Decentralized Noise Handling in Medical Imaging: Encoder-Decoder Based Federated Imputation for Robust Training

Yunyoung Chang[1], Yeonwoo Noh[2], Sang-Woong Lee[1*], Minwoo Lee[3], and Wonjong Noh[4*]

[1] School of Computing, Gachon University, Seongnam, Republic of Korea
jangyyoung88@gachon.ac.kr, slee@gachon.ac.kr
[2] College of Medicine, Gachon University, Inchon, Republic of Korea
[3] Neurology, Hallym University Sacred Heart Hospital, Anyang, Republic of Korea
[4] School of Information Science, Hallym University, Chunchon, Republic of Korea
wonjong.noh@hallym.ac.kr

Abstract. Noise in medical imaging is an inevitable challenge, often stemming from acquisition artifacts, varying imaging protocols, and external interference. While some studies suggest that noise can enhance model robustness, excessive or unstructured noise degrades training quality and classification performance. This issue is further exacerbated in federated learning settings, where individual clients have limited local data, making it difficult to train robust models independently. Federated imputation has been explored as a solution, yet existing methods do not fully leverage federated learning settings for optimal noise reconstruction. In this work, we introduce a novel encoder-decoder based federated imputation method, designed to replace noisy images with more representative reconstructions before training. Experimental results demonstrate that classification models trained with images imputed by the proposed method consistently outperform those trained with raw noisy images and without noisy images, highlighting the importance of effective noise handling in federated learning-based medical imaging.

Keywords: Data Imputation · Federated Learning · Medical Image Classification · Variational Autoencoder · Swin Transformer

1 Introduction

Medical imaging plays a crucial role in modern diagnostics. However, medical images are often affected by various types of noise due to acquisition artifacts, differing imaging protocols, and external interference [1]. Although a modest level of noise may occasionally promote model robustness [2], excessive or erratic noise typically impairs both image quality and subsequent classification performance.

[*] Corresponding authors. Emails: slee@gachon.ac.kr, wonjong.noh@hallym.ac.kr

To address these challenges, various image inpainting techniques[3,4,5,6,7,8,9] have been developed, ranging from traditional methods to modern deep learning-based approaches. Bertalmio et al. [3] was one of the first studies to propose an image inpainting technique for restoring occluded regions in digital images. It introduced a partial differential equation (PDE)-based inpainting approach, making significant contributions to traditional inpainting methods. Hassanpour et al. [4] introduced edge-aware coarse-to fine GAN (E2F-GAN) to achieve more natural reconstructions. Kingma et al. [5] and Vaswani et al. [6] introduced Variational Auto Encoder (VAE) and Transformer model which led to significant improvement in generating plausible image structures [7] . Unlike traditional models, these approaches learn high-level semantic features, allowing more coherent image restoration. Recent generative models leverage encoder-decoder architectures to reconstruct missing or noisy regions more effectively. Jeevan et al [8] introduced WavePaint model leveraging multi-resolution token mixing using 2D-discrete wavelet transform (DWT) [9] to reconstruct occluded regions effectively. Despite these advancements, many inpainting methods still show tiling or patch-like artifacts in the filled regions and rely on externally provided masks as an additional input alongside the corrupted image. This dependency complicates the training pipeline and limits model adaptability in real-world scenarios where such masks are unavailable.

However, in federated learning (FL) settings, noise-related issues are exacerbated due to the limited local data available to each client [10]. This data scarcity hampers independent model training, making robust learning particularly challenging [11]. Recent studies [12,13] have investigated federated imputation methods to enhance image reconstruction before training. Wu et al. [12] have presented a novel approach to enhance noise robustness in FL by sampling confidence-based learning weight adjustment and noise-aware global optimization. Balelli et al. [13] introduced Fed-MIWAE leveraging a VAE-based generative model within FL to effectively compensate for missing values in each client. These approaches improve imputation performance by combining the strengths of generative modeling and privacy-preserving learning. However, existing studies do not fully leverage FL's decentralized setting.

In this work, we introduce an encoder-decoder-based federated imputation method designed to enhance noise reconstruction in FL environments. The proposed approach incorporates the following key advancements:

- **End-to-End Learning**: We developed an end-to-end framework in which the model automatically identifies noise and utilizes it for inpainting. Unlike conventional methods, the proposed model predicts occlusion masks directly from the input image, streamlining the imputation process.
- **Swin Transformer-based WaveMix Module**: We employed a Swin Transformer [15] for token mixing, effectively reducing repetitive artifacts and enhancing fine-grained image details. It combines an attention-based structure with a probability map for occluded regions, which enables more natural and accurate reconstructions, improving FL-based medical imaging tasks.

This novel framework not only eliminates the need for manual mask annotations but also significantly enhances the quality of imputed images, leading to more robust model training in federated medical imaging environments.

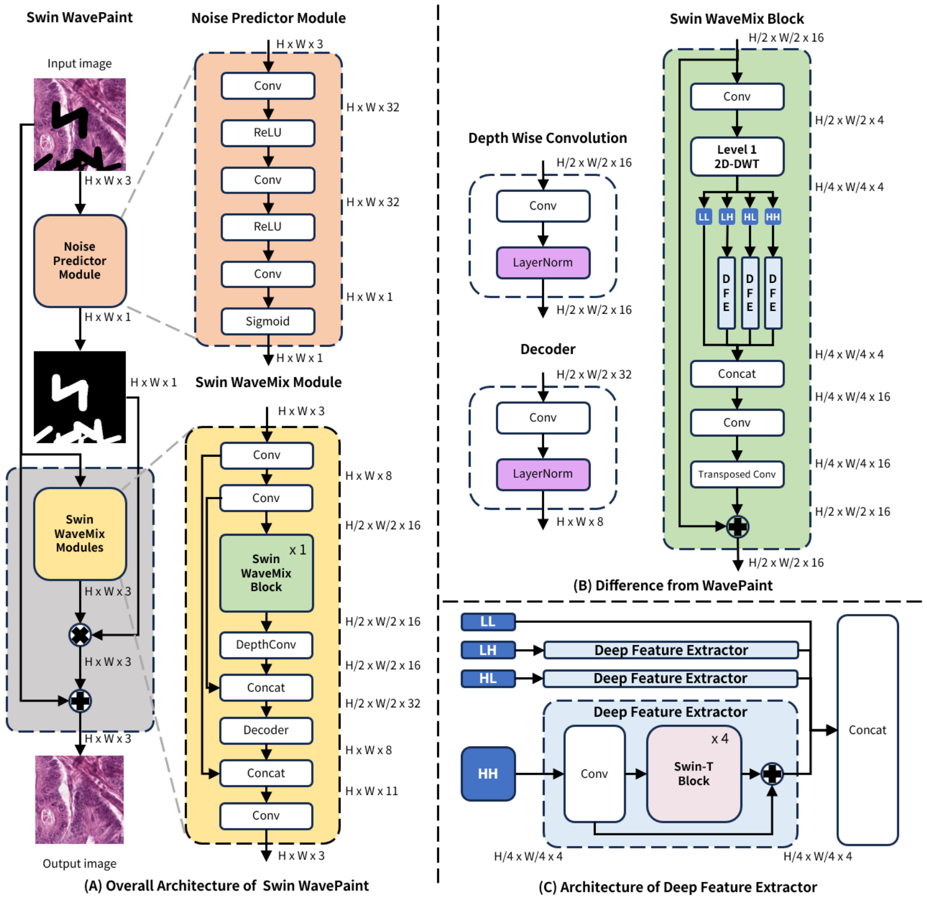

Fig. 1. (a) The overall architecture of a Swin WavePaint; (b) difference from WavePaint; (c) architecture of Deep Feature Extractor (DFE). Swin WavePaint consists of Noise Predictor Module and Swin WaveMix Modules. Swin WaveMix Module uses one Swin WaveMix Block consists of level 1 2D-DWT [14] and DFEs. DFE uses Swin Transformer Block [15] to extract deep features from the results of wavelet trasforms.

2 The Proposed Model

2.1 Imputation Model

The proposed imputation model is built on some major modules and functions: noise prediction module, layer normalization, wavelet decomposition, and swin-transformer based token mixing, which are shown in Fig. 1. We train the model in both centralized and federated settings to demonstrate the effectiveness of our modules.

Noise Prediction Module. We employ a noise predictor module that estimates the location of missing or corrupted regions without relying on external noise information. The module contains three convolutional layers and a sigmoid activation that produces a probability map indicating which pixels belong to the occluded region. The predicted mask is then used internally to guide the inpainting process, enabling an end-to-end pipeline that only requires the noisy image.

Wavelet Decomposition. We employ 2D-discrete wavelet transforms (DWT) to perform efficient spatial token mixing. A single-level wavelet decomposition splits the input features into a low-frequency (LL) component, capturing global structure, and multiple high-frequency (LH, HL, HH) bands that encode fine textures and edges. Because wavelet transforms do not introduce additional learnable parameters, they help keep the model lightweight. Furthermore, the inherent downsampling in wavelet decomposition expands the effective receptive field quickly, which is advantageous for capturing larger contextual information in inpainting tasks.

Deep Feature Extractor. We employ DFE blocks in conjunction with wavelet decomposition to extract deep feature through Swin Transformer-based Token Mixing. We integrate 4 Swin Transformer layer in the high-frequency branch, leveraging window-based multi-head self-attention and hierarchical feature extraction. In particular, each window is processed locally, and subsequent layers apply shifted windows to encourage cross-window interactions. This mechanism helps reduce tiling or patch-like artifacts that can occur when the model repeatedly applies the same learned patterns to fill large occlusions.

Layer Normalization. We employ Layer Normalization (LN) [16] instead of Batch Normalization (BN) to ensure batch-independent stability, particularly in FL environments, where institutions have varying computational resources and hardware constraints. In an FL setting, some institutions (clients) may have limited computing power, making it difficult to process large batch size efficiently. Since BN relies on stable batch statistics, small or inconsistent batch sizes can degrade performance and hinder generalization. In contrast, LN normalizes activations per sample, ensuring stability regardless of batch size. This makes it suitable for institutions with limited computational resources, enabling effective training with smaller batches.

2.2 Classification Model

The classification model contains a total of seven hidden layers, consisting of five convolutional blocks and two fully connected (FC) layers. The five convolutional layers serve as the primary feature extractors by applying learnable filters to capture local patterns in the input image. As the network deepens, these build hierarchical representations from simple to complex features, with pooling operations reducing spatial dimensions and emphasizing key information. The FC layers combine features, which are extracted from the covolutional layers, through non-linear transformations to produce class probabilities. The model is trained only in federated setting.

3 Experiments

3.1 Dataset

The proposed model is evaluated on the PathMNIST dataset from MedMNIST, which is a benchmark for medical image classification [17]. PathMNIST provides images in multiple resolutions – 28x28, 64x64, 128x128, and 224x224 – for 2D images, and we use only the 224x224 resolution images to better leverage standard deep learning architectures and improve feature extraction. The dataset provides a robust benchmark for developing resource-efficient models in medical image classification under diverse imaging conditions. The dataset is partitioned into 89,996 training images, 10,004 validation images, and 7,180 test images.

3.2 Noise Generation

There are three types of noise: the narrow mask occludes small areas, requiring detailed restoration; the medium mask covers moderately sized areas, needing more information to be filled in; and the wide mask covers large areas, helping the model handle more complex restoration tasks. The noise is created in binary form, with missing parts marked as 1 and the remaining parts as 0, and are used as inputs to the model to guide it in reconstructing only the occluded areas. In addition to these three types of noise generation methods from the previous WavePaint model, we incorporated the Noisy Mask, which randomly places noise points at the pixel level to create irregular and unpredictable occlusions. This type of noise further challenges the model by introducing non-uniform and random missing regions, requiring the model to handle more diverse and complex restoration tasks. We show types of noise in Fig. 2.

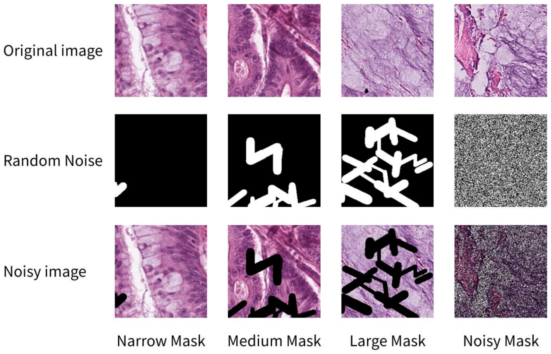

Fig. 2. Types of Noise

3.3 Federated Learning

We configured the number of clients to 10 under a Quantity Skew setting. The 89,996 training images are partitioned into 10 disjoint client datasets. In contrast, the 10,004 validation images are not partitioned across clients but are used as a centralized shared validation set for all clients during training. To further emulate real-world non-IID distributions, each clients is assigned a favored label, with approximately 70% of its data drawn from that class, while the remaining 30% is randomly sampled from other classes. This strategy not only mirrors the inherent quantity imbalance across institutions but also introduces label skew, ensuring that each client's dataset uniquely reflects the varying prevalence of conditions without any duplicate samples.

We assume that half of each client's data is corrupted with random noise while the other half remains clean. The standard PathMNIST validation set is used to monitor the model during training, and the final performance is evaluated on the test set. To integrate the global model, we employed FedAVG [18] for federated optimization, aggregating model updates from distributed clients. We set the number of global rounds for both federated inpainting and classification to 10, with each client performing 10 epochs of local training per round. The experiments were conducted on four GeForce RTX 2080 Ti GPUs.

3.4 Inpainting

We use hybrid loss function, which integrates LPIPS, L1, and MSE loss functions [8]. The LPIPS component captures perceptual similarities, while the L1 and MSE terms ensure pixel-level accuracy, yielding a balanced evaluation of both visual fidelity and quantitative performance. In both centralized and federated settings we employ AdamW optimizer with a learning rate of 0.001. The beta coefficients were set to (0.9, 0.999), epsilon to 1e-8, and weight decay to 0.01.

Table 1. Inpainting Results on Centralized setting

Model	Hybrid Loss	LPIPS	MSE	L1
WavePaint	0.508	0.0376	0.0128	0.0136
LN(Proposed)	0.406	0.0335	0.0020	0.0121
Swin(Proposed)	0.431	0.0356	0.0021	0.0129

Table 2. Inpainting Results on Federated setting

Model	Hybrid Loss	LPIPS	MSE	L1
LN(Proposed)	0.728	0.0624	0.0032	0.0176
Swin(Proposed)	0.705	0.0607	0.0029	0.0167

3.5 Classification

Classification is performance using raw noise data, deleted data, and imputed (by previous inpainting model and the proposed inpaining model) data in federated setting. We employ Cross Entropy Loss as a loss function for training classification model, and Stochastic Gradient Descent(SGD) with learning rate of 0.001 and momentum of 0.9. We evaluate the performance by Accuracy (ACC) and Area Under the ROC Curve (AUC).

Table 3. Classification Results for PathMNIST on Federated setting.

Noise Handling	ACC	AUC
Drop Noise	0.5287	0.9261
With Noise	0.7082	0.9553
LN(Proposed)	0.7173	0.9575
Swin(Proposed)	0.7097	0.9569

3.6 Results and Discussions

In Table 1, the performance of the proposed LN (using LN instead of BN in WavePaint) and Swin methods is compared against the baseline WavePaint model in terms of Hybrid Loss, LPIPS, MSE, and L1 metrics, where lower values signify better performance. Both proposed models outperform the WavePaint baseline, with the LN method achieving the lowest Hybrid Loss, LPIPS, MSE and L1. This result indicates that the proposed LN is particularly effective at reconstructing noisy images in centralized setting.

On the other hand, Table 2 compares the performance in FL setting. It is shown that the proposed Swin method consistently exhibits marginally lower values across all evaluation metrics relative to the LN method. This suggests that the Swin architecture is better suited to address the inherent challenges of FL—namely, handling heterogeneous data distributions and communication constraints—thereby yielding improved inpainting quality.

Table 3 assesses the impact of different noise-handling strategies on classification performance using the PathMNIST dataset in a federated environment. The strategy of dropping noisy data yields the lowest ACC, while training with all data (with noise) improves both ACC and AUC. Notably, the proposed imputation-based approaches (LN and Swin) further enhance classification performance, as evidenced by higher ACC and AUC values compared to the simpler noise-handling methods. These findings underscore the efficacy of imputation techniques in mitigating the adverse effects of noise on classification outcomes in FL settings.

4 Conclusion

In this work, we present a novel encoder-decoder based federated imputation method designed to address the challenges of noisy data in medical imaging, particularly within FL environments. The proposed methods have two key advancements, end-to-end federated imputation and deep feature extractor using Swin-Transformer-based token mixing. We performed evaluations using the PathMNIST dataset. In centralized environment, the proposed LN method demonstrated superior performance. On the other hand, in distributed environement, such as in non-IID FL environment, the proposed SWIN method achieved enhanced performance than the LN method, suggesting that its attention-based architecture is better suited to handle non-IID distributions and communication constraints inherent in FL. Furthermore, classification experiments reveal that imputation-based noise handling significantly improves model performance, demonstrating that effective imputation can mitigate the adverse effects of noise on classification outcomes in decentralized environments. Overall, the experiments highlight the potential of federated imputation as a robust strategy for enhancing medical image reconstruction and classification.

Acknowledgments. This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF), funded by the Korean government (MSIT) (No. RS-2023-00223501).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Krupa, Katarzyna, and Monika Bekiesińska-Figatowska. "Artifacts in magnetic resonance imaging." Polish journal of radiology 80 (2015): 93.
2. Havsteen, Inger, et al. "Are movement artifacts in magnetic resonance imaging a real problem?—a narrative review." Frontiers in neurology 8 (2017): 232.

3. Bertalmio, Marcelo, et al. "Image inpainting." Proceedings of the 27th annual conference on Computer graphics and interactive techniques. 2000.
4. Hassanpour, Ahmad, et al. "E2F-GAN: Eyes-to-face inpainting via edge-aware coarse-to-fine GANs." IEEE Access 10 (2022): 32406-32417.
5. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." 20 Dec. 2013,
6. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
7. Esser, Patrick, et al. "Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis." Advances in neural information processing systems 34 (2021): 3518-3532.
8. Jeevan, Pranav, Dharshan Sampath Kumar, and Amit Sethi. "Wavepaint: Resource-efficient token-mixer for self-supervised inpainting." arXiv preprint arXiv:2307.00407 (2023).
9. Strang, Gilbert, and Truong Nguyen. Wavelets and filter banks. SIAM, 1996.
10. Wu, Chenrui, et al. "Learning Critically in Federated Learning with Noisy and Heterogeneous Clients."
11. Lu, Yang, et al. "Federated learning with extremely noisy clients via negative distillation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 13. 2024.
12. Wu, Nannan, et al. "Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity." arXiv preprint arXiv:2305.05230 (2023).
13. Balelli, Irene, et al. "Fed-miwae: Federated imputation of incomplete data via deep generative models." arXiv preprint arXiv:2304.08054 (2023).
14. Sethi, Amit. "WaveMix: Multi-Resolution Token Mixing for Images." (2021).
15. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
16. Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer normalization." arXiv preprint arXiv:1607.06450 (2016).
17. Yang, Jiancheng, et al. "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification." Scientific Data 10.1 (2023): 41.
18. McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.