FedEvi: Improving Federated Medical Image Segmentation via Evidential Weight Aggregation

Jiayi Chen[1]*, Benteng Ma[2]*, Hengfei Cui[1], and Yong Xia[1,3,4](✉)

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
yxia@nwpu.edu.cn
[2] Hong Kong University of Science and Technology, Hong Kong SAR, China
[3] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China
[4] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China

Abstract. Federated learning enables collaborative knowledge acquisition among clinical institutions while preserving data privacy. However, feature heterogeneity across institutions can compromise the global model's performance and generalization capability. Existing methods often adjust aggregation weights dynamically to improve the global model's generalization but rely heavily on the local models' performance or reliability, excluding an explicit measure of the generalization gap arising from deploying the global model across varied local datasets. To address this issue, we propose FedEvi, a method that adjusts the aggregation weights based on the generalization gap between the global model and each local dataset and the reliability of local models. We utilize a Dirichlet-based evidential model to disentangle the uncertainty representation of each local model and the global model into epistemic uncertainty and aleatoric uncertainty. Then, we quantify the global generalization gap using the epistemic uncertainty of the global model and assess the reliability of each local model using its aleatoric uncertainty. Afterward, we design aggregation weights using the global generalization gap and local reliability. Comprehensive experimentation reveals that FedEvi consistently surpasses 12 state-of-the-art methods across three real-world multi-center medical image segmentation tasks, demonstrating the effectiveness of FedEvi in bolstering the generalization capacity of the global model in heterogeneous federated scenarios. The code will be available at https://github.com/JiayiChen815/FedEvi.

Keywords: Federated learning · Uncertainty estimation · Medical image segmentation

* J. Chen and B. Ma contributed equally.

1 Introduction

Federated learning (FL) holds the promise of collaborative learning across multiple clinical institutions (*i.e.*, clients) to develop a unified global model on a server through model aggregation while preserving the data privacy of each client [22,33]. However, inevitable variations in scanner vendors, imaging protocols, and other factors usually result in divergent data distributions across clients. This divergence, arguably, undermines the generalization capability of the global model in real-world federated applications of medicine [10,37,20].

Existing mitigation strategies for such distribution shifts mainly fall into two categories, namely, regularization-based and aggregation-based methods. **Regularization-based methods** aim to soften the impact of local training drift by constraining parameter differences [14], feature embeddings [32,35], flatness of loss landscapes [24], or prediction consistency [8,18]. However, these methods employ fixed aggregation weights when forming the global model from local ones [38], potentially ignoring the contribution of clients with significantly heterogeneous data [10]. This oversight, unfortunately, weakens the global model's generalization capability. To address this issue, **aggregation-based methods** dynamically adjust aggregation weights based on client contribution estimation [10], performance of proxy or validation dataset [16,19,39], performance enhancement through aggregation [38], parameter divergence [25], model affinity [5], or local uncertainty estimation [34]. These methods predominantly derive aggregation weights solely from local performance metrics [10,16,38,39], excluding the explicit quantification of the generalization gap that emerges when deploying the global model across heterogeneous local datasets. This oversight renders them somewhat incapable of addressing the challenges posed by heterogeneous FL scenarios. Moreover, data complexity variations across different clients tend to affect the reliability of locally trained models, a factor vital in the design of aggregation weights. Although FedUAA [34] utilized the confidence of local model predictions to devise an uncertainty-aware weight aggregation strategy, it fails to distinguish between epistemic uncertainty and aleatoric uncertainty, making it challenging to directly measure the reliability of the local model on its local dataset.

In this paper, we propose a novel method, termed FedEvi, to adjust aggregation weights by considering not only the generalization gap between the global model and local datasets but also the reliability of local models. FedEvi has three key components: a Dirichlet-based evidential model, an evidential weight aggregation strategy, and an evidential model training scheme. It leverages a Dirichlet-based evidential model to separate overall uncertainty into epistemic and aleatoric uncertainties. With the evidential weight aggregation strategy, we first establish a surrogate global model derived from trained local models and previous aggregation weights. Then, we quantify the generalization gap by the epistemic uncertainty within the surrogate global model on local datasets. Next, we measure the local reliability by assessing aleatoric uncertainty in local models. FedEvi, therefore, increases the aggregation weights of the clients characterized by substantial generalization gaps in the global model and high reliability in

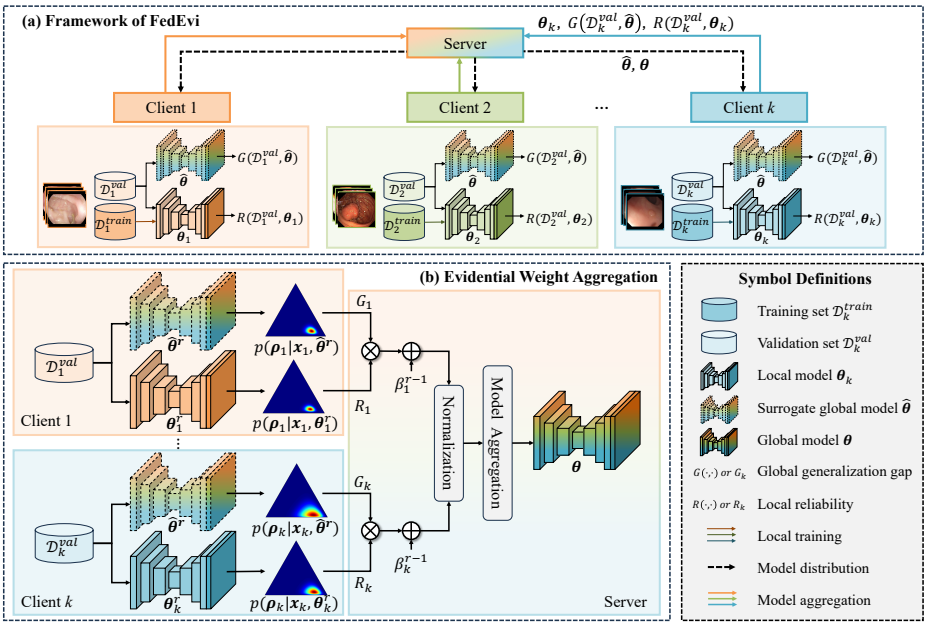

Fig. 1. Illustration of FedEvi. (a) Framework of FedEvi. (b) Evidential weight aggregation strategy. It dynamically adjusts aggregation weights based on both the global generalization gap $G(\mathcal{D}_k^{val}, \hat{\boldsymbol{\theta}})$ and local reliability $R(\mathcal{D}_k^{val}, \boldsymbol{\theta}_k)$.

their local models. Furthermore, FedEvi integrates a regularization loss to enhance local reliability.

The main contributions are three-fold. (1) We employ a Dirichlet-based evidential model to disentangle overall uncertainty into epistemic and aleatoric components, providing a detailed uncertainty representation. (2) We propose FedEvi, a novel aggregation-based FL method, based on global generalization gap and local reliability. The global generalization gap is measured by epistemic uncertainty within the surrogate global model, while the local reliability is assessed via aleatoric uncertainty within local models. (3) FedEvi outperforms 12 SOTA methods on three real multi-center medical image segmentation datasets.

2 Methodology

2.1 Overview

As depicted in Fig. 1, we consider the FL framework that involves K local models $\{\boldsymbol{\theta}_k\}_{k=1}^K$ on clients and a global model $\boldsymbol{\theta}$ on the server. The k-th client maintains a local dataset $\mathcal{D}_k = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N_k}$, where $N_k = |\mathcal{D}_k|$ is the total number of samples in the k-th client. FedEvi comprises three key components: (1) a Dirichlet-based evidential model for decoupling total uncertainty into epistemic uncertainty and

aleatoric uncertainty (Sec. 2.2); (2) an evidential weight aggregation strategy (Sec. 2.3) involving both global generalization and local reliability; and (3) an evidential model training scheme (Sec. 2.4) to improve local reliability. FedEvi conducts R federated rounds, each comprising E epochs of local training. The algorithm of FedEvi is presented in Alg. A1. We now delve into its details.

2.2 Decoupled Uncertainty in Dirichlet-based Evidential Model

In the C-class segmentation task, given a sample $\boldsymbol{x}$, the segmentation model f parameterized with $\boldsymbol{\theta}$ maps $\boldsymbol{x}$ into C-dimensional logits $f(\boldsymbol{\theta}, \boldsymbol{x})$ for each pixel. Dirichlet-based evidential model treats the categorical prediction $\boldsymbol{\rho}$ as a random variable following a Dirichlet distribution $Dir(\boldsymbol{\rho}|\boldsymbol{\alpha})$, which allows multiple potential predictions for a sample and enables a decoupled representation of uncertainty. The probability density function of $\boldsymbol{\rho}$ [21,36] is formulated as:

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

where $\Gamma(\cdot)$ denotes the Gamma function and $\boldsymbol{\alpha}$ is the parameter of the Dirichlet distribution for sample $\boldsymbol{x}$. The Dirichlet paramter $\boldsymbol{\alpha}$ can be expressed as $\boldsymbol{\alpha} = \boldsymbol{e} + 1 = \mathcal{A}(f(\boldsymbol{\theta}, \boldsymbol{x})) + 1$, where $\boldsymbol{e}$ is the evidence quantifiing the support for model predictions [27]. $\mathcal{A}(\cdot)$ denotes a non-negative activation and we utilized the exponential activation $exp(\cdot)$ in experiments.

Therefore, the expected probability of class c can be denoted as:

$$P(y = c|\boldsymbol{x}, \boldsymbol{\theta}) = \int p(y = c|\boldsymbol{\rho}) \cdot p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\rho} = \frac{\alpha_c}{\sum_{j=1}^{C} \alpha_j} = \overline{\rho}_c. \tag{2}$$

The total uncertainty $U_{total}(\boldsymbol{x}, \boldsymbol{\theta})$ is quantified as the Shannon entropy $\mathcal{H}(\cdot)$ of the expected probability $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ [21,28], which can be decoupled into epistemic uncertainty $U_{epi}(\boldsymbol{x}, \boldsymbol{\theta})$ and aleatoric uncertainty $U_{ale}(\boldsymbol{x}, \boldsymbol{\theta})$ [6], as follows:

$$\underbrace{\mathcal{H}[P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})]}_{U_{total}} = \underbrace{\mathcal{I}[\boldsymbol{y}, \boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}]}_{U_{epi}} + \underbrace{\mathbb{E}_{p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta})}[\mathcal{H}[P(\boldsymbol{y}|\boldsymbol{\rho})]}_{U_{ale}}, \tag{3}$$

$$U_{epi}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{c=1}^{C} \overline{\rho}_c [\psi(\alpha_c + 1) - \psi(\sum_{j=1}^{C} \alpha_j + 1)] + \sum_{c=1}^{C} \overline{\rho}_c \log \overline{\rho}_c, \tag{4}$$

$$U_{ale}(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{c=1}^{C} \overline{\rho}_c [\psi(\sum_{j=1}^{C} \alpha_j + 1) - \psi(\alpha_c + 1)], \tag{5}$$

where $\psi(\cdot)$ is the digamma function.

Epistemic uncertainty captures a model's limited knowledge about data due to distribution shifts [4,36]. Thus, high epistemic uncertainty within the global model indicates a large generalization gap between it and local datasets. Aleatoric uncertainty evaluates the inherent data complexity and model reliability. Therefore, low aleatoric uncertainty in local models reflects high local reliability.

2.3 Evidential Weight Aggregation

Generalization Gap of Surrogate Global Model. In the r-th federated round, we obtain K trained local models $\{\boldsymbol{\theta}_k^r\}_{k=1}^K$. Given the aggregation weights $\{\beta_k^{r-1}\}_{k=1}^K$ in the previous round, the surrogate global model is calculated as $\hat{\boldsymbol{\theta}}^r = \sum_{k=1}^K \beta_k^{r-1}\boldsymbol{\theta}_k^r$. We then utilize epistemic uncertainty (Eq. 4) within the surrogate global model on each local validation dataset $\mathcal{D}_k^{val}$ to quantify the generalization gap between the surrogate global model and local dataset as follows:

$$G(\mathcal{D}_k^{val}, \hat{\boldsymbol{\theta}}^r) = \frac{1}{N_k^{val}} \sum_{i=1}^{N_k^{val}} U_{epi}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^r), \quad (\boldsymbol{x}_i \in \mathcal{D}_k^{val}), \tag{6}$$

Reliability of Local Model. We assessed the reliability of the trained local model $\boldsymbol{\theta}_k^r$ by calculating the average reciprocal of aleatoric uncertainty (Eq. 5) on the local validation set $\mathcal{D}_k^{val}$, formulated as follows:

$$R(\mathcal{D}_k^{val}, \boldsymbol{\theta}_k^r) = \frac{1}{N_k^{val}} \sum_{i=1}^{N_k^{val}} \frac{1}{U_{ale}(\boldsymbol{x}_i, \boldsymbol{\theta}_k^r)}, \quad (\boldsymbol{x}_i \in \mathcal{D}_k^{val}). \tag{7}$$

Aggregation Weight. In heterogeneous FL, the server should increase aggregation weights for clients where the global model exhibits inadequate generalization. Simultaneously, the adjustment necessitates careful consideration of local reliability to ensure the robustness of the global model. Hence, given the estimated generalization gap of the surrogate global model and the reliability of local models, FedEvi adjusts the aggregation weight as follows:

$$\beta_k^r = \beta_k^{r-1} + \delta \cdot G(\mathcal{D}_k^{val}, \hat{\boldsymbol{\theta}}^r) \cdot R(\mathcal{D}_k^{val}, \boldsymbol{\theta}_k^r) \tag{8}$$

where δ denotes the magnitude of weight adjustment. Noteably, we set $\beta_k^0 = \frac{N_k}{\sum_{k=1}^K N_k}$. Subsequently, the aggregation weight is normalized as $\beta_k^r = \frac{\beta_k^r}{\sum_{k=1}^K \beta_k^r}$.

2.4 Evidential Model Training (EMT)

The loss function of FedEvi comprises two terms, *i.e.*, the target loss $\mathcal{L}_{dice}$ for segmentation and the regularization loss $\mathcal{L}_{reg}$ to enhance local reliability. Specifically, we followed [13] to employ the Bayes risk of Dice loss as the target loss, formulated as follows:

$$\mathcal{L}_{dice}(\boldsymbol{\theta}_k, \mathcal{D}_k^{train}) = \int (1 - \frac{2}{C}\sum_{c=1}^{C} \frac{|\boldsymbol{y}_c \cdot \boldsymbol{\rho}_c|}{|\boldsymbol{y}_c^2| + |\boldsymbol{\rho}_c^2|}) \cdot p(\boldsymbol{\rho}|\boldsymbol{x}, \boldsymbol{\theta}_k)\ \mathrm{d}\boldsymbol{\rho}. \tag{9}$$

To enhance local reliability, we mitigate the evidence $\boldsymbol{e}$ of incorrect predictions by reducing the corresponding Dirichlet parameter $\tilde{\boldsymbol{\alpha}}$ to its minimum $\mathbf{1}$ as follows:

$$\mathcal{L}_{reg}(\boldsymbol{\theta}_k, \mathcal{D}_k^{train}) = KL[Dir(\boldsymbol{\rho}|\tilde{\boldsymbol{\alpha}})\|Dir(\boldsymbol{\rho}|\mathbf{1})], \tag{10}$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (\mathbf{1} - \boldsymbol{y}) \odot \boldsymbol{\alpha}$ and $KL(\cdot)$ denotes the KL divergence [11].

The overall loss function for the k-th client is:

$$\mathcal{L}(\boldsymbol{\theta}_k, \mathcal{D}_k^{train}) = \mathcal{L}_{dice}(\boldsymbol{\theta}_k, \mathcal{D}_k^{train}) + \lambda \cdot \mathcal{L}_{reg}(\boldsymbol{\theta}_k, \mathcal{D}_k^{train}). \tag{11}$$

3 Experiments

3.1 Experimental Settings

Datasets and Evaluation Metrics. We evaluated FedEvi on three real-world multi-center medical image segmentation datasets. (1) The endoscopic polyp dataset is collected from 4 centers [9,29,31,2] for polyp segmentation. (2) The prostate MRI dataset is gathered from 6 medical centers [17,3,12] for prostate segmentation. (3) The retinal fundus dataset is sourced from 6 centers [30,7,23,1] for joint segmentation of the optic disc and optic cup. Each center was treated as a local client, and the data was partitioned randomly into training, validation, and test sets with a ratio of 70%/10%/20% at the patient level for each client. All images were resized to the resolution of 384×384 pixels following [18]. Details of these datasets including data sources and sample sizes are summarized in Tab. A5. We employed the Dice coefficient (Dice) and the 95% Hausdorff Distance (HD95) to evaluate segmentation results quantitatively.

Implemental Details. We utilized 2D U-Net [26] as the backbone following [10,18] and employed random flipping for data augmentation. We trained local models using the Adam optimizer with a learning rate of $1e{-}3$, betas of $(0.9, 0.99)$, and a weight decay of $1e{-}5$. We conducted $T = 200$ federated rounds, with the local training epochs set to $E = 2$. The weight adjustment magnitude δ is defaulted to 1.0 across all three datasets. Meanwhile, the trade-off weight λ is configured to $1e{-}2$ for both endoscopic polyp and prostate MRI datasets and $1e{-}5$ for the retinal fundus dataset. Each experiment was run three times with different random seeds, and the average results were reported.

Comparison Methods. We compared our FedEvi against 12 state-of-the-art FL methods, including: (1) the baseline FedAvg (AISTATS17) [22], (2) five regularization-based methods: FedProx (MLSys20) [14], FedDG (CVPR21) [18], FedProto (AAAI22) [32], FedSAM (ICML22) [24], and FedBR (ICLR23) [8], and (3) six aggregation-based methods: FedBN (ICLR21) [15], FedLAW (ICML23) [16], FedCE (CVPR23) [10], FedGA (CVPR23) [38], L-DAWA (ICCV23) [25], and FedUAA (MICCAI23) [34].

3.2 Comparison with SOTA Methods

The results for endoscopic polyp, prostate MRI, and retinal fundus segmentation are summarized in Tab. 1, Tab. 2, and Tab. 3, respectively. We can see that across all three datasets, both regularization-based and aggregation-based methods exhibited improvements in average performance compared to FedAvg. Compared with these methods, FedEvi achieved higher average performance and obtained improvement on most clients in terms of Dice and HD95 metrics for prostate MRI segmentation. This benefits from the consideration of both the global generalization gap and local reliability. For endoscopic polyp segmentation, FedEvi improved the average Dice by 9.81% and 3.68% and reduced the average HD95 by 16.31 and 8.1 pixels compared to the baseline FedAvg and the second-best method FedUAA, respectively. Visualization comparisons for these three datasets are provided in Fig. A1, Fig. A2, and Fig. A3, respectively.

Table 1. Performance comparison of FedEvi and 12 competing methods for endoscopic polyp segmentation. For each evaluation metric, we present the performance of each client C_k ($k \in [4]$), along with the average performance (Avg) and the standard deviation (Std) across clients. The best results are highlighted in **bold**.

Method	Dice (%)						HD95					
	C_1	C_2	C_3	C_4	Avg↑	Std↓	C_1	C_2	C_3	C_4	Avg↓	Std↓
FedAvg [22]	85.02	62.35	71.06	88.92	76.84	12.46	38.54	51.22	49.99	20.73	40.12	14.50
FedProx [14]	84.94	68.34	72.06	87.00	78.08	9.27	37.27	44.00	49.56	23.51	38.58	11.33
FedDG [18]	85.44	68.74	73.77	88.66	79.15	9.52	37.51	44.11	47.86	21.68	37.79	11.67
FedProto [32]	85.45	64.58	71.56	88.36	77.49	11.36	38.22	54.26	48.60	21.06	40.53	15.08
FedSAM [24]	85.49	68.48	73.31	88.86	79.04	9.80	37.97	44.10	44.50	20.49	36.77	11.53
FedBR [8]	85.23	66.06	71.65	88.20	77.78	10.67	38.22	52.13	49.00	21.61	40.24	13.91
FedBN [15]	85.66	74.41	76.17	90.44	81.67	7.67	38.71	41.20	42.38	17.32	34.90	13.39
FedLAW [16]	83.26	63.89	71.85	88.92	76.98	11.31	42.23	54.02	48.68	20.78	41.43	14.87
FedCE [10]	85.11	75.79	76.76	90.19	81.96	6.91	37.40	31.89	40.63	18.63	32.14	**10.01**
FedGA [38]	82.55	83.17	67.54	90.44	80.93	9.63	41.33	25.27	55.97	17.87	35.11	17.03
L-DAWA [25]	78.80	66.56	68.00	85.52	74.72	9.14	48.92	43.99	57.12	27.22	44.31	13.07
FedUAA [34]	85.20	79.97	76.46	90.24	82.97	6.26	37.86	30.63	40.53	18.63	31.91	11.08
FedEvi (Ours)	**85.73**	**87.92**	**81.11**	**91.83**	**86.65**	**4.51**	**36.12**	**11.88**	**31.79**	**15.44**	**23.81**	11.98

Table 2. Performance comparison of FedEvi and 12 competing methods for prostate MRI segmentation. The best results are highlighted in **bold**.

Method	Dice (%)								HD95							
	C_1	C_2	C_3	C_4	C_5	C_6	Avg↑	Std↓	C_1	C_2	C_3	C_4	C_5	C_6	Avg↓	Std↓
FedAvg [22]	80.25	88.26	90.69	83.16	87.07	85.17	85.77	3.83	16.12	9.64	5.80	14.06	10.17	**8.60**	10.73	3.80
FedProx [14]	84.80	88.33	90.67	85.03	88.02	84.91	86.96	2.48	11.04	9.59	5.87	12.74	8.94	9.11	9.55	2.37
FedDG [18]	86.71	88.01	89.31	84.74	86.56	83.60	86.49	2.13	10.34	9.97	6.74	12.78	10.39	11.39	10.27	2.13
FedProto [32]	85.35	87.73	89.86	85.94	87.94	85.10	86.99	1.98	12.10	10.05	6.05	12.76	8.87	9.23	9.84	2.61
FedSAM [24]	83.01	88.40	90.19	82.10	88.01	84.86	86.09	3.37	13.33	9.65	5.98	15.27	9.11	9.24	10.43	3.37
FedBR [8]	86.08	88.47	91.03	83.50	88.30	82.31	86.62	3.34	11.67	9.63	6.25	14.58	9.24	10.32	10.28	2.82
FedBN [15]	83.06	88.43	91.00	87.72	87.77	82.51	86.75	3.35	15.81	9.77	5.63	10.65	9.59	12.09	10.59	3.44
FedLAW [16]	84.43	88.38	91.10	84.41	87.09	84.95	86.73	2.71	13.05	9.74	5.49	13.02	9.89	8.97	10.03	2.94
FedCE [10]	83.31	88.71	**91.55**	84.18	88.40	82.71	86.48	3.89	12.95	9.32	**5.35**	13.90	8.56	11.79	10.31	3.50
FedGA [38]	85.84	87.86	89.91	81.26	84.93	83.61	85.57	3.16	11.59	10.29	6.64	16.36	11.77	11.85	11.42	3.26
L-DAWA [25]	83.94	87.79	89.66	79.92	87.03	**86.55**	85.82	3.46	13.34	10.07	6.95	18.26	9.93	8.64	11.20	4.12
FedUAA [34]	85.53	**88.77**	91.24	82.28	88.15	84.10	86.68	3.35	11.95	**9.25**	5.46	15.95	8.77	9.31	10.12	3.57
FedEvi (Ours)	**87.09**	88.52	90.23	**88.40**	**89.11**	85.25	**88.10**	**1.80**	**9.61**	9.39	6.02	**10.01**	**7.65**	8.35	**8.51**	**1.51**

Table 3. Performance comparison of FedEvi and 12 competing methods for retinal fundus segmentation. The best results are highlighted in **bold**.

Method	Dice (%)								HD95							
	C_1	C_2	C_3	C_4	C_5	C_6	Avg↑	Std↓	C_1	C_2	C_3	C_4	C_5	C_6	Avg↓	Std↓
FedAvg [22]	88.23	73.57	90.60	92.03	90.84	92.21	87.91	7.19	27.38	33.92	8.84	6.14	7.11	5.34	14.79	12.58
FedProx [14]	91.48	77.79	90.85	92.10	91.02	92.29	89.25	5.66	16.61	27.28	8.53	5.96	7.09	5.29	11.79	8.80
FedDG [18]	91.33	82.91	91.33	92.24	90.09	91.60	89.92	3.52	21.06	17.79	8.55	5.90	7.71	5.87	11.15	6.59
FedProto [32]	87.73	73.05	91.11	92.25	90.66	91.97	87.79	7.42	28.35	33.85	8.22	5.56	7.29	5.56	14.80	13.01
FedSAM [24]	89.51	75.87	91.01	92.36	90.70	91.76	88.54	6.30	20.96	27.54	8.28	5.54	7.04	5.44	12.47	9.55
FedBR [8]	87.09	74.17	90.00	89.36	89.03	91.74	86.90	6.47	24.35	30.72	9.70	10.53	9.18	5.87	15.06	10.14
FedBN [15]	89.35	82.99	**92.10**	92.17	89.36	91.94	89.65	3.62	21.45	18.52	**7.42**	5.62	9.94	5.52	11.41	7.33
FedLAW [16]	91.09	80.12	91.24	91.50	90.51	91.83	89.38	4.56	17.05	20.33	8.53	7.47	7.20	5.62	11.03	6.15
FedCE [10]	88.50	74.93	90.85	**92.48**	90.97	91.84	88.26	6.68	25.43	27.56	9.23	**5.57**	7.11	5.50	13.40	10.41
FedGA [38]	89.20	81.94	89.92	89.78	89.08	91.67	88.60	3.42	22.42	19.45	9.05	7.97	7.46	5.64	12.00	7.16
L-DAWA [25]	90.85	79.33	90.14	91.33	**91.03**	92.06	89.12	4.84	20.37	21.10	9.83	6.94	7.21	5.43	11.81	7.09
FedUAA [34]	92.54	79.99	90.51	91.83	90.67	92.37	89.65	4.81	13.59	20.59	8.96	6.24	**6.84**	**5.10**	10.22	6.02
FedEvi (Ours)	**93.30**	**87.44**	90.32	90.83	89.76	**92.48**	**90.69**	**2.09**	**11.21**	**11.62**	8.98	6.62	6.89	5.23	**8.43**	**2.64**

Table 4. Ablation study of components. Excluding all components refers to FedAvg.

Component			Endoscopic Polyp Segmentation						Prostate MRI Segmentation							
EMT	$G(\mathcal{D}_k^{val}, \hat{\boldsymbol{\theta}})$	$R(\mathcal{D}_k^{val}, \boldsymbol{\theta}_k)$	C_1	C_2	C_3	C_4	Avg↑	Std↓	C_1	C_2	C_3	C_4	C_5	C_6	Avg↑	Std↓
✗	✗	✗	85.02	62.35	71.06	88.92	76.84	12.46	80.25	88.26	90.69	83.16	87.07	85.17	85.77	3.83
✓	✗	✗	84.83	68.04	72.45	89.07	78.60	9.99	85.15	88.37	90.46	84.00	88.38	83.76	86.69	2.84
✓	✓	✗	85.94	83.89	80.24	90.53	85.15	4.32	87.10	88.48	90.86	84.80	89.41	85.23	87.65	2.40
✓	✗	✓	83.59	84.93	80.34	91.95	85.20	4.95	84.55	88.38	90.28	84.90	89.70	84.52	87.06	2.78
✓	✓	✓	85.73	87.92	81.11	91.83	86.65	4.51	87.09	88.52	90.23	88.40	89.11	85.25	88.10	1.80

Table 5. Ablation study on initial aggregation weight β_k^0.

β_k^0	Endoscopic Polyp Segmentation (Dice)						Prostate MRI Segmentation (Dice)							
	C_1	C_2	C_3	C_4	Avg↑	Std↓	C_1	C_2	C_3	C_4	C_5	C_6	Avg↑	Std↓
Best Competitor	*85.20*	*79.97*	*76.46*	*90.24*	*82.97*	*6.26*	*85.35*	*87.73*	*89.86*	*85.94*	*87.94*	*85.10*	*86.99*	*1.98*
#Client-Aware	85.53	84.06	80.93	91.41	85.48	4.46	86.14	88.59	90.58	86.11	88.80	84.55	87.46	2.37
#Sample-Aware	85.73	87.92	81.11	91.83	86.65	4.51	87.09	88.52	90.23	88.40	89.11	85.25	88.10	1.80

3.3 Ablation Study

Effect of Components. We analyzed the effect of each component in Tab. 4. 'EMT' denotes training with both Dice loss $\mathcal{L}_{dice}$ and regularization loss $\mathcal{L}_{reg}$. Building upon FedAvg (row 1), integrating $\mathcal{L}_{reg}$ into local training (row 2) mitigates incorrect evidence and enhances local reliability, leading to performance improvements of 1.76% and 0.92% for endoscopic polyp segmentation and prostate MRI segmentation, respectively. Additionally, adjusting aggregation weights by jointly considering the global generalization gap and local reliability yielded superior performance compared to considering each factor individually.

Effect of Initial Aggregation Weight β_k^0. We compared two initial aggregation weights β_k^0: #Client-Aware assigns uniform weights ($\beta_k^0 = \frac{1}{K}$) for clients, while #Sample-Aware allocates weights according to local dataset size ($\beta_k^0 = \frac{N_k}{\sum_{k=1}^{K} N_k}$). In Tab. 5, both settings of β_k^0 surpassed the *Best Competitor* in terms of average performance and standard deviation, indicating the efficacy and robustness of FedEvi. Furthermore, as #Sample-Aware outperformed #Client-Aware, we adopted the #Sample-Aware strategy in experiments.

Effect of Weight Adjustment Magnitude δ. We identified the optimal weight adjustment magnitude δ for the endoscopic polyp dataset from the candidate set of $\delta = \{0.25, 0.5, 1.0, 1.25, 1.5\}$. As depicted in Fig. 2, the peak performance was attained at $\delta = 1.0$.

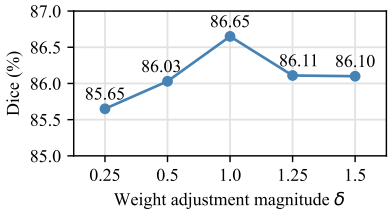

Fig. 2. Effect of hyperparameter δ.

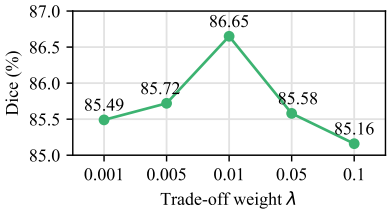

Fig. 3. Effect of trade-off weight λ.

Effect of Trade-Off Weight λ. As illustrated in Fig. 3, we determined the optimal trade-off weight $\lambda = 1e{-}2$ for the endoscopic polyp dataset from the candidate set of $\lambda = \{1e{-}3, 5e{-}3, 1e{-}2, 5e{-}2, 1e{-}1\}$.

4 Conclusion

We proposed a novel method FedEvi to adjust aggregation weights based on global generalization gap and local reliability for heterogeneous FL. FedEvi decomposes uncertainty into epistemic and aleatoric parts using a Dirichlet-based evidential model. It then adjusts aggregation weights by considering the global generalization gap via the surrogate global model's epistemic uncertainty and the local reliability through local models' aleatoric uncertainty. Experiment results demonstrate that FedEvi outperforms other methods on three benchmarks.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grants 62171377, in part by Shenzhen Science and Technology Program under Grants JCYJ20220530161616036, in part by the Ningbo Clinical Research Center for Medical Imaging under Grant 2021L003 (Open Project 2022LYKFZD06), and in part by the Innovation Foundation for Master Dissertation of Northwestern Polytechnical University under Grant PF2024013.

Disclosure of Interests. The authors declare no relevant competing interests.

References

1. Almazroa, A., Alodhayb, S., Osman, E., et al.: Retinal fundus images for glaucoma analysis: the riga dataset. In: Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications. vol. 10579, pp. 55–62 (2018)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., et al.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imag. Grap. **43**, 99–111 (2015)
3. Bloch, N., Madabhushi, A., Huisman, H., et al.: Nci-isbi 2013 challenge: automated segmentation of prostate structures. TCIA **370**(6), 5 (2015)
4. Chen, J., Ma, B., Cui, H., Xia, Y.: Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In: CVPR (2024)
5. Deng, Z., Li, D., Tan, S., et al.: Fedgrav: An adaptive federated aggregation algorithm for multi-institutional medical image segmentation. In: MICCAI (2023)
6. Depeweg, S., Hernández-Lobato, J.M., Doshi-Velez, F., Udluft, S.: Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems. Stat **1050**, 11 (2017)
7. Fumero, F., Alayón, S., Sanchez, J.L., et al.: Rim-one: An open retinal image database for optic nerve evaluation. In: CBMS (2011)
8. Guo, Y., Tang, X., Lin, T.: Fedbr: improving federated learning on heterogeneous data via local learning bias reduction. In: ICML (2023)
9. Jha, D., Smedsrud, P.H., Riegler, M.A., et al.: Kvasir-seg: A segmented polyp dataset. In: MMM (2020)

10. Jiang, M., Roth, H.R., Li, W., et al.: Fair federated medical image segmentation via client contribution estimation. In: CVPR (2023)
11. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)
12. Lemaître, G., Martí, R., Freixenet, J., et al.: Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. Comput. in Bio. and Med. **60**, 8–31 (2015)
13. Li, H., Nan, Y., Del Ser, J., Yang, G.: Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. Neural Comput. Appl. pp. 1–15 (2022)
14. Li, T., Sahu, A.K., et al.: Federated optimization in heterogeneous networks. MLSys **2**, 429–450 (2020)
15. Li, X., Jiang, M., et al.: Fedbn: Federated learning on non-iid features via local batch normalization. In: ICLR (2021)
16. Li, Z., Lin, T., Shang, X., Wu, C.: Revisiting weighted aggregation in federated learning with neural networks. arXiv preprint arXiv:2302.10911 (2023)
17. Litjens, G., Toth, R., Van De Ven, W., et al.: Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. Med. Image. Anal. **18**(2), 359–373 (2014)
18. Liu, Q., Chen, C., Qin, J., et al.: Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: CVPR (2021)
19. Ma, B., Feng, Y., Chen, G., et al.: Federated adaptive reweighting for medical image classification. Pattern Recogn. **144**, 109880 (2023)
20. Ma, B., Zhang, J., Xia, Y., Tao, D.: Vnas: Variational neural architecture search. Int. J. Comput. Vis. pp. 1–25 (2024)
21. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. NeurIPS **31** (2018)
22. McMahan, B., Moore, E., et al.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS (2017)
23. Orlando, J.I., Fu, H., Breda, J.B., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Med. Image. Anal. **59**, 101570 (2020)
24. Qu, Z., Li, X., Duan, R., et al.: Generalized federated learning via sharpness aware minimization. In: ICML (2022)
25. Rehman, Y.A.U., Gao, Y., De Gusmão, P.P.B., et al.: L-dawa: Layer-wise divergence aware weight aggregation in federated self-supervised visual representation learning. In: ICCV. pp. 16464–16473 (2023)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
27. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. NeurIPS **31** (2018)
28. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
29. Silva, J., Histace, A., Romain, O., et al.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. **9**, 283–293 (2014)
30. Sivaswamy, J., Krishnadas, S., Chakravarty, A., 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)

31. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging. **35**(2), 630–644 (2015)
32. Tan, Y., Long, G., Liu, L., et al.: Fedproto: Federated prototype learning across heterogeneous clients. In: AAAI (2022)
33. Wang, J., Jin, Y., Wang, L.: Personalizing federated medical image segmentation via local calibration. In: ECCV (2022)
34. Wang, M., Wang, L., Xu, X., et al.: Federated uncertainty-aware aggregation for fundus diabetic retinopathy staging. arXiv preprint arXiv:2303.13033 (2023)
35. Wu, N., Yu, L., Yang, X., Cheng, K.T., Yan, Z.: Fediic: Towards robust federated learning for class-imbalanced medical image classification. In: MICCAI (2023)
36. Xie, M., Li, S., Zhang, R., Liu, C.H.: Dirichlet-based uncertainty calibration for active domain adaptation. In: ICLR (2023)
37. Zhang, L., Wang, X., Yang, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE Trans. on Med. Imaging. **39**(7), 2531–2540 (2020)
38. Zhang, R., Xu, Q., Yao, J., et al.: Federated domain generalization with generalization adjustment. In: CVPR (2023)
39. Zhou, Q., Zheng, G.: Fedcontrast-gpa: Heterogeneous federated optimization via local contrastive learning and global process-aware aggregation. In: MICCAI (2023)