

Federated Learning for Histopathological Image Segmentation: A Privacy-Preserving Study Using the EBHI-Seg Dataset

Mohammadreza Azimi
0000-0002-9036-2122

Riga Stradins University
Institute of Microbiology and Virology
Ratsupites iela 5 Riga, LV-1067, Latvija
Email: Mohammadreza.Azimi@rsu.lv

Henrik Gabrielyan
0009-0009-8048-8379

Riga Stradins University
Institute of Microbiology and Virology
Ratsupites iela 5 Riga, LV-1067, Latvija
Email: Henrik.Gabrielyan@rsu.lv

Abstract—Federated learning (FL) enables collaborative model training without centralizing sensitive medical data, making it attractive for digital pathology. This study evaluates FL for histopathological image segmentation using the EBHI-Seg dataset under independent and identically distributed (IID) and class-stratified non-IID client partitions.

We simulate six decentralized clients and compare three segmentation architectures (UNet, DeepLabV3, FCN) with three FL strategies (FedAvg, FedProx, FedOptimizer), yielding 18 experimental configurations. Performance is assessed using Dice coefficient, Intersection over Union (IoU), loss, and between-round convergence volatility. The results show that the tested aggregation algorithms achieve similar final segmentation accuracy, while class-stratified non-IID partitioning lowers Dice scores and substantially increases convergence oscillation. These findings indicate that client-level heterogeneity, rather than the choice among the tested aggregators, is the dominant challenge for privacy-preserving federated histopathology segmentation.

Keywords: Federated Learning, Privacy-Preserving Data Spaces, Digital Pathology, AI in Healthcare, Bio Image Processing.

I. INTRODUCTION

THE integration of artificial intelligence (AI) into digital pathology has significantly advanced the field of computational histopathology, enabling automated analysis of whole-slide images for tasks such as tumor detection, grading, and tissue segmentation. In particular, deep learning models, such as convolutional neural networks (CNNs), have demonstrated remarkable performance in medical image segmentation, with architectures like U-Net becoming a standard due to their ability to capture both local and contextual features [1]. Despite these advancements, the success of such models is heavily dependent on access to large, diverse, and well-annotated datasets.

However, in medical domains, data sharing across institutions remains a major challenge. Strict privacy regulations, such as General Data Protection Regulation (GDPR), along with ethical and legal constraints, limit the centralization of patient data [2]. Additionally, institutional policies and concerns over data ownership further hinder collaborative model

development. As a result, many AI models in digital pathology are trained on limited and potentially biased datasets, which can negatively impact their generalizability and robustness across different clinical settings [3].

To address these limitations, Federated Learning (FL) has emerged as a promising distributed learning framework. Initially proposed by Brendan McMahan et al. [4], FL enables multiple institutions (clients) to collaboratively train a shared global model without exchanging raw data. Instead, model updates are computed locally and aggregated centrally, preserving data privacy while leveraging knowledge from decentralized sources. This paradigm is particularly well-suited for healthcare applications, where data sensitivity and heterogeneity are critical concerns.

Recent studies have explored the application of FL in medical imaging, including radiology and pathology, demonstrating its potential to achieve performance comparable to centralized training under certain conditions [5], [6]. However, several challenges remain, especially in the context of histopathological image segmentation. These include non-identically distributed (non-IID) data across institutions, variations in staining protocols, scanner differences, and class imbalance, all of which can affect model convergence and performance [7]. Moreover, the impact of such heterogeneity on segmentation tasks—compared to classification—remains underexplored.

In this work, we investigate the feasibility of federated learning for histopathological image segmentation using the EBHI-Seg dataset. By simulating a multi-institutional environment with varying data distributions, we aim to systematically evaluate the performance of a federated U-Net model relative to a centrally trained baseline. Our study further analyzes the influence of client variability and domain shifts on model performance, providing practical insights into the deployment of FL systems in real-world digital pathology workflows.

II. LITERATURE REVIEW

The rapid advancement of deep learning has significantly improved performance in medical image analysis, particularly

in tasks such as classification, detection, and segmentation. Convolutional neural networks (CNNs), especially architectures like U-Net, have become the standard for histopathological image segmentation due to their ability to capture both local and global contextual features. However, these approaches typically rely on centralized datasets, which are often difficult to construct in the medical domain due to privacy concerns, regulatory constraints, and institutional data silos.

To address these limitations, Federated Learning (FL) has emerged as a promising distributed learning framework that enables collaborative model training without sharing raw data. Instead, participating institutions train local models and share only model updates, thereby preserving data privacy while still benefiting from multi-institutional knowledge. Recent surveys have highlighted the growing importance of FL in medical imaging, particularly for overcoming data scarcity and enhancing model generalization across diverse clinical environments [8].

Several comprehensive reviews have analyzed the application of FL in medical image analysis. For instance, studies published in *Diagnostics* and *Electronics* emphasize that FL enables the use of distributed datasets while addressing privacy and security concerns inherent in healthcare systems [9]. These works also highlight key challenges, including communication overhead, system heterogeneity, and performance degradation under non-identically distributed (non-IID) data conditions. Similarly, research in radiology demonstrates that FL can facilitate multi-center collaborations without requiring data pooling, making it particularly suitable for sensitive imaging modalities such as MRI and CT scans [10].

Despite its advantages, federated learning introduces several technical challenges, especially in medical image segmentation. Variability in imaging protocols, annotation standards, and scanner characteristics across institutions leads to significant domain shifts that negatively impact model performance. Recent systematic reviews have shown that segmentation tasks are particularly sensitive to such heterogeneity, as differences in resolution, staining, and labeling quality can reduce model accuracy and convergence stability [11]. These issues are further exacerbated in histopathology, where staining variability and tissue morphology differ substantially between laboratories.

To mitigate these challenges, several advanced FL methodologies have been proposed. For example, domain generalization techniques such as those introduced in FedDG aim to improve model robustness across unseen institutions by explicitly modeling inter-client distribution shifts. These approaches leverage episodic training strategies to enhance generalization in decentralized environments. Additionally, cross-domain federated learning frameworks have demonstrated the feasibility of training models across heterogeneous datasets while maintaining acceptable segmentation performance, even in multi-organ and multi-modal scenarios [12].

Recent experimental studies further validate the effectiveness of FL in medical image segmentation tasks. For instance,

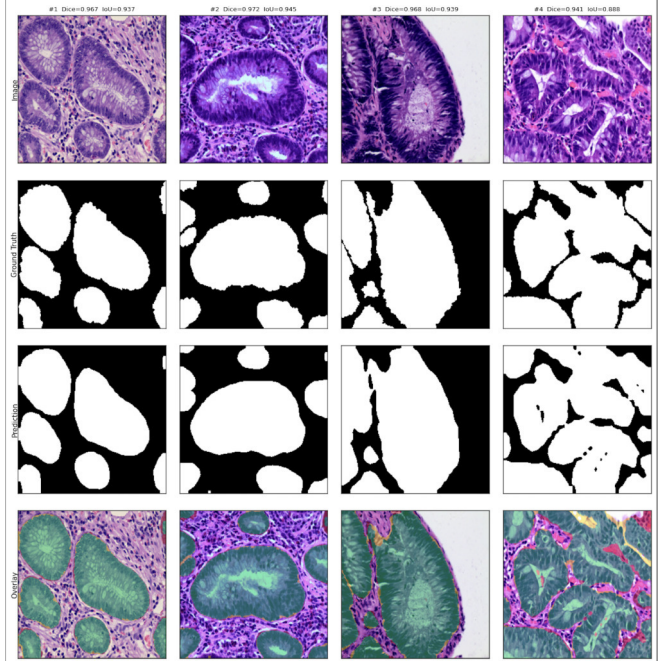


Fig. 1. Representative EBHI-Seg test predictions from federated DeepLabV3 trained with FedAvg: original images, ground-truth masks, model predictions, and overlay visualizations [15]

a study published in *Scientific Reports* demonstrated that federated learning combined with attention-based architectures can achieve competitive performance in brain tumor segmentation, highlighting the potential of FL for complex clinical applications [13]. Moreover, emerging research on large-scale medical imaging models suggests that FL can support continuous model updates and integration of new data sources without violating privacy constraints, making it suitable for real-world deployment in healthcare systems [14].

Overall, the literature indicates that federated learning represents a transformative approach for medical image analysis, particularly in scenarios where data sharing is restricted. While significant progress has been made, challenges related to data heterogeneity, communication efficiency, and model generalization remain open research problems. Addressing these limitations is essential for enabling robust and scalable FL-based solutions in digital pathology and other medical imaging domains.

III. METHODS AND MATERIALS

A. Methods

1) *Data Partitioning Strategy*: The EBHI-Seg image-mask pairs are first shuffled with a fixed seed and split into 70% training, 15% validation, and 15% testing. Only the training split is distributed across six virtual clients. In the non-IID setting, each client receives one tissue class (Table I); in the IID setting, the same training samples are randomly shuffled and divided into approximately equal client chunks. Validation and testing remain shared held-out sets spanning all classes.

TABLE I
CLIENT-LEVEL CLASS ASSIGNMENT FOR THE NON-IID PARTITION

Client	Assigned Class	Local Training Samples
C1	Adenocarcinoma	from global training split
C2	Low-grade IN	from global training split
C3	Polyp	from global training split
C4	Serrated adenoma	from global training split
C5	Normal	from global training split
C6	High-grade IN	from global training split
Validation and test sets are shared global held-out sets; exact counts are printed by the final run.		

2) *Segmentation Models*: We evaluate UNet, DeepLabV3, and FCN. UNet is a custom lightweight encoder-decoder with 32 base channels and approximately 1.9M parameters. DeepLabV3 and FCN use ResNet-50-based segmentation backbones with approximately 39M and 35M parameters, respectively. All models are initialized from scratch; for the ResNet-50-based models, batch normalization is replaced by group normalization to improve stability with small local batches.

3) *Federated Learning Algorithms*: We compare FedAvg, FedProx, and FedOptimizer. FedAvg aggregates client weights by sample-size-weighted averaging. FedProx adds a proximal penalty during local training to reduce client drift under non-IID data. FedOptimizer applies server-side adaptive first- and second-moment estimates to the aggregated update direction.

4) *Training Configuration*: All federated experiments use 20 communication rounds, 3 local epochs per round, 6 participating clients, batch size 16, learning rate 0.001, weight decay 0.0001, and a weighted BCE-Dice loss with equal weights. Local training uses Adam; FedOptimizer additionally applies adaptive moment estimates at the server. Experiments were run on an NVIDIA T4 GPU with 15 GB VRAM in Google Colab.

5) *Implementation Details and Reproducibility*: The implementation is a custom PyTorch-based single-machine FL simulation. The seed is fixed to 42 for Python's `random`, `NumPy`, `PyTorch`, and `CUDA` when available. Images are resized to 256×256 and normalized with ImageNet statistics; binary masks are thresholded at 127. Training augmentation includes horizontal and vertical flips, random 90° rotation, color jitter, and Gaussian blur, while evaluation uses only resizing and normalization. The public implementation is available at https://github.com/HenriARM/federated_learning.

6) *Evaluation Metrics*: Segmentation is evaluated at the final communication round using Dice similarity coefficient, $DSC = \frac{2TP}{2TP+FP+FN}$, Intersection over Union, $IoU = \frac{TP}{TP+FP+FN}$, and BCE-Dice loss.

B. Materials

This study uses EBHI-Seg [15], a hematoxylin-and-eosin histopathology dataset of gastrointestinal endoscopic biopsy image patches with expert pixel-level segmentation masks. The dataset covers six tissue classes and includes variations in morphology, staining intensity, and tissue structure, mak-

TABLE II
ALGORITHM PERFORMANCE COMPARISON AVERAGED ACROSS ARCHITECTURES

Algorithm	Dice (IID)	Dice (Non-IID)	Mean Dice
FedAvg	0.9439	0.9410	0.9425
FedProx	0.9436	0.9412	0.9424
FedOptimizer	0.9435	0.9408	0.9422

ing it suitable for evaluating segmentation robustness under simulated federated client heterogeneity.

IV. RESULTS

We conducted 18 federated experiments across three architectures, three aggregation algorithms, and two data-distribution settings, focusing on final accuracy and convergence behavior under heterogeneity. Across all algorithms and distributions, UNet achieved the highest mean Dice score (0.9468 ± 0.0015), followed by FCN (0.9404 ± 0.0015) and DeepLabV3 (0.9398 ± 0.0012).

A. Federated Learning Algorithm Comparison

Table II reports Dice scores averaged across the three segmentation architectures. FedAvg achieved the highest overall mean Dice coefficient (0.9425 ± 0.0033), followed very closely by FedProx (0.9424 ± 0.0036) and FedOptimizer (0.9422 ± 0.0034). Within each distribution setting, the differences between algorithms were extremely small: FedAvg was marginally highest under IID partitioning, while FedProx was marginally highest under non-IID partitioning. These results indicate that aggregation algorithm choice had negligible effect on final segmentation accuracy compared with the effect of data distribution.

B. Data Distribution Effects on Convergence

The IID partition achieved a mean Dice coefficient of 0.9437 ± 0.0033 , while the class-stratified non-IID partition yielded 0.9410 ± 0.0031 , a 0.27% degradation. More importantly, non-IID experiments showed approximately two-fold higher convergence oscillation (0.0122 versus 0.0053), indicating that heterogeneous clients introduce substantial between-round volatility even when final accuracy remains high.

C. Qualitative Segmentation Results

Figure 1 shows representative predictions from the federated DeepLabV3 model trained with FedAvg. The examples demonstrate visually consistent segmentation across diverse tissue appearances.

V. DISCUSSION

The 0.27% Dice gap between IID and non-IID settings is numerically small, but the two-fold increase in convergence oscillation shows that heterogeneity affects training stability. In the class-stratified setting, each client optimizes on a biased tissue distribution, so aggregated updates can reflect conflicting local objectives. In contrast, IID clients sample from the full class distribution and produce more stable consensus updates.

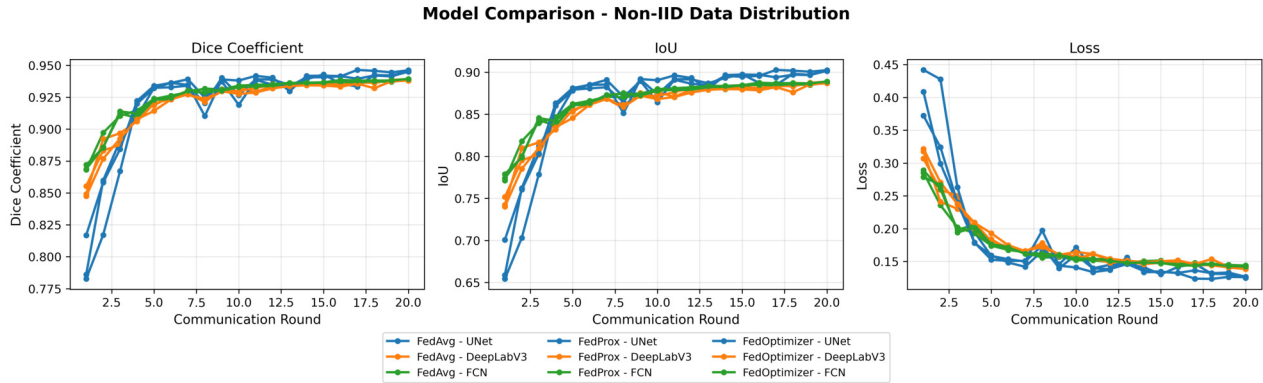


Fig. 2. Performance trajectories under class-stratified non-IID partitioning. The curves show Dice, IoU, and loss over 20 communication rounds for the tested model–algorithm combinations, with visible oscillations caused by client heterogeneity.

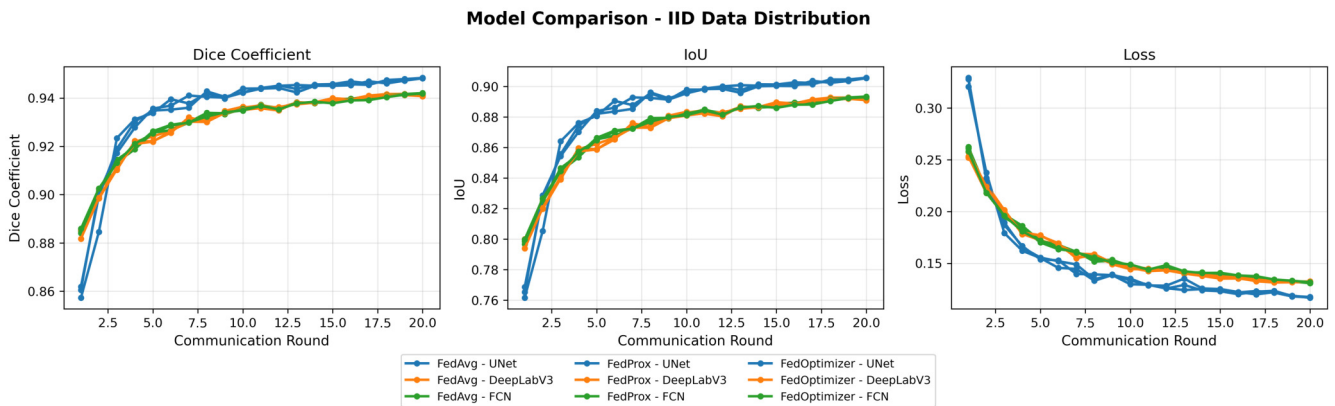


Fig. 3. Performance trajectories under IID random partitioning. Compared with Fig. 2, the IID setting shows smoother convergence and slightly higher final performance.

FedProx and FedOptimizer did not provide a practically meaningful improvement over FedAvg despite their mechanisms for reducing client drift or smoothing server updates. FedAvg was marginally best overall, while FedProx was marginally best in the non-IID subset, but the differences were below the level likely to affect deployment decisions. One explanation is that complete class stratification may be more severe than the regime where a proximal term is helpful. Another is that tissue segmentation in this setup is sufficiently constrained for weighted averaging to remain competitive. The smoothness result also cautions against using convergence curves alone as a robustness criterion: FedOptimizer produced smoother trajectories, but did not achieve the best final Dice.

For practical federated medical imaging, client heterogeneity should be treated as an expected deployment condition rather than removed by weakening privacy guarantees. IID partitioning is useful as an experimental upper bound, but real institutions generally cannot randomize patient populations across sites. Deployments should therefore document client distributions and consider balanced client sampling, class-aware validation, stain normalization, adaptive aggregation, personalization, and site-specific performance monitoring.

VI. CONCLUSION

This study evaluated FedAvg, FedProx, and FedOptimizer for EBHI-Seg histopathology segmentation under IID and class-stratified non-IID partitions. Across 18 configurations, UNet achieved the best mean segmentation performance, while the three aggregation algorithms produced nearly identical final Dice scores: FedAvg 0.9425, FedProx 0.9424, and FedOptimizer 0.9422 on average. Data distribution had a clearer effect, with non-IID partitioning reducing Dice by 0.27% and approximately doubling convergence oscillation. These results suggest that simple federated averaging remains a strong baseline for this setting, and that practical FL deployment should emphasize client-distribution documentation and heterogeneity mitigation without reducing privacy guarantees. Future work should study heterogeneity beyond class distribution, adaptive methods, and validation in genuinely distributed clinical environments.

REFERENCES

- [1] Y. Tian, Y. Zheng, Y. Liu, and Y. Wang, “GDC-Net: a U-Net for precise brain vessel segmentation with global–local and depthwise attention plus content-aware upsampling,” *The Journal of Supercomputing*, vol. 82, article 215, 2026, doi: <http://dx.doi.org/10.1007/s11227-026-08376-x>.

- [2] S. M. Turjya and M. Fawakherji, "Federated lung nodule segmentation using a hybrid transformer-U-Net architecture," *Scientific Reports*, vol. 16, article 5228, 2026, doi: <http://dx.doi.org/10.1038/s41598-026-35243-9>.
- [3] C. Chauhan et al., "The critical role of standards for AI in digital pathology: Digital Pathology Association Concept Paper," *Journal of Pathology Informatics*, vol. 21, article 100645, 2026, doi: <http://dx.doi.org/10.1016/j.jpi.2026.100645>.
- [4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR, vol. 54, pp. 1273–1282, 2017. Available: <https://proceedings.mlr.press/v54/mcmahan17a.html>.
- [5] S. Bakas, X. Li, P. Shah, and H. R. Roth, "Federated learning in healthcare: From research to real-world deployment," *Annual Review of Biomedical Engineering*, vol. 28, 2026, doi: <http://dx.doi.org/10.1146/annurev-bioeng-080125-041414>.
- [6] E. Markodimitrakakis et al., "Federated learning on magnetic resonance imaging: a critical review," *Artificial Intelligence Review*, vol. 59, article 97, 2026, doi: <http://dx.doi.org/10.1007/s10462-026-11508-7>.
- [7] A. R. Aurnob, S. A. Tanim, T. E. Shrestha, M. F. Mridha, and D. Mistry, "FedFusionNet: Advancing oral cancer recurrence prediction through federated fusion modeling," *Information Fusion*, vol. 132, article 104205, 2026, doi: <http://dx.doi.org/10.1016/j.inffus.2026.104205>.
- [8] H. Guan, P. T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, vol. 151, article 110424, 2024, doi: <http://dx.doi.org/10.1016/j.patcog.2024.110424>.
- [9] S. Nazir and M. Kaleem, "Federated learning for medical image analysis with deep neural networks," *Diagnostics*, vol. 13, no. 9, article 1532, 2023, doi: <http://dx.doi.org/10.3390/diagnostics13091532>.
- [10] M. H. U. Rehman, W. H. L. Pinaya, P. Nachev, J. T. Teo, S. Ourselin, and M. J. Cardoso, "Federated learning for medical imaging radiology," *The British Journal of Radiology*, vol. 96, no. 1150, article 20220890, 2023, doi: <http://dx.doi.org/10.1259/bjr.20220890>.
- [11] D. Ghosh, M. Mehjabin, M. E. Rayed, M. F. Mridha, and M. M. Kabir, "Advancements and challenges of federated learning in medical imaging: a systematic literature review," *Artificial Intelligence Review*, vol. 59, article 87, 2026, doi: <http://dx.doi.org/10.1007/s10462-025-11489-z>.
- [12] V. S. Parekh, S. Lai, V. Braverman, J. Leal, S. Rowe, J. J. Pillai, and M. A. Jacobs, "Cross-domain federated learning in medical imaging," *arXiv preprint arXiv:2112.10001*, 2021, doi: <http://dx.doi.org/10.48550/arXiv.2112.10001>.
- [13] S. Alphonse, F. Mathew, K. Dhanush, and V. Dinesh, "Federated learning with integrated attention multiscale model for brain tumor segmentation," *Scientific Reports*, vol. 15, article 11889, 2025, doi: <http://dx.doi.org/10.1038/s41598-025-96416-6>.
- [14] M. Sun, Z. Yang, Y. Huang, H. Yu, Y. Chen, S. Qi, A. B. J. Teoh, and Y. Zhang, "Federated learning for large models in medical imaging: a comprehensive review," *arXiv preprint arXiv:2508.20414*, 2025, doi: <http://dx.doi.org/10.48550/arXiv.2508.20414>.
- [15] L. Shi et al., "EBHI-Seg: a novel enteroscopy biopsy histopathological hematoxylin and eosin image dataset for image segmentation tasks," *Frontiers in Medicine*, vol. 10, article 1114673, 2023, doi: <http://dx.doi.org/10.3389/fmed.2023.1114673>.