \cdot \frac{\rho_c^{\alpha_c} (1 - \rho_c)^{S - \alpha_c - 1}}{\mathcal{B}(\alpha_c + 1, S - \alpha_c)} \, \mathrm{d}\rho_c \\ &= -\sum_{c=1}^{C} \frac{\Gamma(\alpha_c + 1) \cdot \Gamma(S)}{\Gamma(S + 1) \cdot \Gamma(\alpha_c)} \mathbb{E}_{\rho_c \sim Beta(\rho_c|\alpha_c + 1, S - \alpha_c)}[\log \rho_c] \\ &= \sum_{c=1}^{C} \frac{\alpha_c}{S} \cdot [\psi(S + 1) - \psi(\alpha_c + 1)], \end{aligned} \quad (14)$$

where $\mathcal{H}(\cdot)$ denotes the Shannon entropy [29] and $\psi(\cdot)$ represents the digamma function. The aleatoric uncertainty of the sample $\boldsymbol{x}$ in the local model $\boldsymbol{\theta}_k$ can also be calculated as $U_{\text{ale}}(\boldsymbol{x}, \boldsymbol{\theta}_k)$ according to Eq. 14.

Epistemic uncertainty. The differential entropy of a Dirichlet distribution is employed to quantify the inherent randomness in categorical distributions [21]. This metric is beneficial in depicting epistemic uncertainty, which arises due to the global model's lack of knowledge, often caused by domain shifts. Given a sample $\boldsymbol{x}$ and the global model $\boldsymbol{\theta}$, the epistemic uncertainty is defined as:

$$\begin{aligned} U_{\text{epi}}(\boldsymbol{x}, \boldsymbol{\theta}) &= \mathcal{H}[p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta})] \\ &= -\int p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \log p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\rho} \\ &= \sum_{c=1}^{C} \log \frac{\Gamma(\alpha_c)}{\Gamma(S)} - (\alpha_c - 1) \cdot [\psi(\alpha_c) - \psi(S)]. \end{aligned} \quad (15)$$

B.3. Evidential Model Training

Task loss for classification. Dirichlet-based evidential models treat the prediction of a sample as a distribution, allowing for multiple potential predictions to occur with specific probabilities. Taking into account all potential predictions for a sample, we employ the Bayes risk of cross-entropy loss [28] as the task loss for classification. Given an input pair $(\boldsymbol{x}, \boldsymbol{y})$, the task loss for classification is de-

rived as follows:

$$\begin{aligned}\mathcal{L}_{\text{task}}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}) &= \mathbb{E}_{p(\boldsymbol{\rho}|\boldsymbol{x},\boldsymbol{\theta}_k)}[\mathcal{L}_{CE}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y})] \\ &= \int [\sum_{c=1}^{C} -y_c \log(\rho_c)] \cdot p(\boldsymbol{\rho}|\boldsymbol{x},\boldsymbol{\theta}_k)\, \mathrm{d}\boldsymbol{\rho} \\ &= -\sum_{c=1}^{C} y_c \int \log(\rho_c) \cdot p(\rho_c|\boldsymbol{x},\boldsymbol{\theta}_k)\, \mathrm{d}\rho_c \\ &= -\sum_{c=1}^{C} y_c \cdot \mathbb{E}_{\rho_c \sim Beta(\rho_c|\alpha_c, S-\alpha_c)}[\log \rho_c] \\ &= \sum_{c=1}^{C} y_c \cdot [\psi(S) - \psi(\alpha_c)].\end{aligned} \tag{16}$$

Task loss for segmentation. We leverage the Bayes risk of Dice loss as the task loss for segmentation following [17]. Given an input pair $(\boldsymbol{x}, \boldsymbol{y})$, the task loss for segmentation is denoted as follows:

$$\begin{aligned}\mathcal{L}_{\text{task}}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}) &= \mathbb{E}_{p(\boldsymbol{\rho}|\boldsymbol{x},\boldsymbol{\theta}_k)}[\mathcal{L}_{Dice}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y})] \\ &= \mathbb{E}_{p(\boldsymbol{\rho}|\boldsymbol{x},\boldsymbol{\theta}_k)}[1 - \frac{2}{C}\sum_{c=1}^{C} \frac{\|\boldsymbol{y}_c \circ \boldsymbol{\rho}_c\|_1}{\|\boldsymbol{y}_c^2\|_1 + \|\boldsymbol{\rho}_c^2\|_1}] \\ &= \mathbb{E}_{p(\boldsymbol{\rho}|\boldsymbol{x},\boldsymbol{\theta}_k)}[1 - \frac{2}{C}\sum_{c=1}^{C} \frac{\sum_{m=1}^{M} y_{mc} \cdot \rho_{mc}}{\sum_{m=1}^{M}(y_{mc}^2 + \rho_{mc}^2)}] \\ &= 1 - \frac{2}{C}\sum_{c=1}^{C} \frac{\sum_{m=1}^{M} y_{mc} \cdot \mathbb{E}[\rho_{mc}]}{\sum_{m=1}^{M}(y_{mc}^2 + \mathbb{E}[\rho_{mc}^2])},\end{aligned} \tag{17}$$

where $\circ$ is the Hadamard product and the image $\boldsymbol{x}$ is comprised of M pixels. y_{mc} and ρ_{mc} represent the label indicator and categorical probability of pixel x_m *w.r.t.* class c, respectively. By using the following equation:

$$\mathbb{E}[\rho_{mc}^2] = \mathbb{E}[\rho_{mc}]^2 + \text{Var}(\rho_{mc}), \tag{18}$$

Eq. 17 can be updated to:

$$\begin{aligned}\mathcal{L}_{\text{task}}(\boldsymbol{x},\boldsymbol{\theta}_k,\boldsymbol{y}) &= 1 - \frac{2}{C}\sum_{c=1}^{C} \frac{\sum_{m=1}^{M} y_{mc} \cdot \overline{\rho}_{mc}}{\sum_{m=1}^{M}[y_{mc}^2 + \overline{\rho}_{mc}^2 + \frac{\overline{\rho}_{mc}(1-\overline{\rho}_{mc})}{S_m+1}]} \\ &= 1 - \frac{2}{C}\sum_{c=1}^{C} \frac{\|\boldsymbol{y}_c \circ \overline{\boldsymbol{\rho}}_c\|_1}{\|\boldsymbol{y}_c^2\|_1 + \|\overline{\boldsymbol{\rho}}_c^2\|_1 + \|\frac{\overline{\boldsymbol{\rho}}_c \circ (\mathbf{1} - \overline{\boldsymbol{\rho}}_c)}{\boldsymbol{S}+\mathbf{1}}\|_1}.\end{aligned} \tag{19}$$

C. Experiments

C.1. Experimental Settings

Datasets. We verified the effectiveness of FEAL across five real-world medical image datasets, two for classification and three for segmentation. Detailed information, including the data source, the number of samples, and the resolution of each sample, is summarized in Tab. 5. Illustrative samples from each data source within the five multi-center medical image datasets are showcased in Fig. 8. Note that the Camelyon17 [2] dataset comprises five distinct data sources with varying stains, therefore, we partitioned them into five subsets to construct the Fed-Camelyon dataset within the FAL framework. Similarly, the REFUGE [24] dataset contains data from two separate sources, each of which was treated as an individual client within the Fed-Fundus dataset. For the three segmentation datasets,

Dataset	Data source	# Train	# Test	Resolution
Fed-ISIC	Client 1: BCN [23]	9,930	2,483	224×224
	Client 2: HAM_vidir_molemax [23]	3,163	791	
	Client 3: HAM_vidir_modern [23]	2,691	672	
	Client 4: HAM_rosendahl [23]	1,807	452	
Fed-Camelyon	Client 1: Camelyon17 [2]	47,548	11,888	96×96
	Client 2: Camelyon17 [2]	27,923	6,981	
	Client 3: Camelyon17 [2]	68,043	17,011	
	Client 4: Camelyon17 [2]	103,870	25,968	
	Client 5: Camelyon17 [2]	117,377	29,345	
Fed-Polyp	Client 1: Kvasir [11]	800	200	384×384
	Client 2: ETIS [30]	157	39	
	Client 3: ColonDB [32]	304	75	
	Client 4: ClinicDB [3]	490	122	
Fed-Prostate	Client 1: BIDMC [19]	225	36	384×384
	Client 2: BMC [4]	306	78	
	Client 3: HK [19]	134	24	
	Client 4: I2CVB [16]	387	81	
	Client 5: RUNMC [4]	337	84	
	Client 6: UCL [19]	152	23	
Fed-Fundus	Client 1: Drishti-GS [31]	81	20	384×384
	Client 2: RIM-ONE-r3 [8]	128	31	
	Client 3: REFUGE [24]	320	80	
	Client 4: REFUGE [24]	320	80	

Table 5. Details of multi-center datasets utilized in our study.

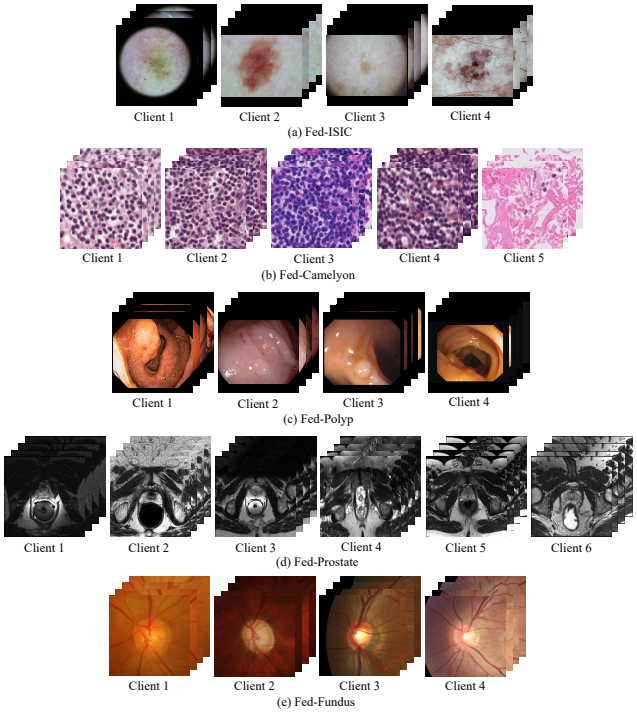


Figure 8. Illustrative samples from each data source within five multi-center medical image datasets utilized in our study.

we followed [34] to resize the images and ground truth to 384×384 pixels. In the training phase, we implemented data augmentation by randomly cropping patches of 320×320 pixels. Subsequently, we evaluated the segmentation results on the entire image of 384×384 pixels during the inference phase.

Evaluation metrics. For classification tasks, we assessed the Balanced Multi-class Accuracy (BMA) for skin lesion classification [6] and measured accuracy (ACC) for breast cancer histology classification. Regarding segmentation tasks, we used the Dice score and the 95% Hausdorff Distance (HD95) to assess segmentation results.

Implemental details. We conducted $R = 5$ rounds of FAL involving federated model training and data annotation. Federated model training comprises local training and model communication. During local training, we followed the previous work [12, 34, 36] to utilize EfficientNet-B0 [33] for the Fed-ISIC dataset, DenseNet-121 [9] for the Fed-Camelyon dataset, and U-Net [20, 26] for segmentation datasets. Notably, both EfficientNet-B0 and DenseNet-121 were pre-trained on ImageNet [7]. In FEAL, we employed the $ReLU(\cdot)$ activation as the non-negative activation function $\mathcal{A}(\cdot)$ for both global and local models. We trained local models using the Adam optimizer [14] with a learning rate of $5e{-}4$. The weight decay was set to $5e{-}4$ for the Fed-ISIC dataset and $1e{-}5$ for the other four datasets. In terms of communication, we employed the FedAvg algorithm [15], with all clients participating in each communication round. We conducted $T = 100$ rounds of communication to attain a robust global model. Regarding data annotation, we followed the previous work [5, 13] to a uniform annotation budget B_k across all local clients k. Specifically, annotation budget B_k is allocated based on the size of each dataset. The annotation budget B_k was set to 500 for both Fed-ISIC and Fed-Camelyon datasets, 50 for the Fed-Polyp dataset, and 20 for both Fed-Prostate and Fed-Fundus datasets. Furthermore, to account for the different sizes of datasets among local clients, we established a maximum annotation ratio of 85%. It means that the clients whose number of labeled samples achieves the threshold ceased further data annotation. All the experiments were conducted three times using different random seeds, and the average results were reported.

Comparison methods. We compared FEAL with eight state-of-the-art FAL approaches, including random sampling (Random), entropy-based sampling (Entropy) [29], TOD [10], Gradnorm [35], CoreSet [27], BADGE [1], LoGo [13], and KAFAL [5]. The first six methods are designed for standard active learning, whereas LoGo and KAFAL are specifically tailored for federated scenarios. For a comprehensive comparison, we implemented the first six sampling strategies in three distinct manners: (1) utilizing the global model for data evaluation (denoted as G), (2) employing the local model for data evaluation (L), and (3) integrating an ensemble technique that harnesses both global and local models for evaluation (E). The details of comparison methods and ensemble techniques are summarized as follows.

- **Random**: Randomly select B_k unlabeled samples for local client k in each FAL round.
- **Entropy**: Entropy is an uncertainty-based sampling strategy, which prioritizes the top-B_k unlabeled samples with the highest entropy scores in model predictions. Beyond just utilizing the global or local model for data evaluation, we also implemented a simple ensemble strategy that aggregates the entropy scores in both models.
- **TOD**: TOD is an uncertainty-based sampling strategy, which leverages the temporal output discrepancy to quantify the uncertainty of unlabeled samples. It selects the top-B_k unlabeled samples with the highest cyclic output discrepancy (COD) scores. In our implementation, we incorporated an additional ensemble technique [13], fine-tuning, into TOD. Specifically, local client k downloads the global model $\boldsymbol{\theta}^r$ and fine-tunes it with the available labeled samples. Subsequently, local client k employs both the fine-tuned global model and its historical local model $\boldsymbol{\theta}_k^{r-1}$ to compute the COD score, thereby integrating insights from both global and local models for more effective sampling.
- **Gradnorm**: Gradnorm initially estimates the pseudo loss of unlabeled samples, leveraging either pseudo labels or entropy scores derived from model predictions. This loss estimation is then backpropagated to determine the gradient norm across all model parameters, serving as a measure of data uncertainty to evaluate the potential value of each unlabeled sample. In our study, we aggregated the gradient norm from both models to achieve a fundamental ensemble setting. This approach enabled us to leverage the knowledge of both global and local models, thereby facilitating a more comprehensive evaluation.
- **CoreSet**: CoreSet is a diversity-based sampling strategy that identifies and selects B_K unlabeled samples with feature embeddings exhibiting the greatest dissimilarity to available labeled samples. In a basic ensemble setting, we averaged the feature embeddings from both global and local models for each sample and applied CoreSet on the interpolated feature embedding.
- **BADGE**: BADGE is a hybrid sampling strategy that simultaneously considers uncertainty and diversity metrics. It begins by extracting gradient embeddings of unlabeled samples to ensure uncertainty. Following this, it employs k-means clustering on these gradient embeddings to maintain a diverse selection of samples.

Table 6. Comparison results (mean$_{\pm \text{std}}$) on medical image datasets for classification. We evaluated the balanced multi-class accuracy (BMA) for Fed-ISIC and the accuracy for Fed-Camelyon and presented the mean result and standard deviation of three random seeds. Red and blue highlights the Top-1 and Top-2 results, respectively.

Model	Method	Fed-ISIC (%)				Fed-Camelyon (%)			
		R2	R3	R4	R5	R2	R3	R4	R5
-	Random	$61.59_{\pm 1.45}$	$64.90_{\pm 1.53}$	$65.53_{\pm 1.31}$	$64.99_{\pm 1.43}$	$94.82_{\pm 0.30}$	$95.40_{\pm 0.24}$	$96.02_{\pm 0.12}$	$96.34_{\pm 0.07}$
G	Entropy [29]	$61.82_{\pm 1.38}$	$65.74_{\pm 1.62}$	$65.99_{\pm 0.35}$	$64.44_{\pm 1.21}$	$94.91_{\pm 0.54}$	$95.98_{\pm 0.14}$	$96.53_{\pm 0.16}$	$96.84_{\pm 0.14}$
	TOD [10]	$56.63_{\pm 3.05}$	$64.13_{\pm 3.50}$	$65.32_{\pm 1.10}$	$64.54_{\pm 0.76}$	$93.48_{\pm 1.23}$	$95.47_{\pm 0.54}$	$96.41_{\pm 0.31}$	$96.81_{\pm 0.16}$
	Gradnorm [35]	$63.20_{\pm 0.49}$	$65.72_{\pm 1.29}$	$65.31_{\pm 2.03}$	$64.61_{\pm 0.76}$	$94.10_{\pm 0.18}$	$94.95_{\pm 0.23}$	$95.27_{\pm 0.15}$	$95.73_{\pm 0.18}$
	CoreSet [27]	$62.19_{\pm 1.57}$	$66.73_{\pm 0.49}$	$66.33_{\pm 0.30}$	$65.62_{\pm 1.09}$	$93.89_{\pm 0.53}$	$94.15_{\pm 0.26}$	$95.14_{\pm 0.35}$	$96.04_{\pm 0.18}$
	BADGE [1]	$62.26_{\pm 1.74}$	$64.98_{\pm 1.56}$	$65.46_{\pm 1.94}$	$64.39_{\pm 0.63}$	$94.87_{\pm 0.38}$	$95.55_{\pm 0.19}$	$96.07_{\pm 0.16}$	$96.29_{\pm 0.25}$
L	Entropy [29]	$62.61_{\pm 3.39}$	$64.95_{\pm 1.87}$	$66.93_{\pm 1.44}$	$65.76_{\pm 2.34}$	$95.07_{\pm 0.24}$	$96.05_{\pm 0.03}$	$96.68_{\pm 0.07}$	$96.73_{\pm 0.12}$
	TOD [10]	$58.52_{\pm 2.89}$	$65.11_{\pm 2.04}$	$64.88_{\pm 2.00}$	$64.81_{\pm 3.11}$	$93.79_{\pm 1.18}$	$95.29_{\pm 0.47}$	$96.46_{\pm 0.34}$	$96.83_{\pm 0.11}$
	Gradnorm [35]	$62.38_{\pm 1.90}$	$64.96_{\pm 3.21}$	$65.85_{\pm 2.79}$	$64.39_{\pm 1.44}$	$94.22_{\pm 0.38}$	$94.98_{\pm 0.39}$	$95.37_{\pm 0.54}$	$95.92_{\pm 0.26}$
	CoreSet [27]	$63.16_{\pm 0.70}$	$66.84_{\pm 0.49}$	$66.43_{\pm 0.89}$	$66.40_{\pm 0.36}$	$93.92_{\pm 0.40}$	$94.12_{\pm 0.24}$	$95.24_{\pm 0.34}$	$95.87_{\pm 0.17}$
	BADGE [1]	$63.12_{\pm 0.72}$	$65.54_{\pm 1.47}$	$65.41_{\pm 1.77}$	$64.80_{\pm 2.39}$	$95.09_{\pm 0.23}$	$95.73_{\pm 0.27}$	$96.14_{\pm 0.21}$	$96.50_{\pm 0.05}$
E	Entropy [29]	$63.21_{\pm 0.59}$	$64.86_{\pm 1.09}$	$66.35_{\pm 0.14}$	$65.57_{\pm 1.92}$	$95.03_{\pm 0.01}$	$96.08_{\pm 0.23}$	$96.52_{\pm 0.20}$	$96.88_{\pm 0.18}$
	TOD [10]	$58.10_{\pm 1.95}$	$66.56_{\pm 0.36}$	$66.26_{\pm 1.22}$	$65.51_{\pm 0.75}$	$93.17_{\pm 0.87}$	$95.27_{\pm 0.07}$	$96.07_{\pm 0.09}$	$96.50_{\pm 0.12}$
	Gradnorm [35]	$63.23_{\pm 1.25}$	$66.14_{\pm 1.51}$	$67.02_{\pm 1.00}$	$66.52_{\pm 0.75}$	$94.40_{\pm 0.06}$	$94.85_{\pm 0.21}$	$95.64_{\pm 0.04}$	$95.89_{\pm 0.13}$
	CoreSet [27]	$62.53_{\pm 1.50}$	$65.91_{\pm 0.78}$	$66.61_{\pm 0.20}$	$66.84_{\pm 0.21}$	$93.90_{\pm 0.25}$	$93.95_{\pm 0.31}$	$94.94_{\pm 0.09}$	$95.85_{\pm 0.13}$
	BADGE [1]	$59.45_{\pm 0.67}$	$64.27_{\pm 0.74}$	$66.73_{\pm 0.46}$	$64.71_{\pm 1.07}$	$94.97_{\pm 0.41}$	$95.62_{\pm 0.11}$	$96.25_{\pm 0.12}$	$96.37_{\pm 0.10}$
	LoGo [13]	$62.36_{\pm 2.30}$	$66.43_{\pm 0.69}$	$66.12_{\pm 2.64}$	$66.26_{\pm 0.50}$	$94.98_{\pm 0.07}$	$95.60_{\pm 0.15}$	$96.20_{\pm 0.26}$	$96.51_{\pm 0.05}$
	KAFAL [5]	$62.34_{\pm 0.3}$	$65.36_{\pm 1.15}$	$66.26_{\pm 1.22}$	$66.24_{\pm 1.31}$	$95.06_{\pm 0.17}$	$96.08_{\pm 0.07}$	$96.76_{\pm 0.11}$	$96.92_{\pm 0.04}$
	FEAL (Ours)	$65.18_{\pm 0.41}$	$67.77_{\pm 1.31}$	$68.41_{\pm 1.01}$	$68.46_{\pm 0.37}$	$95.79_{\pm 0.68}$	$96.54_{\pm 0.40}$	$97.04_{\pm 0.28}$	$97.29_{\pm 0.35}$

- **LoGo**: LoGo is specifically designed to address class-imbalance issues in federated active learning, harnessing insights from both global and local models. It first performed k-means clustering with gradient embeddings extracted from the local model. Then it applied cluster-wise sampling to select the most uncertain sample within each cluster, identified by the highest entropy score in the global model.
- **KAFAL**: KAFAL is another sampling strategy tailored for federated active learning with class imbalance. It employs a knowledge-specialized KL divergence, calculated between the global and local models, to quantify the informativeness of unlabeled samples for both models. For fair comparisons in our study, we implemented KAFAL by exclusively using the supervised loss component.

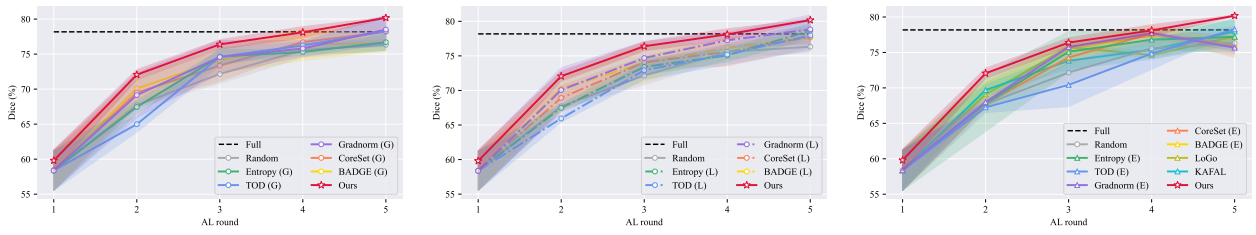
C.2. Results

Image classification. Tab. 6 summarizes the quantitative results of image classification. From the second to the fifth FAL rounds, FEAL demonstrates superior performance over the second-best method, achieving margins of 1.95%, 0.92%, 1.39%, and 1.62% on the Fed-ISIC dataset, respectively. Furthermore, on the Fed-Camelyon dataset, FEAL maintains a consistent performance advantage, surpassing the second-best method by margins of 0.69%, 0.46%, 0.3%, and 0.37% in these rounds. Additional results with different annotation budgets/ratios on the Fed-ISIC dataset are presented in Sec. C.3.

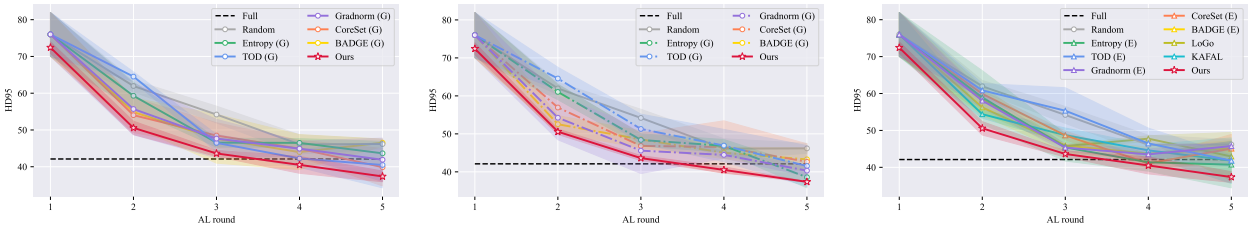
Image segmentaion. The mean results and standard deviations of Dice and HD95 metrics for the Fed-Polyp, Fed-Prostate, and Fed-Fundus datasets are illustrated in Fig. 9, Fig. 10, and Fig. 11, respectively. Extensive results in Fig. 9 - Fig. 11 indicate that FEAL yields superior performance on three multi-center segmentation datasets, as evidenced by its higher Dice scores and lower HD95 metrics. Notably, FEAL outperforms the second-best method by the margin of 1.78%, 0.63%, 0.89%, and 1.34% from the second to the fifth rounds on the Fed-Polyp dataset.

C.3. Discussions

Effect of uncertainty calibration. In addition to the ablation study conducted on the Fed-ISIC dataset for classification, we expanded our analysis to include the Fed-Polyp dataset, specifically examining the impact of uncertainty calibration in segmentation tasks. The results, including the average Dice score and the corresponding standard deviation, are summarized in Tab. 7. Here, U^G_{epi}, U^G_{ale}, and U^L_{ale} represent the epistemic uncertainty in the global model, the aleatoric uncertainty in the global model, and the aleatoric uncertainty in the local model, respectively. As can be seen, the optimal performance is attained when incorporating U^G_{epi}, U^G_{ale}, and U^L_{ale}, demonstrating the effectiveness of the proposed uncertainty calibration method. Furthermore, we visualize the aleatoric uncertainty in both global and local models on the Fed-Polyp dataset in Fig 12. As depicted in Fig 12, U^G_{ale} and U^L_{ale} highlight different regions of a sample, underscoring the significance of integrating the

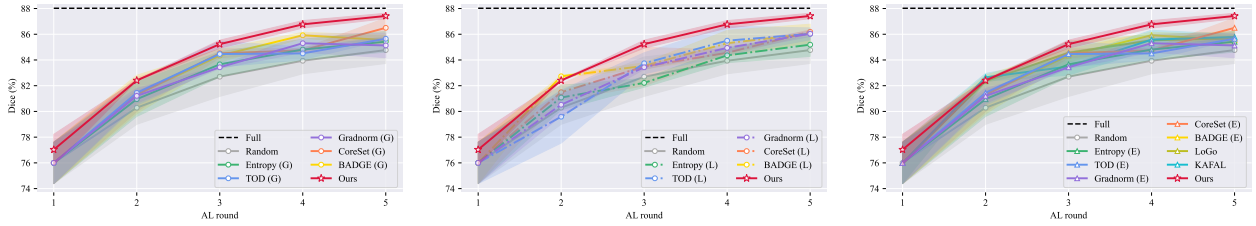


(a) Dice of Fed-Polyp (G) (b) Dice of Fed-Polyp (L) (c) Dice of Fed-Polyp (E)

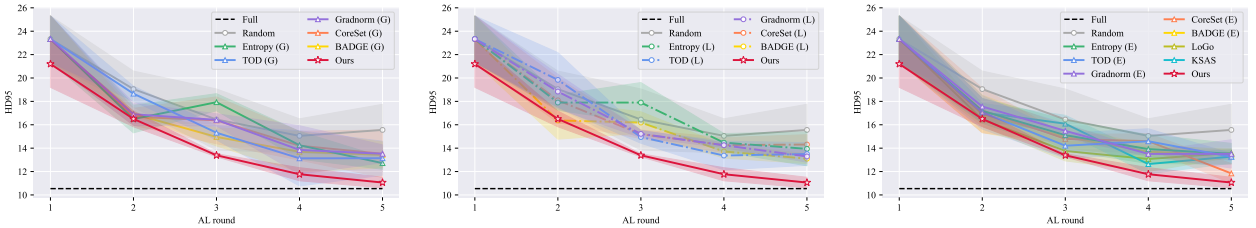


(d) HD95 of Fed-Polyp (G) (e) HD95 of Fed-Polyp (L) (f) HD95 of Fed-Polyp (E)

Figure 9. Comparison results on the Fed-Polyp dataset.

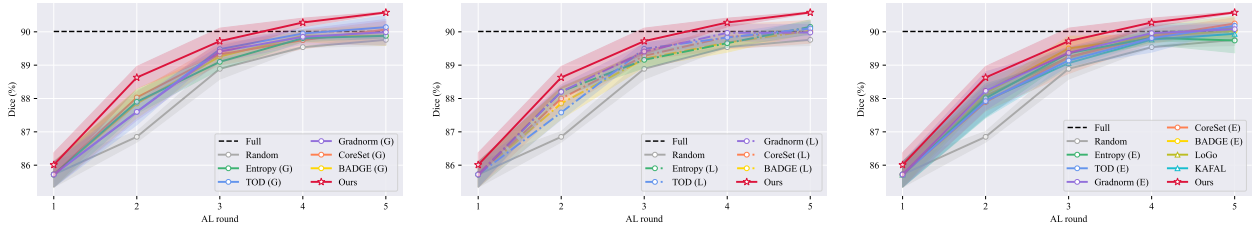


(a) Dice of Fed-Prostate (G) (b) Dice of Fed-Prostate (L) (c) Dice of Fed-Prostate (E)

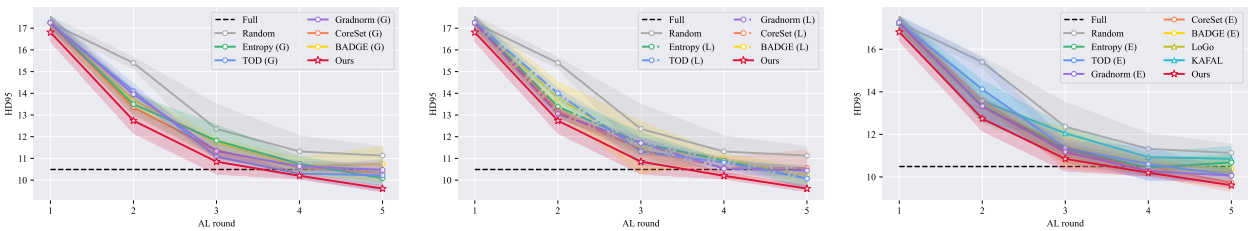


(d) HD95 of Fed-Prostate (G) (e) HD95 of Fed-Prostate (L) (f) HD95 of Fed-Prostate (E)

Figure 10. Comparison results on the Fed-Prostate dataset.



(a) Dice of Fed-Fundus (G) (b) Dice of Fed-Fundus (L) (c) Dice of Fed-Fundus (E)



(d) HD95 of Fed-Fundus (G) (e) HD95 of Fed-Fundus (L) (f) HD95 of Fed-Fundus (E)

Figure 11. Comparison results on the Fed-Fundus dataset.

aleatoric uncertainty in both global and locals for a thorough assessment.

Table 7. Ablation study of uncertainty calibration on Fed-Polyp.

U^G_{epi}	U^G_{ale}	U^L_{ale}	Round 2	Round 3	Round 4	Round 5
-	✓	-	$68.61_{\pm 0.48}$	$73.12_{\pm 2.38}$	$75.19_{\pm 1.10}$	$78.00_{\pm 1.14}$
-	-	✓	$69.13_{\pm 1.15}$	$75.19_{\pm 1.29}$	$77.85_{\pm 1.20}$	$78.12_{\pm 1.29}$
-	✓	✓	$66.89_{\pm 3.41}$	$74.48_{\pm 1.00}$	$76.56_{\pm 1.49}$	$76.86_{\pm 0.43}$
✓	-	-	$70.61_{\pm 4.22}$	$74.45_{\pm 2.29}$	$75.07_{\pm 2.42}$	$78.25_{\pm 1.29}$
✓	✓	-	$69.41_{\pm 2.17}$	$75.18_{\pm 2.48}$	$74.66_{\pm 0.87}$	$78.91_{\pm 1.04}$
✓	-	✓	$69.29_{\pm 2.51}$	$75.93_{\pm 1.41}$	$76.74_{\pm 1.08}$	$78.28_{\pm 0.67}$
✓	✓	✓	$72.06_{\pm 0.72}$	$76.39_{\pm 0.66}$	$78.62_{\pm 0.81}$	$80.18_{\pm 0.10}$

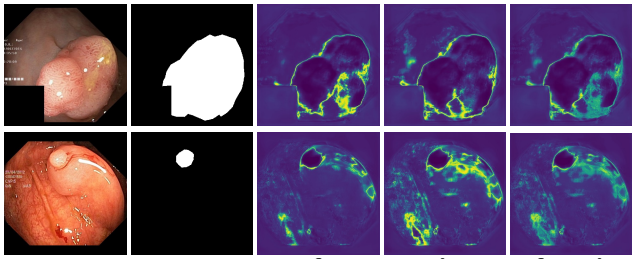
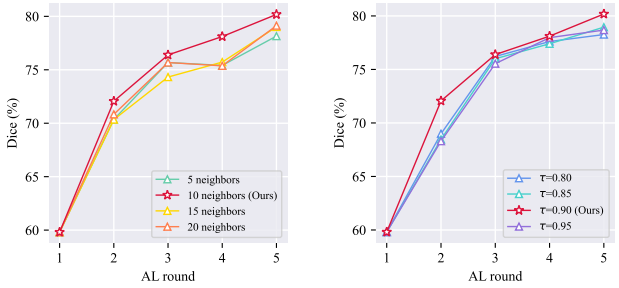


Figure 12. Visualization of aleatoric uncertainty on Fed-Polyp. U^G_{ale} and U^L_{ale} denote the aleatoric uncertainty in global and local models, respectively.

Effect of diversity relaxation. An ablation study was carried out on the Fed-Polyp dataset to investigate the impact of diversity relaxation. As illustrated in Fig. 13, the optimal performance is attained when setting the minimum neighbor size to $n = 10$ and the cosine similarity threshold to $\tau = 0.90$ on the Fed-Polyp dataset.



(a) Minimum neighbor size n. (b) Similarity threshold τ.

Figure 13. Ablation study of diversity relaxation on Fed-Polyp.

Effect of evidential model training. We conducted experiments to compare the evidential loss ($\mathcal{L}$ in Eq. 9) against the cross-entropy loss (CE) on two classification datasets and against dice loss (Dice) on three segmentation datasets. As summarized in Tab. 8, the proposed evidential model training yields an average performance gain of 1.03% on the Fed-ISIC dataset, 0.29% on the Fed-Camelyon dataset, 1.16% on the Fed-Polyp dataset, 1.17% on the Fed-Prostate dataset, and 0.36% on the Fed-Fundus dataset.

Table 8. Ablation study of loss function.

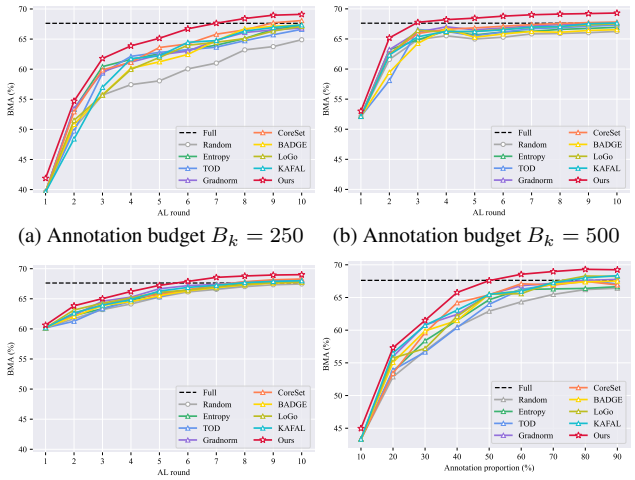
Dataset	Loss	Round 2	Round 3	Round 4	Round 5
Fed-ISIC	CE	$64.28_{\pm 1.64}$	$66.69_{\pm 0.95}$	$67.32_{\pm 1.16}$	$67.40_{\pm 0.22}$
	$\mathcal{L}$	$65.18_{\pm 0.41}$	$67.77_{\pm 1.31}$	$68.41_{\pm 1.01}$	$68.46_{\pm 0.37}$
Fed-Camelyon	CE	$95.24_{\pm 0.03}$	$96.21_{\pm 0.04}$	$96.80_{\pm 0.07}$	$97.26_{\pm 0.06}$
	$\mathcal{L}$	$95.79_{\pm 0.17}$	$96.54_{\pm 0.08}$	$97.04_{\pm 0.02}$	$97.29_{\pm 0.02}$
Fed-Polyp	Dice	$70.14_{\pm 0.10}$	$75.77_{\pm 0.67}$	$77.23_{\pm 0.21}$	$79.48_{\pm 0.62}$
	$\mathcal{L}$	$72.06_{\pm 0.72}$	$76.39_{\pm 0.66}$	$78.62_{\pm 1.44}$	$80.18_{\pm 0.10}$
Fed-Prostate	Dice	$81.43_{\pm 0.75}$	$84.50_{\pm 1.02}$	$85.32_{\pm 0.60}$	$86.50_{\pm 0.59}$
	$\mathcal{L}$	$82.94_{\pm 0.04}$	$85.29_{\pm 0.31}$	$86.77_{\pm 0.29}$	$87.42_{\pm 0.21}$
Fed-Fundus	Dice	$88.11_{\pm 0.3}$	$89.68_{\pm 0.22}$	$89.84_{\pm 0.23}$	$90.15_{\pm 0.19}$
	$\mathcal{L}$	$88.63_{\pm 0.3}$	$89.72_{\pm 0.39}$	$90.28_{\pm 0.13}$	$90.58_{\pm 0.02}$

Effect of trade-off weight λ. We additionally conducted experiments to determine the optimal value for the hyper-parameter λ on the Fed-Polyp dataset, choosing from the candidate set $\{1e{-}5, 5e{-}5, 1e{-}4, 5e{-}4, 1e{-}3\}$. The findings, summarized in Tab. 9, reveal the optimal performance is achieved when $\lambda = 1e{-}4$ for the Fed-Polyp dataset.

Table 9. Ablation study of trade-off weight λ on Fed-Polyp.

λ	Round 2	Round 3	Round 4	Round 5
$1e{-}5$	$70.91_{\pm 1.89}$	$76.27_{\pm 0.36}$	$75.98_{\pm 1.09}$	$78.58_{\pm 1.85}$
$5e{-}5$	$70.27_{\pm 2.64}$	$74.68_{\pm 1.59}$	$76.82_{\pm 1.00}$	$78.26_{\pm 0.76}$
$1e{-}4$	$72.06_{\pm 0.72}$	$76.39_{\pm 0.66}$	$78.62_{\pm 1.44}$	$80.18_{\pm 0.10}$
$5e{-}4$	$70.90_{\pm 1.82}$	$75.59_{\pm 2.02}$	$77.63_{\pm 0.84}$	$79.76_{\pm 1.08}$
$1e{-}3$	$69.78_{\pm 1.64}$	$74.54_{\pm 2.56}$	$76.92_{\pm 1.91}$	$77.64_{\pm 0.88}$

Effect of annotation budget B_k. **(1) Fixed-number.** To validate the effectiveness and robustness of FEAL, we further analyzed the impact of annotation budget B_k on the Fed-ISIC dataset. We conducted $R = 10$ rounds of FAL with three configurations of B_k, *i.e.* $B_k = \{250, 500, 750\}$, for the Fed-ISIC dataset. As depicted in Fig. 14(a)-(c), FEAL consistently outperforms other counterparts across all three annotation budget configurations B_k, demonstrating its effectiveness and resilience. **(2) Fixed-ratio.** Furthermore, considering the imbalanced dataset sizes among local clients, we also implemented the fixed-ratio strategy to annotate 10% of the samples for each client in every FAL round. The results depicted in Fig. 14(d) demonstrate that FEAL surpasses other competing methods even under the fixed-ratio setting. Remarkably, both fixed-number and fixed-ratio strategies exhibit similar performance in later rounds, indicating that FEAL is robust against variations in dataset sizes across local clients.



(a) Annotation budget $B_k = 250$ (b) Annotation budget $B_k = 500$

(c) Annotation budget $B_k = 750$ (d) Annotation budget $B_k = \frac{N_k}{10}$

Figure 14. Ablation study of annotation budget B_k on Fed-ISIC.

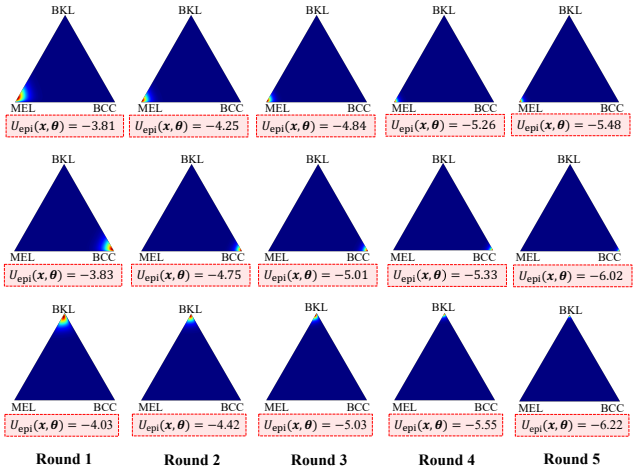


Figure 15. Visualization of the Dirichlet simplex for unlabeled samples across five FAL rounds using FEAL. The unlabeled samples are predicted to belong to MEL, BCC, and BKL from the first to the third rows, respectively. Please zoom in for details.

Analysis of Dirichlet simplex. We analyzed the Dirichlet simplex on a subset of the Fed-ISIC, specifically encompassing three classes: MEL, BCC, and BKL. Within the Dirichlet simplex, a concentrated red region indicates low epistemic uncertainty U_{epi}, while a red region near the corner suggests a low aleatoric uncertainty U_{ale}. As depicted in Fig. 15 and Fig. 16, when selecting samples with FEAL, the Dirichlet distribution becomes narrower and more concentrated for unlabeled local data from the first to the fifth FAL round. This trend suggests a reduction in epistemic uncertainty within the global model, validating the effectiveness of calibrated evidential sampling in mitigating domain shifts. Moreover, starting with an identical set of labeled samples, we tracked the selection of samples in the second FAL round utilizing multiple FAL methods. The resulting Dirichlet simplexes, corresponding to these different methods, are depicted in Fig. 17. A critical observation from this analysis is that the Dirichlet distribution of samples selected via FEAL exhibits a notably broader spread across the simplex. This broader spread indicates that FEAL effectively models the global model's understanding of local data and prioritizes the selection of samples characterized by high epistemic uncertainty.

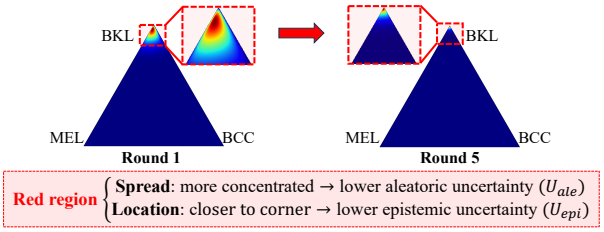


Figure 16. Comparison of the Dirichlet simplex for unlabeled samples in the first and fifth FAL round using FEAL.

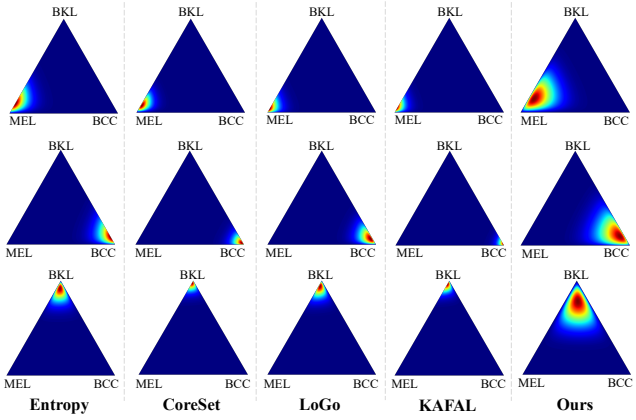


Figure 17. Visualization of the Dirichlet simplex for samples selected in the second FAL round using multiple sampling strategies. The unlabeled samples are predicted to belong to MEL, BCC, and BKL from the first to the third rows, respectively.

C.4. Evaluation Time Costs

All experiments were conducted using a NVIDIA GeForce RTX 2080Ti GPU. The average time cost for one round of data selection across all clients is presented in Tab. 10. Note that we reported the time cost for the ensemble settings of Entropy, CoreSet, TOD, Gradnorm, and BADGE.

Table 10. Time cost (in seconds) for one round of data selection.

Method	Fed-ISIC	Fed-Camelyon	Fed-Polyp	Fed-Prostate	Fed-Fundus
Entropy	23.49	189.97	8.11	14.20	8.39
CoreSet	81.27	340.24	9.68	14.56	8.30
TOD	23.26	331.01	14.50	55.71	28.53
Gradnorm	1126.95	5335.42	72.49	128.49	67.42
BADGE	24.51	178.66	68.12	119.74	69.17
LoGo	96.22	378.81	86.52	80.21	43.20
KAFAL	23.27	175.95	14.54	13.35	7.45
FEAL (Ours)	21.76	191.71	14.88	13.59	13.98

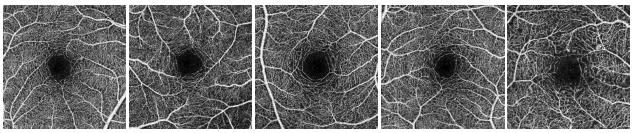
D. Investigations on OCTA Datasets

D.1. Experimental Settings

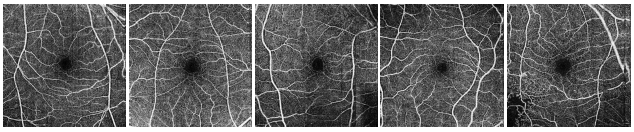
Dataset. We further validated the effectiveness of FEAL on another medical image dataset OCTA-500 [18] for foveal avascular zone (FAZ) segmentation. The OCTA-500 dataset comprises two subsets, namely OCTA_3M and OCTA_6M, each providing a distinct field of view (FOV). Specifically, the field of views (FOV) for OCTA_3M is $3mm \times 3mm \times 2mm$, while that for OCTA_6M is $6mm \times 6mm \times 2mm$. In our study, we regarded each subset as an individual local dataset within federated scenarios and then divided each local dataset into training and test sets using an 8:2 ratio. Note that we utilized the projection maps between the Internal Limiting Membrane (ILM) layer and the Outer Plexiform Layer (OPL) in experiments. Details of the OCTA-500 dataset are provided in Tab. 11. Illustrative samples from both data sources within the OCTA-500 dataset are shown in Fig. 18.

Dataset	Data source	# Train	# Test	Resolution
OCTA-500	Client 1: OCTA_3M [18]	160	40	304×304
	Client 2: OCTA_6M [18]	240	60	400×400

Table 11. Details of multi-center datasets utilized in our study.



(a) Client1: OCTA_3M



(b) Client2: OCTA_6M

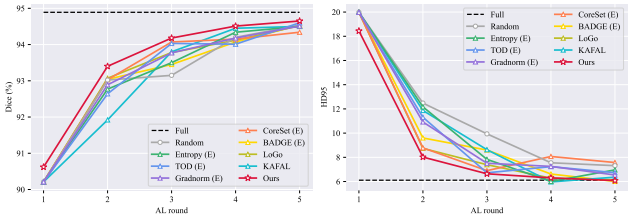
Figure 18. Illustrative samples from each data source within the OCTA-500 dataset.

Evaluation metrics. In line with the aforementioned three image segmentation datasets, we employed the Dice score and the 95% Hausdorff Distance (HD95) as metrics to quantify segmentation results.

Implemental details. We conducted $R = 5$ rounds of FAL, which comprises federated model training and data annotation. For federated model training, We adopted U-Net [20, 26] as the backbone and employed the $ReLU(\cdot)$ as the non-negative activation function $\mathcal{A}(\cdot)$ for both global and local models. We trained local models using the Adam optimizer [14] with a learning rate of $5e{-}4$ and a weight decay of $1e{-}5$. The federated learning process comprises $T = 100$ rounds of communication to attain a robust global model, with each local training session lasting for 1 epochs. Regarding data annotation, the annotation budget B_k was set to 20 for the OCTA-500 dataset.

D.2. Results

We compared FEAL with eight state-of-the-art FAL approaches and present the results in Fig. 19. The results in Fig. 19 verify the effectiveness of FEAL on the OCTA-500 dataset, characterized by superior Dice scores and lower HD95 metrics. It is noteworthy that FEAL achieves a Dice score of 94.18% while utilizing only 24% of annotated samples, which is equivalent to 99.25% of the fully supervised performance.



(a) Dice of OCTA-500 (E) (b) HD95 of OCTA-500 (E)

Figure 19. Comparison results on the OCTA-500 dataset.

E. Discussion with EDL and DUC

Discussion with EDL. While Sensoy *et al.* [28] leveraged evidential deep learning (EDL) to quantify the overall uncertainty of training samples and alleviate model overconfidence, our FEAL differs significantly in both objectives and approaches. Specifically, we employed evidential deep learning to decompose the overall uncertainty into aleatoric and epistemic components, aiming to measure the uncertainty of unlabeled samples and reduce annotation costs in federated learning scenarios with domain shifts.

Discussion with DUC. Although DUC [37] employs evidential deep learning to differentiate between aleatoric uncertainty and epistemic uncertainty, there are notable differences in both objectives and methods compared to

our FEAL. In terms of objectives, DUC aims to enhance domain adaptation by annotating partial samples from the target domain, whereas FEAL focuses on reducing annotation costs for local clients in realistic medical federated scenarios. Regarding informativeness measurement, DUC evaluates the data informativeness based on a model trained on the source domain, whereas FEAL quantifies data informativeness leveraging both global and local models in federated scenarios. Furthermore, DUC is an uncertainty-based method that overlooks the diversity of selected target samples. By contrast, FEAL identifies and annotates samples considering both uncertainty and diversity measures.

References

[1] 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. 4, 5

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Trans. Med. Imaging.*, 38(2):550–560, 2018. 3

[3] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imag. Grap.*, 43:99–111, 2015. 3

[4] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015. 3

[5] 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. 4, 5

[6] 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. 4

[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. 4

[8] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *CBMS*, pages 1–6. IEEE, 2011. 3

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 4

[10] 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. 4, 5

[11] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MMM*, pages 451–462, 2020. 3

[12] 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. 4

[13] 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. 4, 5

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 9

[15] 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. 4

[16] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Comput. Biol. Med.*, 60:8–31, 2015. 3

[17] 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. 3

[18] Mingchao Li, Kun Huang, Qiuzhuo Xu, Jiadong Yang, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Octa-500: a retinal dataset for optical coherence tomography angiography study. *Med. Image. Anal.*, 93:103092, 2024. 9

[19] Geert Litjens, Robert Toth, Wendy Van De Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram Van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Med. Image. Anal.*, 18(2):359–373, 2014. 3

[20] 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. 4, 9

[21] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *NeurIPS*, 31, 2018. 2

[22] Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. Dirichlet and related distributions: Theory, methods and applications. 2011. 2

[23] 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. 3

[24] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluat-

ing automated methods for glaucoma assessment from fundus photographs. *Med. Image. Anal.*, 59:101570, 2020. 3

[25] Deep Shankar Pandey and Qi Yu. Learn to accumulate evidence from all training samples: Theory and practice. In *ICML*, pages 26963–26989, 2023. 2

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 4, 9

[27] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 4, 5

[28] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *NeurIPS*, 31, 2018. 1, 2, 9

[29] Claude Elwood Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948. 2, 4, 5

[30] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.*, 9:283–293, 2014. 3

[31] Jayanthi Sivaswamy, S Krishnadas, Arunava Chakravarty, G Joshi, A Syed Tabish, et al. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1): 1004, 2015. 3

[32] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging.*, 35(2):630–644, 2015. 3

[33] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019. 4

[34] Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *ECCV*, pages 456–472, 2022. 4

[35] 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. 4, 5

[36] 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. 4

[37] Mixue Xie, Shuang Li, Rui Zhang, and Chi Harold Liu. Dirichlet-based uncertainty calibration for active domain adaptation. *arXiv preprint arXiv:2302.13824*, 2023. 1, 9