Fed-FGS: Domain-Generalized Federated Polyp Segmentation via Fourier Hard-Thresholding and Gradient Scaling

Ahmed M. Gaber[1], Noha S. Tawfik[2], Salema Fayed[1], Noha Ghatwary[1]

[1]Computer Engineering Department, Arab Academy for Science and Technology, Smart Village Campus, Giza P.O. Box 2033, Egypt.
[2]Computer Engineering Department, Arab Academy for Science and Technology, Abou Kir, Alexandria P.O. Box 1029, Egypt
noha.ghatwary@aast.edu

Abstract. Federated learning (FL) enables collaborative training of medical image segmentation models without sharing sensitive patient data, making it well-suited for multi-center applications. However, severe data heterogeneity due to differences in imaging devices, acquisition, and protocols leads to domain shifts that degrade the generalization ability of methods. In this paper, we propose *Fed-FGS*, a frequency-aware federated learning framework for robust colorectal polyp segmentation. Fed-FGS combines three complementary strategies: (1) difficulty-aware federated gradient scaling emphasizes clinically challenging small polyps to the global optimization; (2) a Discrete Fourier Transform Hard Thresholding (DFT-HT) augmentation to suppress domain-specific noise while preserving structural semantics; and (3) a composite Dice-BCE-Focal loss to address foreground-background imbalance. We evaluate Fed-FGS on the multi-center PolypGen benchmark using five-fold cross-validation across Centers 1-5 with Center 6 reserved for unseen-center evaluation. The proposed Fed-FGS achieves a Dice score of 0.75±0.02, outperforming the FedGS baseline (0.73 ± 0.01), while demonstrating greater stability under the Small Best selection criterion (0.73 ± 0.01 vs 0.66 ± 0.16). These results suggest that Fed-FGS provides a practical and privacy-preserving framework for deploying robust medical image segmentation models across heterogeneous clinical centers.

Keywords: Federated Learning · Polyp Segmentation · Discrete Fourier Transform · Composite Loss · Frequency-Domain learning

1 Introduction

colorectal cancer ranks among the leading causes of cancer-related mortality worldwide; nevertheless, it is preventable through the early detection and removal of precancerous polyps during colonoscopy [1]. Recent advances in deep learning have enhanced automated polyp segmentation, providing the potential to assist clinicians in accurately localizing and diagnosing lesions [2]. However, models trained on data from a single institution lack generalizability across

clinical centers due to significant variability in imaging devices, acquisition techniques, and lesion characteristics [3]. At the same time, strict privacy regulations limit the sharing of patients across institutions, making centralised training impractical and motivating the adoption of federated learning (FL) for collaborative medical image analysis [4, 5].

FL addresses privacy constraints by allowing multiple institutions to jointly train a shared model without exchanging raw patient data [5, 7]. However, medical imaging datasets are typically non-IID, with substantial differences in lesion distributions, imaging conditions, and demographics across centers, which produce inconsistent local updates that degrade global model quality [6, 13, 14]. These variations produce heterogeneous local updates that can degrade the global models when conventional aggregation methods such as Federated Averaging (FedAvg) [7] are used.

Recent methods have aimed to improve federated learning under a heterogeneous data distribution. FedGS [8] assigns greater importance to updates associated with clinically challenging samples. Other approaches have also explored stronger feature representations, including federated foundation models for gastrointestinal endoscopy [6, 14] and federated parameter-efficient adaptation for polyp segmentation [15]. However, these methods do not directly address appearance-level domain shifts caused by differences in imaging conditions.

Frequency-domain approaches provide a complementary strategy for addressing appearance-level domain shifts. Low-frequency Fourier amplitudes capture imaging characteristics such as illumination and color distribution, whereas phase information retains structural content [9, 16]. Pan et al. [9] proposed frequency-based federated domain generalization for polyp segmentation by exchanging low-frequency amplitude information across clients to reduce cross-domain appearance discrepancies. Although frequency-domain approaches can improve robustness to appearance variation, such approaches do not explicitly address the optimization imbalance introduced by heterogeneous lesion distributions and varying sample difficulty across participating clients.

Motivated by these challenges, we propose **Fed-FGS**, a unified federated learning framework that extends FedGS, incorporating difficulty-aware gradient scaling with Discrete Fourier Transform Hard-Thresholding (DFT-HT) augmentation. DFT-HT is applied during local training to reduce appearance-level domain variation, while difficulty-aware gradient scaling increases the contribution of challenging samples to global optimization. A composite Dice-BCE-Focal loss is additionally employed to mitigate foreground-background imbalance. By jointly addressing appearance and optimization heterogeneity, Fed-FGS aims to improve cross-center generalization while preserving the privacy advantage of FL. We evaluated Fed-FGS on the multi-center PolypGen benchmark [10]. The main contributions of the proposed model are threefold: (i) a unified framework that combines frequency-domain augmentation and difficulty-aware federated optimization; (ii) integration of DFT-HT with FedGS to jointly address appearance-level domain shifts and optimization imbalance; and (iii) extensive evaluation of cross-center and unseen-center generalization.

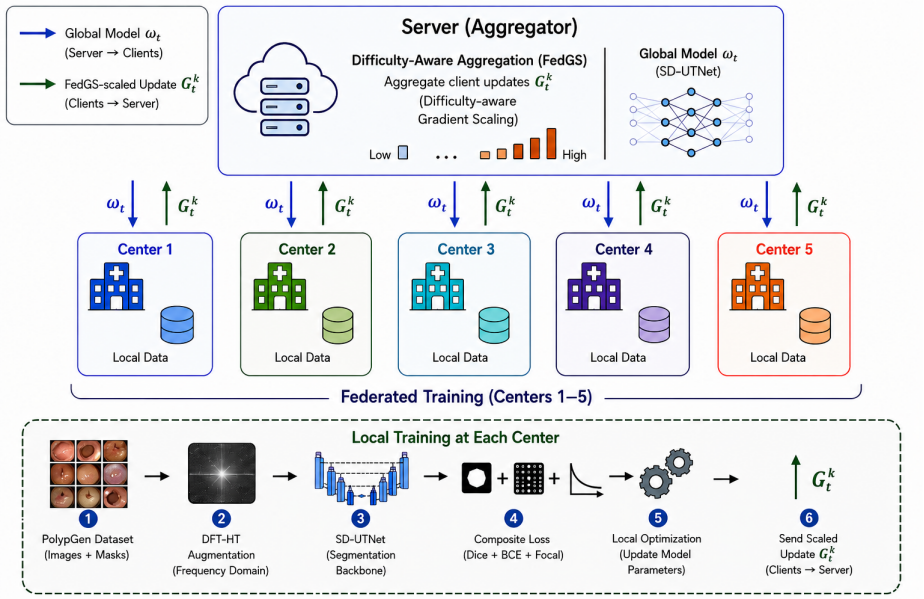

Fig. 1. Overview of the proposed Fed-FGS framework.

2 Methodology

Figure 1 presents the overall architecture of **Fed-FGS**, which consists of a central server coordinating multiple clients, each holding a private local dataset $\mathcal{D}_k$. At every communication round, the server distributes the current global model to all participating clients (k). Each client performs local training using **SD-UTNet** as the segmentation backbone, where input images are first processed through the **DFT-HT augmentation** [9] to reduce appearance variations across clinical centers. The segmentation network is optimized using a **composite Dice-BCE-Focal loss** [17, 15] to address severe class imbalance. After local optimization, model updates are scaled using the FedGS strategy before being transmitted to the server for global aggregation. By combining frequency-domain augmentation with difficulty-aware federated optimization, Fed-FGS aims to improve robustness and generalization across heterogeneous clinical environments.

2.1 Federated Learning Framework

We consider a federated learning setting consisting of K participating clinical centers. Each client k owns a private dataset D_k containing n_k samples, while the total number of training samples is $n = \sum_{k=1}^{K} n_k$. The objective is to optimize a shared global model without exchanging patient data as follows:

$$\min_{\omega} \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\omega). \tag{1}$$

Centers 1-5 of the PolypGen dataset [10] participate in federated training using five-fold cross-validation, while Center 6 is reserved for unseen-center evaluation. Each client adopts SD-UTNet as the local segmentation backbone due to its effectiveness in medical image segmentation under heterogeneous data distributions. During local training, either the original image x or the DFT-HT augmented image $\tilde{x}$ is provided as input, and the predicted segmentation mask is optimized using the composite loss described in the following subsection.

2.2 Frequency-aware local training

To improve robustness against appearance variations across clinical centers, FedFGS incorporates a frequency-domain augmentation strategy based on DFT Hard Thresholding (DFT-HT). The low-frequency amplitude components primarily encode domain-specific characteristics such as illumination, texture, and color distribution, whereas the phase spectrum preserves structural information. Therefore, selectively modifying amplitude information encourages the network to learn domain-invariant representations while maintaining anatomical consistency. For input x, the 2D DFT per channel c is:

$$X(c,u,v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} x(c,h,w)\, e^{-j2\pi\left(\frac{hu}{H}+\frac{wv}{W}\right)}. \tag{2}$$

Hard-thresholding suppresses noisy components:

$$\mathrm{HT}(x,T) = \begin{cases} x, & |x| \geq T \\ 0, & |x| < T \end{cases} \tag{3}$$

The thresholded amplitude is then mixed as follows:

$$\hat{A} = (1-\lambda)\, A \odot (1-M) + \lambda\, \mathrm{HT}(\tilde{A}, T) \odot M, \tag{4}$$

where $T = \alpha \cdot \max(\tilde{A})$, with α=0.05. The modified amplitude is recombined with the original phase and reconstructed via inverse DFT, reducing domain-specific variation while retaining structural fidelity [9, 17]. The reconstructed $\tilde{x}$ is subsequently used for local model training, allowing SD-UTNet to learn representations that are less sensitive to center-specific appearance variation while preserving clinically relevant structural information.

Composite Loss To further improve segmentation performance under severe foreground-background imbalance, the local network is optimized using a combination of Dice-BCE and focal losses:

$$\mathcal{L}_{\mathrm{Total}} = 0.7\, \mathcal{L}_{\mathrm{Dice}} + 0.2\, \mathcal{L}_{\mathrm{BCE}} + 0.1\, \mathcal{L}_{\mathrm{Focal}}, \tag{5}$$

where the Dice, BCE and Focal components are weighted by 0.7, 0.2, and 0.1, respectively.

2.3 Difficulty-aware Federated aggregation

Following local optimization, Fed-FGS employs the FedGS aggregation strategy to reduce optimization imbalance caused by heterogeneous lesion distributions. FedGS assigns larger aggregation weights to updates originating from clinically challenging samples, particularly those containing small polyps, ensuring that these underrepresented cases contribute more strongly during global optimizations process. The cumulative gradient for client k at round t is:

$$G_t^k = G_{t-1}^k + \eta_t \, (\omega_t^k - \omega_{t-1}^k), \tag{6}$$

aggregated by the server as:

$$G_A = \sum_{k=1}^{K} \frac{\text{steps}_k}{\text{steps}_{\text{total}}} \, G_t^k. \tag{7}$$

The difficulty-aware scaling factor η_t is computed from the inverse lesion area:

$$a_x^{-1} = \frac{H \times W}{\sum_{i,j} m_x(i,j)}, \qquad \eta_t = 1 + \frac{2}{N} \sum_{i=1}^{N} f\left(a_{x_i}^{-1}\right), \tag{8}$$

where $f(\cdot)$ is a logarithmic nonlinearity and m_x is the binary mask. Smaller polyps therefore contribute more strongly to global optimization [8, 17]. By integrating FedGS with the proposed DFT-HT augmentation, Fed-FGS simultaneously addresses optimization imbalance and domain shifts, leading to more robust federated learning under clinical conditions.

3 Dataset and Experimental Setup

We evaluate Fed-FGS on PolypGen [10], a multi-center dataset designed to assess the generalizability of polyp segmentation models. That dataset comprises colonoscopy data gathered from six independent clinical centers, exhibiting substantial variability in endoscopic devices, imaging protocols, illumination conditions, image quality, and polyp morphology. Such diversity makes PolypGen a challenging benchmark for evaluating federated learning methods under realistic cross-center domain shifts. Following the PolypGen evaluation protocol, Centers 1-5 are treated as independent federated clients participating in collaborative training using five-fold cross-validation, while Center 6 is excluded during training and used independently for out-of-distribution evaluation. Performance is assessed using overlap-based metrics (*Dice, IoU, Jaccard*), pixel-level metrics (*Recall, Precision, and F_2*), and detection-oriented metrics (Average Precision). To evaluate robustness across lesion scales, Dice was additionally reported separately for large polyps (*DiceL*) and small polyps (*DiceS*). All experiments were implemented in PyTorch. SD-UTNet was used as the local segmentation backbone for all clients. Local models were optimized using the Adam optimizer with an initial learning rate of 1×10^{-4}.

Table 1. Comparison of Fed-FGS with existing frameworks and baselines

Setting	Method	IoU	Recall	Precision	F_2	Jaccard	AP
Internal Validation	FedAvg	0.63	0.68	0.71	0.66	0.63	0.71
	FedGS [8]	0.63	0.67	0.72	0.66	0.63	0.72
	Fed-FGS	**0.66**	**0.69**	**0.73**	**0.68**	**0.66**	**0.81**
OOD generalization	FedAvg	0.65	0.68	0.72	0.67	0.65	0.72
	FedGS [8]	0.65	0.68	0.73	0.67	0.65	0.72
	Fed-FGS	**0.69**	**0.70**	**0.74**	**0.69**	**0.68**	**0.80**

Table 2. Dice score exploration (mean ± std over folds)

Criterion	Method	Dice	DiceL	DiceS
Best	FedAvg	0.72 ± 0.01	0.77 ± 0.03	0.43 ± 0.01
	FedGS [8]	0.73 ± 0.01	0.77 ± 0.01	0.44 ± 0.03
	Fed-FGS†	$\mathbf{0.75 \pm 0.02}$	$\mathbf{0.80 \pm 0.02}$	$\mathbf{0.46 \pm 0.03}$
Small Best	FedAvg	0.72 ± 0.02	0.78 ± 0.03	0.43 ± 0.01
	FedGS [8]	0.66 ± 0.16	0.77 ± 0.01	0.39 ± 0.12
	Fed-FGS†	$\mathbf{0.73 \pm 0.01}$	$\mathbf{0.78 \pm 0.01}$	$\mathbf{0.44 \pm 0.03}$

Table 3. Per-center evaluation on PolypGen. Best per block **bolded**. Split (train/test) is shown under each center label.

Subset	Method	IoU	DSC	Recall	Prec.	F_2	Jacc.	AP
Center 1 (230/26)	FedAvg	0.6592	0.7392	0.7755	0.7648	0.7481	0.6592	0.7433
	FedGS	0.6410	0.7130	0.7412	0.7673	0.7227	0.6410	0.7296
	Fed-FGS	**0.7060**	**0.7774**	**0.8004**	**0.8007**	**0.7793**	**0.7060**	**0.9126**
Center 2 (270/31)	FedAvg	0.6394	0.7049	0.6498	0.7648	0.6522	0.6394	0.7471
	FedGS	**0.6426**	**0.7068**	0.6298	0.7741	0.6390	**0.6426**	0.7526
	Fed-FGS	0.6186	0.6782	**0.6498**	**0.7800**	**0.6406**	0.6186	**0.8642**
Center 3 (441/46)	FedAvg	0.7702	0.8428	0.8374	0.8859	0.8362	0.7702	0.8405
	FedGS	0.7754	0.8459	0.8383	**0.8974**	0.8380	0.7754	0.8468
	Fed-FGS	**0.7818**	**0.8481**	**0.8433**	0.8874	**0.8426**	**0.7818**	**0.9389**
Center 4 (204/23)	FedAvg	0.4917	0.5395	0.3952	0.3605	0.3576	0.4917	0.4936
	FedGS	0.4888	0.5375	0.3912	0.3758	0.3585	0.4888	0.5026
	Fed-FGS	**0.5250**	**0.5619**	**0.3998**	**0.3780**	**0.3596**	**0.5250**	**0.6347**
Center 5 (187/21)	FedAvg	0.4238	0.4983	0.5406	0.5355	0.5077	0.4238	0.4935
	FedGS	0.4480	**0.5261**	**0.5684**	**0.5813**	**0.5324**	0.4480	0.5268
	Fed-FGS	**0.4596**	0.5189	0.5669	0.5545	0.5285	**0.4596**	**0.6654**

4 Results and Discussion

For an in-depth performance analysis of the framework, we evaluate the models under two distinct paradigms: *Internal Validation*, as shown in Table 1, and

Out-of-Distribution (OOD) Generalization. For internal validation, the trained models are evaluated on a unified test set formed by combining the test partitions of Centers 1–5, providing an overall performance estimate across the participating training centers through five-fold cross-validation. For OOD generalization, Center 6 is treated as a completely held-out clinical site to assess the robustness of the models against unseen domain shifts. On the internal validation set, Fed-FGS achieves the highest performance across all evaluation metrics compared with both FedAvg and FedGS. Compared with FedGS, Fed-FGS improves the IoU from 0.63 to 0.66, Recall from 0.67 to 0.69, Precision from 0.72 to 0.73, and F2-score from 0.66 to 0.68. The Jaccard index increases from 0.63 to 0.66, while Average Precision (AP) improves from 0.72 to 0.81. More importantly, these improvements are maintained under the unseen-center evaluation, where Fed-FGS continues to outperform both baselines, achieving an IoU of 0.69, Recall of 0.70, Precision of 0.74, F2-score of 0.69, Jaccard of 0.68, and AP of 0.80. These results demonstrate the consistent performance of Fed-FGS across both participating and unseen clinical centers.

As shown in Table 2, we further analyze the unseen-center behavior of all methods on Center 6 using *Best* checkpoint corresponding to the model achieving the highest overall Dice on the validation fold. The proposed framework improves generalization and stability with a Dice score of 0.75±0.02 on Center 6, surpassing FedGS (0.73±0.01) and FedAvg (0.72±0.01). The improvement is consistent across both target-scale regimes, with DiceL rising from 0.77 to 0.80 and DiceS from 0.44 to 0.46. The most notable advantage of the proposed method is observed under the Small Best selection criterion, which targets the clinically challenging small-polyp subset. Under this criterion, the *Small Best* checkpoint corresponds to the model achieving the highest Dice specifically on small polyps. Under this strict criterion, the FedGS baseline suffers a severe performance drop and high volatility, with its overall Dice falling to 0.66 ± 0.16 and its small-polyp Dice dipping to 0.39 ± 0.12. Conversely, the proposed framework demonstrates greater stability, maintaining an overall Dice of 0.73 ± 0.01 and a small-polyp Dice of 0.44 ± 0.03. These improvements are mainly attributed to the proposed DFT Hard-Thresholding augmentation, which reduces appearance-level domain shifts across clinical centers. In addition, the composite Dice-BCE-Focal loss provides complementary supervision by balancing region overlap, pixel-wise accuracy, and difficult foreground pixels, thereby improving optimization under severe class imbalance, particularly for small polyps.

The proposed Fed-FGS demonstrates competitive performance across all participating centers despite the substantial variability in imaging conditions and lesion characteristics as presented in Table 3. On Center 1, Fed-FGS achieves the highest performance across most evaluation metrics, improving the Dice score from 0.7392 (FedAvg) and 0.7130 (FedGS) to 0.7774, while also substantially increasing AP to 0.9126, indicating improved segmentation quality and prediction confidence. On Center 2, although the overall Dice score is slightly lower than FedGS, Fed-FGS achieves a notable improvement in AP (0.8642), suggesting more reliable probability estimation despite comparable segmentation accuracy.

Table 4. Ablation study of the proposed components on top of the FedGS baseline.

Method	DFT-HT	Composite Loss	Dice	IoU	AP
FedGS (baseline)	–	–	0.7170	0.6438	0.7113
FedGS-DFTHT	✓	–	0.7295	0.6646	0.7019
FedGS-CL	–	✓	0.7423	0.6747	**0.8592**
Fed-FGS	✓	✓	**0.7625**	**0.6905**	0.8481

On Center 3, which represents the strongest-performing center, Fed-FGS further improves the Dice score to 0.8481 and achieves the highest AP (0.9389), demonstrating that the proposed frequency-aware learning strategy remains beneficial even under favorable imaging conditions. On the more challenging Centers 4 and 5, where all methods experience reduced performance, Fed-FGS consistently outperforms the baselines on the primary overlap metrics for Center 4 and achieves the highest AP on both centers, highlighting its improved robustness under difficult clinical scenarios.

The ablation study in Table 4, evaluates the individual and combined contributions of DFT-HT and the composite loss under the same experimental setting. As shown, the composite Dice-BCE-Focal loss provides a stronger individual contribution than DFT-HT, improving Dice from 0.7170 to 0.7423, IoU from 0.6438 to 0.6747, and AP from 0.7113 to 0.8592. This improvement highlights the benefit of combining region-level Dice supervision with BCE and focal weighting for the foreground/background imbalance in polyp segmentation. DFT-HT alone also improves Dice and IoU to 0.7295 and 0.6646, respectively. Combining both components in Fed-FGS achieves the best Dice (0.7625) and IoU (0.6905) with a competitive AP of 0.8481, demonstrating their complementary contributions to segmentation performance.

5 Conclusion

In this work, we presented Fed-FGS, a unified FL framework for robust polyp segmentation across heterogeneous clinical centers. Unlike conventional FL approaches that primarily address optimization or representation learning independently, Fed-FGS jointly tackles optimization imbalance and appearance-level domain shifts through the integration of difficulty-aware gradient scaling and DFT-HT frequncy domain augmentation. Experimental results on the multi-center PolypGen benchmark demonstrate consistent improvements over FedAvg and FedGS in both internal and unseen-center evaluations, with strong robustness for clinically challenging small polyps. These findings indicate that optimization-aware aggregation and frequency-domain adaptation are complementary strategies for improving cross-center generalizability in privacy-preserving medical image segmentation. Future work will investigate adaptive frequency thresholding and evaluate additional federated optimization strategies.

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., et al.: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **68**, 394–424 (2018). doi:10.3322/caac.21492
2. Gupta, M., Mishra, A.: A systematic review of deep learning based image segmentation to detect polyp. Artif. Intell. Rev. **57**, 1 (2024). doi:10.1007/s10462-023-10621-1
3. Ali, S., Ghatwary, N., Jha, D., Isik-Polat, E., Polat, G., Yang, C., Li, W., Galdran, A., Ballester, M.-Á.G., Thambawita, V., et al.: Assessing generalizability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. Sci. Rep. **14**, 2032 (2024). doi:10.1038/s41598-024-52063-x
4. Hasan, M.M., Rahman, M.M.: Privacy-preserving polyp segmentation using federated learning with differential privacy. Smart Health 36, 100551 (2025).
5. Nguyen, D.C., Pham, Q.-V., Pathirana, P.N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., Hwang, W.-J.: Federated learning for smart healthcare: A survey. ACM Comput. Surv. **55**, 1–37 (2022). doi:10.1145/3501296
6. Kirchner, M., Jenke, A.C., Bodenstedt, S., Kolbinger, F.R., Saldanha, O.L., Kather, J.N., Wagner, M., Speidel, S.: Federated EndoViT: Pretraining Vision Transformers via Federated Learning on Endoscopic Image Collections. In: Medical Imaging with Deep Learning (2026). https://openreview.net/forum?id=zBtyakIFxx
7. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS), vol. 54, pp. 1273–1282 (2017)
8. Schutte, P., Corbetta, V., Beets-Tan, R., Silva, W.: FedGS: Federated gradient scaling for heterogeneous medical image segmentation. arXiv preprint arXiv:2408.11701 (2024)
9. Pan, H., Jha, D., Biswas, K., Bagci, U.: Frequency-based federated domain generalization for polyp segmentation. In: Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP) (2025). doi:10.1109/ICASSP49660.2025.10889662
10. Ali, S., et al.: A multi-center polyp detection and segmentation dataset for generalizability assessment. Sci. Data **10**, 75 (2023). doi:10.1038/s41597-023-01981-y
11. Rauniyar, A., Hagos, D.H., Jha, D., Håkegård, J.E., Bagci, U., Rawat, D.B., Vlassov, V.: Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. IEEE Internet Things J. **11**, 7374–7398 (2023).
12. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE Signal Process. Mag. **37**(3), 50–60 (2020)
13. Kairouz, P., et al.: Advances and open problems in federated learning. Found. Trends Mach. Learn. **14**(1–2), 1–210 (2021)
14. Devkota, A., et al.: Federated foundation model for GI endoscopy images. arXiv preprint arXiv:2505.24108 (2025)
15. Chen, X., Zhu, F., Li, D., Li, Q., Anwar, M.S., Shan, G., Jiang, J.: Towards Clinically Applicable Large-Model-Based Privacy-Preserving Polyp Segmentation: A Federated LoRA Approach to Colonoscopy. IEEE Journal of Biomedical and Health Informatics (2025). doi:10.1109/JBHI.2025.3639279.
16. Rieke, N., et al.: The future of digital health with federated learning. NPJ Digit. Med. **3**, 119 (2020)
17. Ou, B., Jia, Z.: FedESS: A federated learning approach with enhanced small-target segmentation for medical images. In: Proc. 2nd Int. Symp. AI Cybersecurity (ISAICS) (2025). doi:10.1109/ISAICS66888.2025.11350236