Hybrid Diffusion Model for Ultrasound Image Augmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose a hybrid diffusion-based augmentation framework to overcome the critical challenge of limited and imbalanced data in medical ultrasound AI. 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 and fine-tuning using LoRA and textual inversion (TI). For the Breast Ultrasound (BUSI) dataset, our method generated realistic, class-consistent images that improved classification accuracy (90.4% $\rightarrow$ 91.7%), F1-score (88.7% $\rightarrow$ 90.4%), and achieved an AUC of 0.985. Incorporating img2img refinement further reduced the Fréchet Inception Distance (FID) to 33.29, enhancing visual fidelity without sacrificing performance. These results demonstrate that hybrid diffusion augmentation produces high-fidelity ultrasound images and strengthens downstream model reliability, offering a scalable solution to one of the most persistent barriers in clinical imaging AI.

1 Introduction

Medical imaging, particularly ultrasound, is a cornerstone of early disease diagnosis [1] due to its affordability, non-invasive nature, and real-time imaging capabilities [2]. Yet, applying deep learning for automated ultrasound interpretation remains challenging. Advances in artificial intelligence (AI) and machine learning (ML) have improved breast cancer diagnosis, enabling automated detection and classification of tumor subtypes such as invasive ductal carcinoma (IDC) and ductal carcinoma in situ (DCIS) [3]. However, progress is consistently hindered by the limited availability and imbalance of annotated ultrasound datasets, which restricts model robustness and generalizability across disease categories.

Traditional augmentation techniques such as flipping, rotations, and intensity variations provide only superficial diversity in medical imaging [4, 5]. Synthetic data generation methods, including Generative Adversarial Networks (GANs), have been explored but often fail to reproduce the speckle noise and fine-grained tissue textures that are diagnostically essential in ultrasound. Diffusion models offer a promising alternative, with greater stability, controllability, and fidelity in generating realistic samples.

Recent advances have demonstrated their effectiveness for high-quality image generation. Text-to-image (text2img) frameworks such as Palette [6] and One-Step models [7] can translate semantic cues into detailed visuals. Textual Inversion (TI) [8, 9] enables domain personalization and fine-grained control, while parameter-efficient strategies such as StyleInject [10] facilitate adaptation to specialized domains. Although text2img and other generative models have shown promising results in general images, they often fail to adapt to the domain-specific features of medical imaging, such as ultrasound. Hence, this leaves a gap in the utility and fidelity of such synthetic images in the medical domain.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In this study, we address the low fidelity of the ultrasound image generation and the critical bottleneck of class imbalance in breast ultrasound imaging, particularly the underrepresentation of malignant cases. We propose a hybrid diffusion augmentation pipeline that combines semantic conditioning (via text2img generation) with structural refinement (via image-to-image synthesis). The framework builds on Stable Diffusion v1.5, enhanced with Low-Rank Adaptation (LoRA) and TI to capture ultrasound-specific characteristics, while img2img refinement improves visual fidelity. By generating high-quality, class-consistent synthetic images for minority classes, we enrich training datasets and improve downstream classification performance. Evaluation includes Fréchet Inception Distance (FID) for realism and standard metrics such as accuracy, F1-score, and AUC-ROC for diagnostic utility.

In summary, we present a hybrid diffusion–based augmentation framework for breast ultrasound that improves the fidelity of synthetic images and mitigates class imbalance. Our contributions are as follows:

- A hybrid diffusion model framework integrating text2img generation with image-to-image refinement for realistic ultrasound synthesis.
- Domain-adaptive fine-tuning with LoRA and Textual Inversion to capture ultrasound-specific noise and texture patterns.
- Comprehensive evaluation on the BUSI dataset, including ablation studies isolating the contributions of each component.

2 Related Work

Ultrasound imaging is non-invasive, and cost-effective nature, but it suffers from speckle noise, acoustic clutter, and low signal-to-noise ratios. Diffusion models have recently been explored for enhancement and denoising in this context. Stevens et al. [11] introduced a diffusion-based dehazing strategy for echocardiography, mitigating acoustic clutter while preserving weak tissue echoes. Zhang et al. [12] combined adaptive beamforming with DDPMs for despeckling, effectively maintaining anatomical fidelity. Stojanovski et al. [13] employed semantic label maps to synthesize echocardiograms that improved segmentation performance, while Asgariandehkordi et al. [14] proposed a plane-wave denoising method that generalized well from simulation to phantom and in vivo data. Collectively, these works highlight the ability of diffusion models to address ultrasound's inherent noise and improve image interpretability.

Recent efforts apply diffusion models directly to breast ultrasound data for augmentation and diagnosis. Freiche et al. [15] explored Stable Diffusion for text2img augmentation, while Oh et al. [16] applied diffusion probabilistic models on the BUSI dataset. Lai et al. [17] introduced a lesion-focused diffusion framework to amplify tumor visibility. Kazerouni et al. [18] further contextualized these studies in a broader survey, noting ultrasound as a growing application area for generative diffusion approaches.

Beyond generative methods, deep learning has long been a cornerstone in breast ultrasound analysis. Shilaskar et al. and others [19] proposed hybrid frameworks coupling VGG-16 for classification with UNet for segmentation, achieving 90% classification and 98% segmentation accuracy. These results demonstrate the advantage of task-specific CNNs within unified diagnostic pipelines and emphasize the complementary strengths of classification and segmentation in improving computer-aided diagnosis.

Overall, the prior work illustrates two converging directions: (i) diffusion models enhance or synthesize ultrasound data to address noise and class imbalance, and (ii) discriminative deep learning models perform well for classification and segmentation. The synergy of these paradigms motivates hybrid strategies, where diffusion-based generation augments traditional classifiers, advancing robust and generalizable models for breast cancer diagnosis.

3 Hybrid Diffusion Model

The overall workflow of the proposed method is illustrated in Figure 1, which consists of three stages: preprocessing, model fine-tuning, and image synthesis.

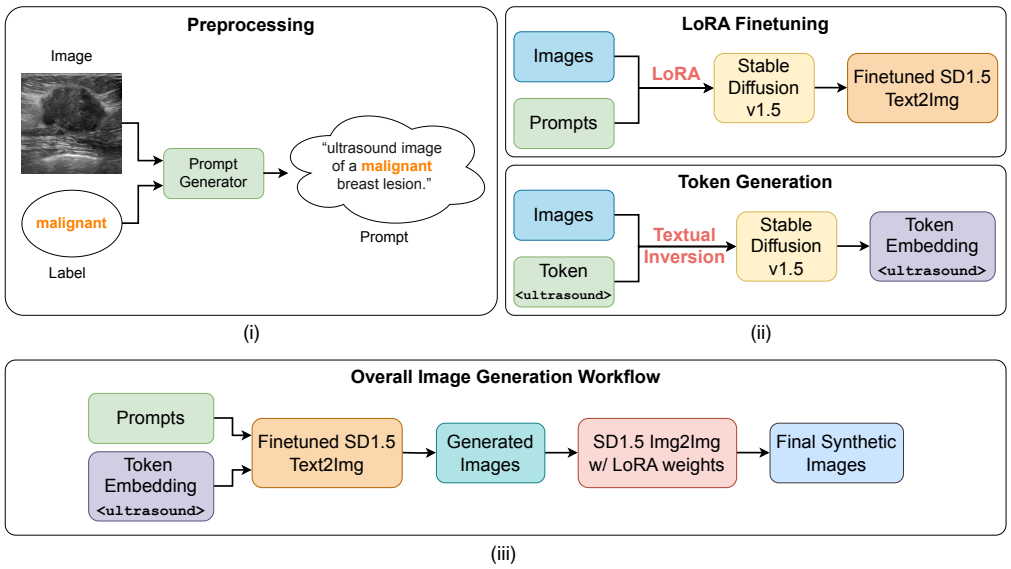

Figure 1: Overview of the proposed hybrid diffusion-based image generation framework for breast ultrasound augmentation. The pipeline consists of three main stages. (i) Preprocessing stage converts the labels into descriptive prompts. (ii) LoRA finetuning and Token Generation adapts the Stable Diffusion v1.5 using LoRA (for image–prompt alignment) and Textual Inversion (for learning domain-specific ultrasound tokens). (iii) In the final workflow, the prompts and learned token `<ultrasound>` are passed through the finetuned Text2Img model to generate synthetic images, which are further refined using an Img2Img stage with LoRA weights, yielding the final synthetic ultrasound images.

3.1 Preprocessing

We incorporated both textual and structural priors to prepare the dataset. To guide the generation process, we used class-specific prompts derived from BUSI labels (normal, benign, malignant), 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." These prompts provided semantic guidance, ensuring the diffusion model generated images aligned with diagnostic categories while preserving medical plausibility.

3.2 Model Fine-Tuning

We utilized Low-Rank Adaptation (LoRA) to fine-tune the attention mechanisms of the Stable Diffusion model efficiently. LoRA enabled parameter-efficient adaptation by injecting trainable low-rank matrices into frozen attention layers. This strategy facilitated the learning of domain-specific features, such as characteristic textures and lesion appearances, while avoiding full model retraining. The fine-tuning was performed using paired prompts and ultrasound images, allowing the model to map textual semantics to appropriate visual features.

In order to improve the model's capacity to interpret ultrasound-specific prompts, we introduced a custom token `<ultrasound>` through TI. This technique optimized a new embedding that encapsulated domain-specific patterns like speckle noise and soft tissue textures based on a curated set of representative BUSI images. The learned embedding was appended to all prompts during both training and inference, enhancing semantic alignment and promoting consistent generation across diagnostic categories.

3.3 Image Synthesis

Following the initial synthesis, we applied Stable Diffusion's img2img pipeline to enhance the realism of the generated outputs. This refinement process was performed with a low denoising strength

of 0.3, which maintained the structural integrity while improving visual quality. It effectively sharpened textures and reinforced ultrasound-specific attributes like speckle noise and soft tissue gradients, which are essential for producing clinically plausible synthetic images suitable for training downstream classification models.

4 Experimental Results

4.1 Dataset and Evaluation Metrics

We use the Breast Ultrasound Images (BUSI) dataset, which contains 780 grayscale ultrasound scans from female patients, categorized as normal (133), benign (437), or malignant (210). For benign and malignant images, the dataset also provides binary lesion segmentation masks. Prior to training, all images were resized to a uniform resolution and normalized to ensure consistency across acquisitions and reduce preprocessing artifacts. The original split (80-20) consisted of 623 training images (349 benign, 168 malignant, 106 normal) and 157 validation images (88 benign, 42 malignant, 27 normal). To mitigate class imbalance, we padded the training set with 1 benign, 182 malignant, and 244 normal synthetic images generated by our hybrid diffusion pipeline. The final training set thus contained 1,050 images evenly distributed across the three classes, while the validation set remained unchanged.

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

4.2 Training Protocol and Hardware

ResNet18 classification model were trained using the Adam optimizer with a learning rate of 0.0001, batch size of 16, and for 30 epochs. Cross-entropy loss was used for optimization. Standard data augmentations, such as horizontal flips and normalization, were applied to improve generalization. Random seeds were fixed to ensure reproducibility.

All experiments were conducted on a workstation equipped with a 12th Gen Intel Core i5-12500 processor (6 cores, 12 threads, base clock 3.0 GHz, turbo up to 4.6 GHz), 128 GB of RAM, and an NVIDIA RTX A4000 GPU with 16 GB VRAM. The system ran on a 64-bit Ubuntu Linux environment. Model training and diffusion-based image generation were implemented in PyTorch, utilizing the Hugging Face *diffusers* library.

Table 1: Ablation study showing the impact of each component in the hybrid diffusion framework on classification and image quality metrics.

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**

4.3 Ablation Study

Table 1 isolates the effect of LoRA finetuning, Textual Inversion (TI), and Img2Img refinement. Using SD1.5 text2img augmentation on top of real data improves the baseline to Acc 0.917, F1 0.905, and yields the highest AUC-ROC (0.986) with FID 45.97. Adding TI delivers the best downstream classification with Acc 0.924, F1 0.912, and PPV 0.906, which indicates that the learned domain token improves class-aware synthesis (FID 45.66). Img2Img consistently lowers FID (from 45.97 to 38.34 without TI and from 45.66 to 37.18 with TI), but modestly reduces classifier accuracy (0.905) and F1 (0.884). Across all settings AUC-ROC remains ≥ 0.975, suggesting stable class ranking despite the trade-off against visual realism.

4.4 Qualitative Analysis

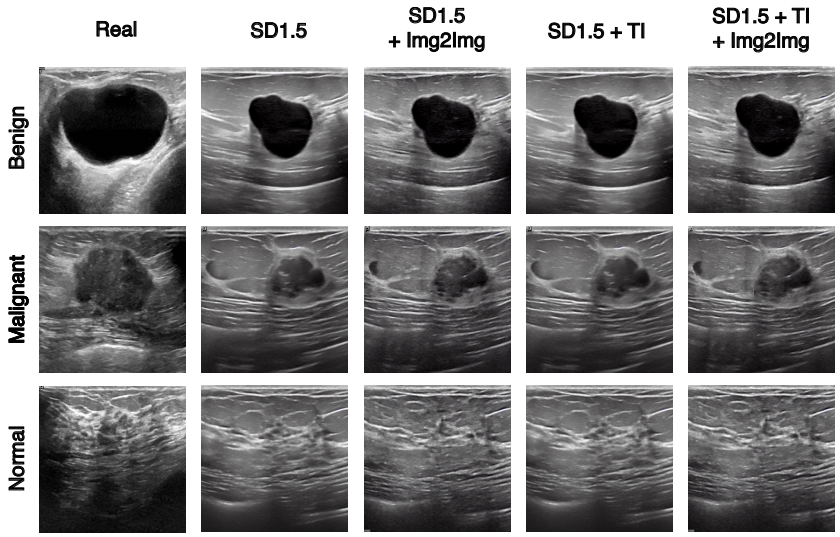

Figure 2: Comparison of real ultrasound images and synthetic variants generated by different Stable Diffusion 1.5–based approaches. Rows correspond to breast lesion categories: benign (top), malignant (middle), and normal (bottom). Columns show (from left to right): original real images, baseline SD1.5 generations, SD1.5 with img2img refinement, SD1.5 with TI, and SD1.5 combined with TI and img2img. The img2img refinement increases the fidelity by improving ultrasound-specific artifacts.

To assess the visual quality and realism of the generated images, we presented a sample grid in Figure 2, displaying synthetic ultrasound images across different diagnostic categories. The images demonstrate that the model captures critical ultrasound features such as lesion boundaries, internal textures, and background anatomy. Notably, malignant lesions display irregular, heterogeneous texture, while benign ones exhibit smoother contours with homogeneous texture. Normal cases show no lesions and variable breast tissues. In addition, img2img refinement helps preserve fine-grained tissue patterns relative to baseline SD1.5, while textual inversion enhances semantic consistency with the intended class label. Some minor artifacts and oversmoothing remain, but overall, the results suggest the synthetic images retain clinically interpretable traits. This visual validation complements the quantitative improvements, highlighting both the strengths and remaining challenges in generating diagnostically meaningful content.

5 Conclusion

We presented a hybrid diffusion–based augmentation framework for breast ultrasound that integrates prompt-driven text-to-img synthesis with LoRA finetuning and Textual Inversion, plus an img2img refinement stage. Applied to BUSI, our method balanced the training set (350 images per class) and improved downstream classification over a real-only baseline. The refinement stage further improved visual quality, preserving ultrasound characteristics relevant for diagnosis. Although after refinement, it achieved the lowest FID, it also reveals a realism–utility trade-off.

Future work will focus on closing this gap via task-aware generation: conditioning on lesion masks or structure-preserving priors, jointly optimizing with a diagnostic encoder, and weighting synthetic samples by classifier confidence. We will also extend validation to multi-institutional, multimodal cohorts that include clinical metadata. Finally, we will assess generalizability to other modalities, including MRI, CT, and X-ray, to evaluate robustness and clinical utility.

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] T. Islam, M. A. Sheakh, M. S. Tahosin *et al.*, "Predictive modeling for breast cancer classification in the context of bangladeshi patients by use of machine learning approach with explainable ai," *Scientific Reports*, vol. 14, p. 8487, 2024.

[3] A. Al-Rohaimi and S. Alshahrani, "Artificial intelligence in breast cancer diagnosis: a comprehensive review," *Cancers*, vol. 12, no. 9, p. 2564, 2020.

[4] F. Garcea, A. Serra, F. Lamberti, and L. Morra, "Data augmentation for medical imaging: A systematic literature review," *Computers in Biology and Medicine*, vol. 152, p. 106391, 2023. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2022.106391

[5] 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

[6] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–10, 2022.

[7] G. Parmar, T. Park, S. Narasimhan, and J.-Y. Zhu, "One-step image translation with text-to-image models," *arXiv preprint arXiv:2303.11305*, 2023.

[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] J. Yang, H. Wang, Y. Zhang, R. Xiao, S. Wu, G. Chen, and J. Zhao, "Controllable textual inversion for personalized text-to-image generation," *arXiv preprint arXiv:2303.08048*, 2023.

[10] Y. Bai, M. Zhou, and Q. Yang, "Styleinject: Parameter efficient tuning of text-to-image diffusion models," *arXiv preprint arXiv:2310.14147*, 2023.

[11] T. S. Stevens, F. C. Meral, J. Yu, I. Z. Apostolakis, J.-L. Robert, and R. J. Van Sloun, "Dehazing ultrasound using diffusion models," *IEEE Transactions on Medical Imaging*, vol. 43, no. 10, pp. 3546–3558, 2024.

[12] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[13] D. Stojanovski, U. Hermida, P. Lamata, A. Beqiri, and A. Gomez, "Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation," in *International Workshop on Advances in Simplifying Medical Ultrasound*. Springer, 2023, pp. 34–43.

[14] 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.

[15] 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.

[16] S.-H. Oh, G. Jung, M. Kim, Y.-M. Kim, H.-J. Lee, S.-Y. Kim, H.-S. Kwon, and H.-M. Bae, "Breast tumor image synthesis based on diffusion probabilistic model," in *2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS)*. IEEE, 2024, pp. 1–4.

[17] Y. Lai, Y. Liu, Q. Zhang, J. Shang, X. Qiu, and J. Yan, "Lesiondiff: enhanced breast cancer classification via dynamic lesion amplification using diffusion models," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, vol. 12, no. 1, 2024.

[18] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof, "Diffusion models in medical imaging: A comprehensive survey," *Medical image analysis*, vol. 88, p. 102846, 2023.

[19] Anonymous, "Breast cancer detection using vgg-16 and unet architectures on ultrasound images," *Unpublished Abstract*, 2024, manuscript excerpt provided in project materials.