Hybrid Diffusion Model for Breast Ultrasound Image Augmentation

Farhan Fuad Abir *
Electrical and Computer Engineering
University of Central Florida
farhan.fuad@ucf.edu

Sanjeda Sara Jennifer*
Department of Computer Science
University of Central Florida
sanjedasara.jennifer@ucf.edu

Niloofar Yousefi
Industrial Engineering and Management Systems
University of Central Florida
niloofar.yousefi@ucf.edu

Laura J. Brattain
Department of Medicine
University of Central Florida
laura.brattain@ucf.edu

Abstract

We propose a hybrid diffusion-based augmentation framework to overcome the critical challenge of limited and imbalanced data in breast ultrasound (BUS) datasets. Unlike conventional augmentations, our approach captures ultrasound-specific features such as speckle noise by combining text-to-image generation with image-to-image (img2img) refinement as well as fine-tuning using LoRA and textual inversion (TI). Our method generated realistic, class-consistent images on an open-source Kaggle breast ultrasound image dataset (BUSI). Incorporating TI and img2img refinement on the Stable Diffusion v1.5 backbone reduced the Fréchet Inception Distance (FID) from 45.97 to 33.29, demonstrating a substantial gain in fidelity while maintaining comparable downstream classification performance. Overall, the proposed framework effectively mitigates the low-fidelity limitations of synthetic ultrasound images and enhances the quality of augmentation for robust diagnostic modeling.

1 Introduction

Ultrasound is a key medical imaging modality for breast cancer diagnosis due to its affordability, non-invasive nature, and real-time imaging capabilities [1]. However, deep learning–based ultrasound interpretation faces persistent challenges due to limited and imbalanced datasets that constrain model generalizability across lesion types. Traditional augmentation techniques, such as flipping, rotation, and intensity variations, offer only limited diversity [2]. Early generative approaches using Generative Adversarial Networks (GANs) often produce artifacts and fail to replicate the speckle noise and subtle tissue textures critical for diagnosis [3].

Diffusion models progressively denoise latent representations, yielding structural detail essential for ultrasound synthesis. These models have been widely used for denoising, despeckling [4] and text-conditioned generation of echocardiograms or breast ultrasound (BUS) scans [5]. These methods showed promising results of using diffusion-based image generation for BUS. However, existing diffusion-based BUS generation tends to produce overly smooth images that lack the characteristic speckle noise and structural complexity of real scans.

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

To address the research gap, we propose a hybrid diffusion augmentation approach to improve the fidelity of ultrasound synthesis and the severe class imbalance in breast ultrasound datasets. Our framework integrates semantic conditioning with text-to-image (text2img) generation and structural refinement with image-to-image (img2img) diffusion [6]. First, we implemented Stable Diffusion v1.5 (SD1.5) [6], fine-tuned with Low-Rank Adaptation (LoRA) [7] and enhanced with textual inversion (TI) [8] to embed domain-specific ultrasound concepts into the model's latent space. Then, we use img2img for retaining fine-grained BUS structures. This hybrid strategy captures speckle noise, tissue heterogeneity, and lesion boundaries more faithfully than other diffusion-based approaches. We used the proposed framework to generate high-quality synthetic samples for underrepresented classes, addressing data imbalance in the an open-source Kaggle breast ultrasound image (BUSI) dataset [9]. By producing class-consistent and anatomically coherent ultrasound images, the method improves downstream classification performance.

2 Methodology

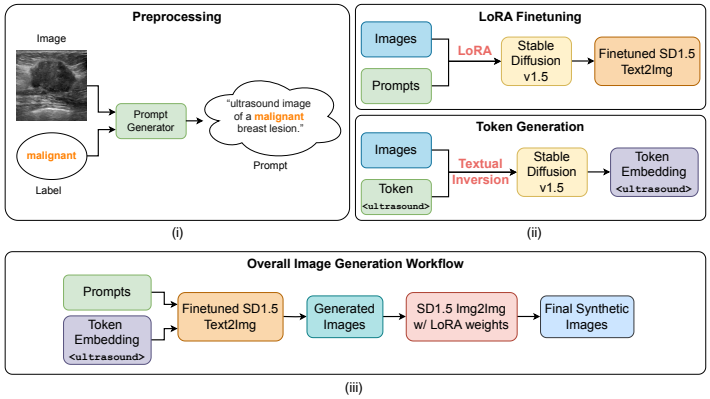

Figure 1: Overview of the proposed hybrid diffusion-based image generation framework for breast ultrasound augmentation.

2.1 Dataset Description

We used the Breast Ultrasound Images (BUSI) dataset consisting of 780 grayscale scans categorized as normal (133), benign (437), and malignant (210) [9]. All images were resized and normalized to ensure consistency. The train-test split (80-20) consisted of 623 training images (349 benign, 168 malignant, 106 normal) and 157 validation images (88 benign, 42 malignant, 27 normal).

2.2 Hybrid Diffusion Model

The overall workflow consists of three stages: preprocessing, model fine-tuning, and image synthesis, as shown in Figure 1.

Prompt Generation. We derived class-specific prompts from BUSI labels (normal, benign, malignant) to guide the generation process. These prompts were mapped to radiology-style descriptions such as "ultrasound image of a benign (or malignant) breast lesion." For the normal class, it is: "ultrasound image of a benign breast tissue."

Model Fine-Tuning. Firstly, we used Low-Rank Adaptation (LoRA) to fine-tune the Stable Diffusion model with the images and associated text prompts. This facilitated domain-specific feature learning while avoiding full-scale model pretraining. Then we introduced a custom token `<ultrasound>` through TI to improve the model's capacity to interpret ultrasound-specific prompts. This technique created a new embedding that encoded domain-specific patterns on a curated set of representative BUSI images. The learned embedding was appended to all prompts, enhancing semantic alignment and promoting consistent generation across BUS classes.

Synthetic Image Generation. At first, we generated the synthetic BUS images from the finetuned text2img SD1.5 model. Then we applied Stable Diffusion's img2img pipeline to enhance the fidelity

Table 1: Impact of each component in the hybrid diffusion framework.

Components	Accuracy ↑	F1-Score ↑	AUC-ROC ↑	PPV ↑	FID ↓
Baseline (Real)	0.904	0.887	0.979	0.890	-
Real + SD1.5	0.917	0.905	**0.986**	0.901	45.97
Real + SD1.5 + img2img	0.898	0.879	0.978	0.878	38.34
Real + SD1.5 + TI	**0.924**	**0.912**	0.980	**0.906**	45.66
Real + SD1.5 + TI + img2img	0.905	0.884	0.975	0.89	**37.18**

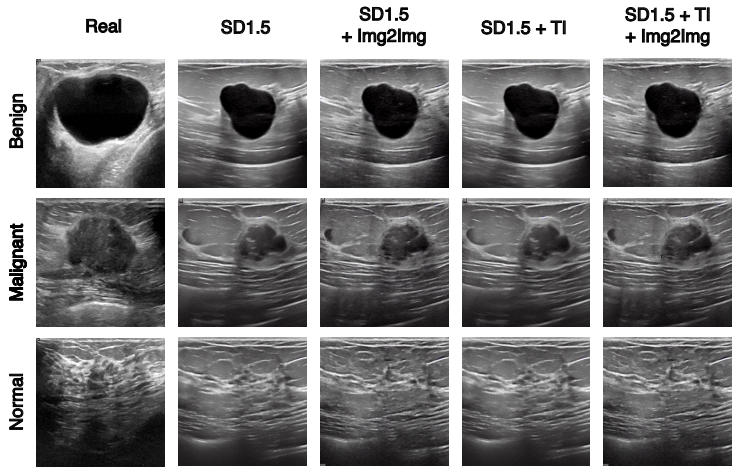

Figure 2: Comparison of real ultrasound images and synthetic variants.

of the generated outputs. This refinement process was performed with a low denoising strength of 0.3. Afterward, we used the hybrid diffusion pipeline to balance the training dataset. We added synthetic images to each class of the training set, yielding 350 samples per class, while the validation set remained unchanged.

2.3 Evaluation Metrics

We used accuracy, F1-score, AUC-ROC, PPV, and recall to evaluate classification performance. Moreover, we used the Fréchet Inception Distance (FID), which quantifies the similarity between real and synthetic image distributions. FID scores were computed using Inception v3 features on 780 real and synthetic images.

3 Experiments and Results

3.1 Experimental Setup

We conducted five experiments using original ultrasound images as the baseline and ResNet18 was used to classify the three classes. (1) Baseline (Real): trained only on original ultrasound images; (2) Original + SD1.5: augmented with synthetic samples generated by SD1.5; (3) Real + SD1.5 + img2img: incorporating img2img translation for style and contrast adaptation; (4) Real + SD1.5 + TI: leveraging textual inversion for domain-specific embedding alignment; and (5) Real + SD1.5 + TI + img2img: combining all generative modules.

ResNet18 was trained with the Adam optimizer (learning rate = 0.0001, batch size = 16, 30 epochs) using cross-entropy loss. Standard augmentations (horizontal flip, normalization) were applied to improve generalization. All experiments ran on a workstation with a 12th Gen Intel Core i5-12500 CPU, 128 GB RAM, and an NVIDIA RTX A4000 (16 GB) GPU, under Ubuntu Linux. Model training and diffusion-based synthesis were implemented in PyTorch [10] using the Hugging Face diffusers library [11].

3.2 Results

Table 1 summarizes the impact of LoRA fine-tuning, TI, and img2img refinement. Using SD1.5 text2img augmentation with real data improved the baseline to Acc 0.917, F1 0.905, and the highest AUC-ROC 0.986 (FID 45.97). Incorporating TI achieved the best overall classification (Acc 0.924, F1 0.912, PPV 0.906), demonstrating that the learned domain token enhances class-aware synthesis (FID 45.66). Img2img consistently lowered FID (to 38.34 without TI and 37.18 with TI) but slightly reduced accuracy and F1. AUC-ROC remained ≥ 0.975 across all settings, indicating stable discriminative performance. Moreover, figure 2 qualitatively illustrates that the generated ultrasound images preserve key diagnostic features showing that TI and img2img enhances visual fidelity.

4 Conclusion

We presented a hybrid diffusion–based augmentation framework for BUS diagnosis that integrates prompt-driven text2img synthesis with LoRA finetuning and Textual Inversion, with an img2img refinement stage. Applied to BUSI, the img2img stage further improved visual quality, preserving ultrasound characteristics relevant for diagnosis. Future work will focus on extending validation to multi-institutional, multimodal cohorts that include clinical metadata. Finally, we will assess generalizability to other modalities, including MRI, CT, and X-ray.

References

[1] A. Khalid, A. Mehmood, A. Alabrah, B. F. Alkhamees, F. Amin, H. AlSalman, and G.-S. Choi, "Breast cancer detection and prevention using machine learning," *Diagnostics (Basel)*, vol. 13, no. 19, p. 3113, 10 2023.

[2] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artificial Intelligence Review*, vol. 56, pp. 12 561–12 605, 2023. [Online]. Available: https://doi.org/10.1007/s10462-023-10453-z

[3] Y. Jiménez-Gaona, D. Carrión-Figueroa, V. Lakshminarayanan, and M. J. Rodríguez-Álvarez, "Gan-based data augmentation to improve breast ultrasound and mammography mass classification," *Biomedical Signal Processing and Control*, vol. 94, p. 106255, 2024.

[4] H. Asgariandehkordi, S. Goudarzi, M. Sharifzadeh, A. Basarab, and H. Rivaz, "Denoising plane wave ultrasound images using diffusion probabilistic models," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2024.

[5] B. Freiche, A. El-Khoury, A. Nasiri-Sarvi, M. S. Hosseini, D. Garcia, A. Basarab, M. Boily, and H. Rivaz, "Ultrasound image generation using latent diffusion models," *arXiv preprint arXiv:2502.08580*, 2025.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[8] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[9] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, W. Berman, Y. Xu, S. Liu, and T. Wolf, "Diffusers: State-of-the-art diffusion models," https://github.com/huggingface/diffusers, 2022.