Data Augmentation for Small Breast Cancer Imaging Datasets Using Generative Adversarial Networks

***Abstract*—The limited availability of annotated medical imaging data remains a critical barrier in developing robust deep learning models, particularly in breast Ultrasound (US). This study presents a novel conditional Generative Adversarial Network (GAN) framework tailored for synthesizing high-quality breast US images in data-constrained environments. Our method is evaluated on 100 US images (on one of the smallest publicly available breast US dataset) from the Open Access Series of Breast US Dataset (OASBUD), simulating real-world clinical scarcity. By generating standardized B-mode images directly from raw radio- frequency (RF) data, our framework minimizes inter-vendor variability and produces diagnostically coherent images to support segmentation and classification tasks. Our framework combines Progressive Growing GAN and StyleGAN architectures, integrating spectral normalization, residual connections, and a Pearson correlation-based loss function to preserve structural features. Performance evaluation shows that images synthesized by the proposed model achieved an accuracy of 82%, with balanced F1-scores of 0.83 for both benign and malignant cases. Additionally, segmentation performance on generated images achieves a Dice score of 0.74. Overall, the proposed GAN-based augmentation strategy effectively enhances deep learning performance in medical imaging tasks and provides a practical solution for training models with limited annotated data and strengthens CAD system performance in both segmentation and classification tasks.**

***Index Terms*—Generative adversarial networks, Data augmentation, Breast Ultrasound, Deep learning.**

I. Introduction

Breast cancer remains one of the most prevalent malignancies affecting women globally, with early detection playing a critical role in improving patient survival rates [1]. Ultrasound (US) imaging has emerged as an important diagnostic tool, particularly valuable for patients with dense breast tissue or contraindications to mammography. Its non-invasive nature, real-time capabilities, absence of radiation exposure, and cost-effectiveness make it attractive for various clinical scenarios, especially in resource-limited settings. However, manual US interpretation remains time-consuming and subject to significant inter-observer variability, potentially leading to diagnostic inconsistencies and delayed decision-making. These limitations have necessitated Computer-Aided Diagnosis (CAD) systems to provide assistance in segmentation as well as classification problems. Contemporary deep learning-based CAD systems demonstrate remarkable performance but require extensive, annotated datasets for optimal results. Unfortunately, assembling such datasets in the medical domain remains a formidable challenge due to strict patient privacy regulations, the high cost and time burden of expert annotations, institutional data-sharing constraints, and the inherent scarcity of certain pathological cases [2]. These factors have led to a growing interest in synthetic data generation to augment existing datasets and support robust CAD model development.

Synthetic data generation, particularly through the application of Generative Adversarial Networks (GANs) has gained significant attention to address these limitations [3]–[5]. GANs are generally better than alternative methods like Variational Autoencoders, diffusion models, or traditional augmentation for small datasets because their adversarial training enables to model complex, high-dimensional data distributions and produce sharper, more realistic samples [6]. GANs offer a unique capability to augment existing datasets, thereby enhancing model robustness and mitigating the risks of overfitting. GAN-based medical image synthesis spans diverse applications including liver lesion generation [7], mammogram translation [8], brain lesion synthesis [9]. While these approaches commonly suffer from training instability, high computational requirements, and dependence on extensive datasets [4], [5], [10]. For low-dimensional datasets, the least-squares GAN (LSGAN) algorithm has been proposed to enhance classification performance in small medical datasets which contains various features and instances. [11]. For US based breast cancer diagnosis, efforts have explored hybrid methods combining GAN-generated augmentations with architectures like U-Net for improved segmentation [12], and semi-supervised GANs for classifying breast masses in US images [13]. Moreover, public breast databases of US and mammography are merged to train different GAN models [3]. However, many of these GAN models have typically been applied to comparatively large image datasets for effective augmentation. Rather than standard B-mode images, US radio frequency (RF) signals have been explored for image segmentation using spectral data augmentation while the reported Dice score remained below 70 % [14].

Accordingly, we propose a novel conditional GAN framework optimized for high-quality synthetic breast US image generation from small datasets. The breast US is chosen as the limited dataset can lead to an imbalance problem for diseases like malignant tumors [15]. Unlike prior works that rely on vendor-specific B-mode images, we generate standardized B-mode images from raw RF signals, thereby ensuring uniform preprocessing, reducing inter-vendor differences, and maintaining consistent image quality. Furthermore, synthesizing B-

mode images from RF data allows us to maintain consistent image characteristics across all samples. Our experiments are conducted on one of the smallest publicly available breast US datasets — the OASBUD dataset, which contains only 100 lesion samples. Our proposed model leverages a combination of Progressive Growing GAN (PGGAN) [16] and StyleGAN architectures to synthesize high-resolution US images suitable for training CAD systems in low-data settings [17]. Our design integrates spectral normalization, residual blocks, progressive growing, and an innovative Pearson correlation-based loss function to preserve structural and pathological features essential for clinical utility. PGGAN enables stable training and fine-grained detail synthesis by increasing image resolution during training, which is particularly advantageous in limited-data regimes. StyleGAN, on the other hand, introduces style-based modulation and stochastic variation control, allowing for disentangled representation learning and finer control over image features such as texture, shape, and pathological attributes. The approach not only addresses the scarcity of annotated medical data but also promotes practical deployment in resource-limited healthcare environments.

The contribution is summarized as follows:

- A combination of PGGAN and StyleGAN architectures has been introduced integrating a Pearson correlation-based auxiliary loss function to synthesize breast US images from extremely limited raw RF datasets.
- By generating B-mode images directly from raw RF signals, our approach minimizes inter-vendor variability, ensures uniform preprocessing, and preserves critical structural and pathological features necessary for accurate diagnosis.
- The synthesized images are evaluated using multiple performance metrics for both benign–malignant classification and segmentation, with a comparative analysis against existing methods for deployment in resource-limited healthcare environments.

Section II presents the methodology containing GAN architecture, training protocol, segmentation workflow. Section III represents results using different metrices, limitations and future scope. Finally concluding remarks in section IV.

II. Methodology

A. Dataset

Our dataset consists of only 100 distinct lesions, each with two corresponding US scans, for a total of 200 images. The OASBUD dataset table [18] contains raw radio-frequency US signals. In Figure 1, we present two distinct representations of the signal. The first panel displays the signal's image representation, while the second panel illustrates its one-dimensional (1D) representation.

For our GAN training:

- Converted RF signals to B-mode images using Hilbert transform
- Resized to 256×256 pixels and normalized to [-1,1]
- Applied rigorous augmentation:

TABLE I
OASBUD Dataset Characteristics

Parameter	Specification
Acquisition Period	Nov 2013 - Oct 2015
Patients	78 women
Scans per Lesion	2 orthogonal views
Transducer	L14-5/38 linear array
Original Resolution	1280×1024 pixels
Annotations	BI-RADS, lesion masks

  - Random flips (p=0.5)
  - Affine transformations ($\pm 10^\circ$ rotation)
  - Gaussian noise ($\sigma = 0.01$)
  - Intensity variations ($\pm 20\%$)
- Class-balanced batches (16 images each)

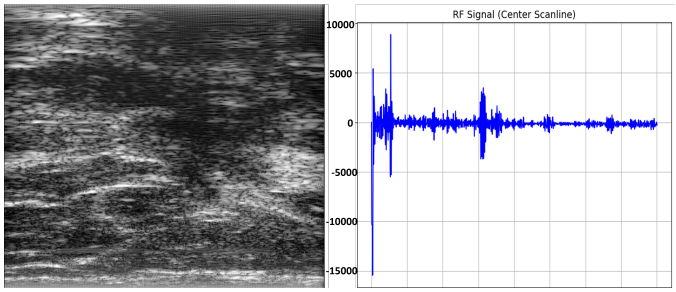

Fig. 1. Visualization of the OASBUD dataset: (Left) Image representation of a breast ultrasound sample showing the lesion region, and (Right) its corresponding one-dimensional (1D) radio-frequency signal representation. This illustrates the dual nature of the data used for GAN training, highlighting both spatial and signal-based information.

B. Network Architecture

1) Generator: $G(z, y)$: The generator is tasked with mapping a latent vector $z \in \mathbb{R}^{512}$, drawn from a standard normal distribution, and a class label $y \in \{0, 1\}$ to a synthetic image. The label y is first embedded into $\mathbb{R}^{512}$ via a learned dense layer. To achieve effective conditioning, we combine this embedding with the latent vector using element-wise multiplication and an affine shift: $z' = z \odot (1 + W_y)$, where W is a learned matrix in $\mathbb{R}^{512\times512}$. This design promotes disentanglement between class information and the underlying generative factors.

The core of the generator features a hierarchical, multi-level structure. It begins at a 16×16 spatial resolution, progressively upsampling to 256×256. Each upsampling step doubles the spatial resolution and halves the number of channels, and each scale contains two residual blocks to facilitate efficient gradient flow and feature reuse [19], [20]. Pixel-wise normalization is applied after each residual block to counteract internal covariate shift [21]. Multi-scale skip connections aggregate feature maps from each resolution, promoting both global and local consistency in the final output. The output head consists of a 1×1 convolution followed by a $\tanh$ activation, restricting the output to the $[-1, 1]$ range.

To better capture subtle stochastic variations such as speckle noise in US images, we also inject learned per-pixel noise at the output.

2) Discriminator: $D(x, y)$: The discriminator receives an image $x \in \mathbb{R}^{256\times256}$ and associated label y. The label is projected via bilinear upsampling to match the spatial dimensions of the image and concatenated along the channel axis, yielding a tensor of shape $2 \times 256 \times 256$. The initial feature extraction is performed by a 7×7 convolution with stride 2, followed by a series of residual blocks that increase the channel width from 64 to 1024. Each block is regularized with spectral normalization [22] and followed by average pooling to reduce spatial dimensions. LeakyReLU activations with a negative slope of 0.2 are used throughout for improved gradient propagation [23].

The discriminator's output layer consists of global average pooling, followed by a spectrally-normalized dense layer. To further encourage diversity and penalize mode collapse, a minibatch standard deviation feature is concatenated before the final decision. The architectures of the discriminator and generator can be seen in Table II and Table III, respectively.

C. Loss Functions and Optimization

The overall generator loss, $\mathcal{L}_G$, harmonizes three objectives:

$$\begin{aligned}\mathcal{L}_G = &\underbrace{\mathbb{E}\big[\log(1 - D(G(z, y)))\big]}_{\mathcal{L}_{\text{adv}}} \\ &+ \lambda_{\text{pix}} \underbrace{\|G(z, y) - x\|_1}_{\mathcal{L}_{\text{pix}}} \\ &+ \lambda_{\text{pear}} \underbrace{(1 - \rho(G(z, y), x))}_{\mathcal{L}_{\text{pear}}}\end{aligned} \tag{1}$$

Here, $\mathcal{L}_{\text{adv}}$ is the adversarial loss, $\mathcal{L}_{\text{pix}}$ is a pixel-wise ℓ_1 loss to ensure fidelity at the voxel level, and $\mathcal{L}_{\text{pear}}$ is a Pearson correlation loss that enforces preservation of structural relationships.

1) WGAN-GP Adversarial Loss: Following the Wasserstein GAN with Gradient Penalty (WGAN-GP) paradigm [24], the adversarial loss is written as:

$$\begin{aligned}\mathcal{L}_{\text{adv}} = &\mathbb{E}[D(x, y)] - \mathbb{E}[D(G(z, y), y)] \\ &+\lambda_{\text{gp}}\mathbb{E}\left[(\|\nabla_{\hat{x}} D(\hat{x}, y)\|_2 - 1)^2\right]\end{aligned} \tag{2}$$

where $\hat{x}$ is an interpolated sample between a real and a generated image: $\hat{x} = \epsilon x + (1 - \epsilon)G(z, y)$, with $\epsilon \sim \mathcal{U}(0, 1)$. The gradient penalty term ensures the discriminator remains within a 1-Lipschitz constraint, which is crucial for style GAN training.

2) Pearson Correlation Loss: To preserve anatomical and structural features, the Pearson correlation coefficient is calculated as:

$$\rho(X, Y) = \frac{\sum_{i,j}(X_{ij} - \mu_X)(Y_{ij} - \mu_Y)}{\sqrt{\sum_{i,j}(X_{ij} - \mu_X)^2 \sum_{i,j}(Y_{ij} - \mu_Y)^2}} \tag{3}$$

This metric is computed within local 11×11 receptive fields, focusing computation within lesion regions via masking, and is adapted for the dynamic intensity range present in US images [25].

TABLE II
STYLEGAN-BASED GENERATOR WITH PGGAN DISCRIMINATOR HYBRID MODEL (MODEL 0)

Component	Architecture Details
Generator (MedicalGenerator)	
Style Mapping	Linear(101, 512) → LeakyReLU → Linear(512, 512) → LeakyReLU
Fully Connected Layer	Linear(512, 65536)
Block 0	Upsample(×2) → Conv2d(1024, 512, k=3, p=1) → BN → LeakyReLU → ResBlock(512, 512)
Block 1	Upsample(×2) → Conv2d(512, 256, k=3, p=1) → BN → LeakyReLU → ResBlock(256, 256)
Block 2	Upsample(×2) → Conv2d(256, 128, k=3, p=1) → BN → LeakyReLU → ResBlock(128, 128)
Block 3	Upsample(×2) → Conv2d(128, 64, k=3, p=1) → BN → LeakyReLU → ResBlock(64, 64)
Final Layer	Conv2d(64, 32, k=3, p=1) → LeakyReLU → Conv2d(32, 1, k=3, p=1) → Tanh
Discriminator (MedicalDiscriminator)	
Initial Layer	Conv2d(2, 64, k=4, s=2, p=1) → LeakyReLU
Main Block	Conv2d(64, 128, k=4, s=2, p=1) → IN → LeakyReLU → Conv2d(128, 256, k=4, s=2, p=1) → IN → LeakyReLU → Conv2d(256, 512, k=4, s=2, p=1) → IN → LeakyReLU → Conv2d(512, 1, k=4, s=1)
Residual Block Details	
ResBlock Structure	Conv2d(k=3, p=1) → BN → LeakyReLU → Conv2d(k=3, p=1) → BN + Skip

Abbreviations: BN: Batch Normalization, IN: Instance Normalization, SN: Spectral Normalization, ReLU: Rectified Linear Unit, LeakyReLU: Leaky Rectified Linear Unit, FC: Fully Connected, Conv2d: 2D Convolution, k: kernel size, p: padding, s: stride, ↓: downsample, ↑: upsample, Tanh: hyperbolic tangent activation, AdaptiveAvgPool2d: adaptive average pooling. **Residual Block (ResBlock)** includes two convolutional layers with normalization and activation, plus a skip connection.

D. Training Protocol

1) Early Stopping and Validation: Generalization is critical when training on limited data. To mitigate overfitting, we employ a stringent early stopping framework, whereby model progress is quantitatively monitored using the Pearson correlation coefficient (PCC) evaluated on a stratified hold-out validation set. This set is constructed from 20% of the data, ensuring proportional representation of both diagnostic classes. Training is mandated to continue for at least 200 epochs, safeguarding against premature convergence and allowing the model to fully exploit the available data. However, the training is halted if PCC not improve by at least 0.01 over any consecutive window of 50 epochs. This dual threshold ensures a balance between sufficient learning and prudent regularization, fostering both expressiveness and stability in the resulting models.

2) Mixed-Precision Training: To optimize memory usage and accelerate training, we utilize NVIDIA's Automatic Mixed Precision (AMP), which allows dynamic computation in both 16-bit and 32-bit floating point formats. We initialize the

TABLE III
PG GAN WITH RESIDUAL BLOCKS ARCHITECTURE (MODEL 1)

Component	Architecture Details
Generator (MedicalGenerator)	
Label Embedding	Linear(1, 100) → LeakyReLU
Dense Layer	SN(Linear(100, 32768))
Block 1 (8×8→16×16)	ResBlock(512, 512) → ResBlock(512, 512)
Block 2 (16×16→32×32)	Upsample(×2) → ResBlock(512, 256) → ResBlock(256, 256)
Block 3 (32×32→64×64)	Upsample(×2) → ResBlock(256, 128) → ResBlock(128, 128)
Block 4 (64×64→128×128)	Upsample(×2) → ResBlock(128, 64) → ResBlock(64, 64)
Block 5 (128×128→256×256)	Upsample(×2) → ResBlock(64, 32) → ResBlock(32, 32)
Final Conv	SN(Conv2d(32, 1, k=3, p=1)) → Tanh
Discriminator (MedicalDiscriminator)	
Label Embedding	Linear(1, 65536) → LeakyReLU
Initial Conv	SN(Conv2d(2, 64, k=3, p=1))
Block 1 (256×256→128×128)	ResBlock(64, 128, ↓) → ResBlock(128, 128)
Block 2 (128×128→64×64)	ResBlock(128, 256, ↓) → ResBlock(256, 256)
Block 3 (64×64→32×32)	ResBlock(256, 512, ↓) → ResBlock(512, 512)
Block 4 (32×32→16×16)	ResBlock(512, 1024, ↓) → ResBlock(1024, 1024)
Final Layers	AdaptiveAvgPool2d(1) → SN(Linear(1024, 1))
Residual Block Details	
ResBlock Structure	SN(Conv2d) → LeakyReLU → SN(Conv2d) + Skip
Skip Connection	SN(Conv2d(1×1)) if channel mismatch, else Identity

Abbreviations: BN: Batch Normalization, IN: Instance Normalization, SN: Spectral Normalization, ReLU: Rectified Linear Unit, LeakyReLU: Leaky Rectified Linear Unit, FC: Fully Connected, Conv2d: 2D Convolution, k: kernel size, p: padding, s: stride, ↓: downsample, ↑: upsample, Tanh: hyperbolic tangent activation, AdaptiveAvgPool2d: adaptive average pooling.

gradient scaler at 2^{16}, incrementing every 200 steps to preclude gradient underflow or overflow [23]. This method enables us to increase the batch size from 8 (in full-precision FP32 mode) to 16, resulting in greater intra-batch diversity and more robust gradient signals. Empirical profiling indicates up to 40% memory savings, thereby facilitating deeper models without additional hardware requirements.

3) Progressive Growing: To further enhance stability and enable detailed synthesis at high resolutions, we introduce a progressive growing strategy, inspired by StyleGAN [26]. The generative network begins at a coarse 64×64 resolution, gradually expanding through five phases to the target 256×256 resolution. Advancement to each successive phase is contingent on the Fréchet Inception Distance (FID) plateauing, defined as a less than 5% improvement over 20 epochs. During phase transitions, new layers are smoothly blended into the model via an exponential moving average, $\alpha_{\text{new}} = 0.1\alpha + 0.9(1-\alpha)$, which ensures gentle adaptation and avoids destabilizing the learning process [16]. This incremental approach allows the generator to first capture global anatomical structure before focusing on fine details, a critical feature for clinical realism.

4) Spectral Normalization: To further stabilize adversarial training and control the Lipschitz constant of the discriminator, we apply spectral normalization (SN) to all convolutional (and dense) layers in both the generator and discriminator [22]. Each weight matrix W is normalized by its maximum singular value, $W_{SN} = W/\sigma(W)$, where $\sigma(W)$ is computed using three power iterations per update. The Lipschitz constant is explicitly targeted at $K = 1.0$. This constraint is essential for preventing the discriminator from becoming excessively sensitive, which can otherwise lead to vanishing gradients for the generator and, ultimately, to mode collapse [23].

5) Label Smoothing: To further regularize the discriminator and reduce overconfidence, we employ label smoothing with a factor $\epsilon = 0.1$. In this setup, real images are assigned a label of 0.9 and generated (fake) images a label of 0.1, rather than the traditional $1.0/0.0$ binary targets. This softening of the targets impedes the discriminator from learning sharp, brittle boundaries, instead encouraging it to focus on more generalizable, semantically meaningful features.

6) Alternating Optimization and Scheduling: The training regimen alternates between discriminator and generator updates, with the discriminator typically updated 3 to 5 times per generator step. This schedule is designed to maintain a power balance between the competing networks. The generator is updated with gradients clipped to the range ± 0.01 to prevent instability. Both networks are optimized using Adam with hyperparameters $\beta_1 = 0$, $\beta_2 = 0.99$, and an initial learning rate of 10^{-4}, which is decayed smoothly via a cosine annealing schedule to promote gradual convergence.

7) Auxiliary Classifier and Evaluation: To quantitatively assess the quality and realism of generated images, an auxiliary deep classifier is trained in parallel with the GAN. This classifier provides additional metrics such as classification accuracy, proxy Inception Score [27], and feature embeddings for FID calculation. In addition to these, we log common image similarity metrics: Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Pearson correlation. Performance curves and representative sample grids are plotted and saved every 10 epochs, offering both quantitative and qualitative feedback throughout training.

Algorithm 1 outlines the StyleGAN training loop. At each iteration, real images and labels are sampled alongside latent noise vectors to generate fake images. The discriminator is trained using the WGAN-GP loss, which includes a gradient penalty term for stability. The generator is updated less frequently (every N_{CRITIC} steps) to match the WGAN setup. It optimizes a composite loss involving adversarial, L1, and Pearson correlation components. Additionally, an auxiliary classifier is trained to enforce label consistency in generated images. Learning rate schedules, evaluation, and model checkpointing are applied at each epoch.

E. Segmentation Workflow

To further assess the fidelity and utility of the generated images, we introduce a segmentation-based evaluation protocol using a U-Net architecture [28]. This workflow serves a dual

Algorithm 1 StyleGAN Training Procedure

Initialize: Generator G, Discriminator D, auxiliary classifier, and history buffers
for epoch = 1 to 1000 **do**
  for each batch in the dataloader **do**
    Sample real images x_{real}, labels y
    Sample latent vectors $z \sim \mathcal{N}(0, I)$
    Generate fake images $\hat{x}_{\text{fake}} = G(z, y)$
    Discriminator update:
    Compute $D(x_{\text{real}}, y)$ and $D(\hat{x}_{\text{fake}}, y)$
    Compute gradient penalty on $\hat{x}$
    Update D with WGAN-GP loss
    if batch index mod $N_{\text{CRITIC}} = 0$ **then**
      Generator update:
      Compute adversarial, L1, and Pearson losses
      Update G with total loss and apply gradient clipping
    end if
    Train auxiliary classifier on real and fake images
  end for
  Update learning rates
  Evaluate metrics on validation set
  Save checkpoints and generated samples
end for

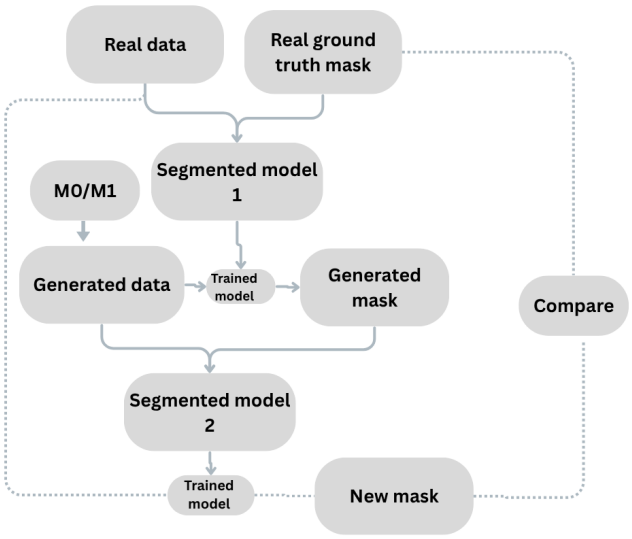

Fig. 2. Illustration of the segmentation-based evaluation workflow: Real ultrasound images with expert-annotated masks are used to train a U-Net segmentation model. The trained U-Net is then used to generate pseudo-ground truth masks for GAN-synthesized images, forming paired synthetic datasets. A segmentation model trained solely on synthetic data is evaluated on real images to assess the clinical realism and anatomical fidelity of the generated samples. This comprehensive workflow bridges GAN-based image generation and downstream clinical segmentation tasks.

purpose: it both quantifies the anatomical realism of synthetic samples and demonstrates their applicability for downstream clinical tasks.

1) Segmentation Model Training: Initially, we train a U-Net segmentation model exclusively on the real dataset, utilizing paired US images and their expert-annotated ground truth masks. The U-Net is optimized with a Dice loss, employing standard data augmentations to enhance robustness. This model establishes a high-quality baseline for lesion segmentation performance on authentic data.

2) Generating Pseudo-Ground Truth for Synthetic Images: Once trained, the segmentation model is deployed to generate pseudo-ground truth masks for the images synthesized by the GAN. Each generated image is passed through the U-Net, producing a corresponding segmentation mask. This approach facilitates objective, pixel-level assessment of anatomical plausibility in synthetic images, as the masks reflect clinically relevant structures learned from real data.

3) Synthetic Data Augmentation and External Validation: To rigorously test the utility of synthetic data, we construct a new dataset comprising generated images paired with their pseudo-ground truth segmentations. A separate segmentation model is then trained solely on this synthetic dataset. The trained model is evaluated on a hold-out set of real images, allowing us to directly measure the transferability and practical value of GAN-generated data.

Comparative analysis is performed by evaluating standard segmentation metrics (e.g., Dice coefficient, Intersection over Union, pixel-wise accuracy) on both real and synthetic-trained models. This closed-loop process provides a robust, quantitative validation of the anatomical and structural integrity of the generative framework. As shown in Figure 2, the segmentation-based evaluation workflow uses synthetic and real datasets to assess the utility of GAN-generated images. The segmentation-centric validation framework closes the loop between generative modeling and clinical application, providing a comprehensive, quantitative, and visual assessment of model performance. By comparing segmentation accuracy on real test images using models trained on both real and synthetic data, we offer strong evidence for the anatomical plausibility and translational potential of the generated images.

F. Classification Performance

Table IV presents the performance of a pretrained classifier on real and GAN-generated medical images across various models. The metrics include accuracy, precision, recall, F1-score, and class support for B and M cases. On real images, the classifier achieved excellent results with 95% accuracy and balanced metrics across both classes, confirming its reliability on authentic data. Model 0 (M0-G) generated images led to an accuracy of 82%, with slightly imbalanced precision and recall but relatively consistent F1-scores (0.83 for B, 0.81 for M). Model 1 (M1-G) shows lower performance (72% accuracy) and greater class imbalance—benign recall was high (0.94), but precision was low (0.65). Malignant recall dropped to 0.50, indicating many misclassifications, likely due to poor image realism. In hybrid dataset (real+generated), M0-H maintained 82% accuracy while balancing performance across classes (F1

around 0.83). M1-H improved over M1-G (79% accuracy), with more balanced F1-scores (0.80 for B, 0.79 for M), showing the benefit of mixing real and generated data.

TABLE IV
CLASSIFICATION PERFORMANCE SUMMARY

Model	Acc	Class	Prec	Rec	F1	Sup
Real	0.95	B	0.95	0.95	0.95	21
		M	0.95	0.95	0.95	19
M0-G	0.82	B	0.79	0.88	0.83	50
		M	0.86	0.76	0.81	50
M1-G	0.72	B	0.65	0.94	0.77	50
		M	0.89	0.50	0.64	50
M0-H	0.82	B	0.81	0.83	0.82	143
		M	0.84	0.82	0.83	154
M1-H	0.79	B	0.75	0.85	0.80	143
		M	0.84	0.74	0.79	154

B: Benign, M: Malignant, M0: Model 0, M1: Model 1, H: Hybrid,G:generated, acc: Accuracy, prec: Precision, rec: Recall, f1: F1-score, sup: Support

III. RESULTS

The result section presents the training metrics, classification outcomes, and segmentation performance for both Model 0 and Model 1.

A. Generated Data Evaluation

Figures 3 to 9 present the key performance metrics over 1000 training epochs, offering a comprehensive assessment of the quality, stability, and fidelity of the synthetic images produced by the proposed GAN models. Each metric highlights a specific dimension of generative performance, as described below:

- **Generator and Discriminator Loss** — Figure 3 illustrates the adversarial loss trends for both the generator and discriminator using M0 and M1 models. Style adversarial training is characterized by smooth convergence with controlled oscillations. Deviations such as sharp spikes or collapses may suggest mode collapse or unstyle learning dynamics. The loss functions used for standard GAN training are defined as:

$$\mathcal{L}_D = -\mathbb{E}[\log D(x)] - \mathbb{E}[\log(1 - D(G(z)))], \quad (4)$$

$$\mathcal{L}_G = -\mathbb{E}[\log D(G(z))], \quad (5)$$

  where D(x) denotes the discriminator's output on real data and D(G(z)) on generated data.
- **Structural Similarity Index (SSIM)** — SSIM quantitatively evaluates perceptual similarity between real and generated images by comparing luminance, contrast, and structure:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (6)$$

  where μ and σ denote the means and standard deviations of images x and y, and C_1, C_2 are stabilization constants. Figure 4 depicts the evolution of SSIM during training, where higher SSIM values indicate improved preservation of structural integrity and texture fidelity in synthetic images [3].
- **Fréchet Inception Distance (FID)** — FID assesses the similarity of feature distributions between real and synthetic images by leveraging deep network embeddings:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}),$$

  where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of features from real and generated image distributions, respectively. In Figure 5, the trajectory of FID over epochs demonstrates how closely the GAN-generated images approach the statistical distribution of real mammograms, with lower FID scores indicating better visual realism and distributional alignment.
- **Pearson Correlation Coefficient** — Figure 6 captures the linear correlation between pixel intensities of real and generated images: A higher correlation value implies stronger structural similarity, indicating that the GAN effectively preserves global and local structures critical for medical interpretation.
- **Peak Signal-to-Noise Ratio (PSNR)** — PSNR evaluates the fidelity of the generated images at the pixel level by measuring the logarithmic ratio between the maximum possible signal and the error introduced during generation:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{\text{MSE}}\right),$$

  where MAX_I represents the maximum pixel intensity (typically 255), and MSE is the mean squared error. Figure 7 shows PSNR trends, where higher values indicate superior preservation of fine image details and reduced distortion in synthetic samples.
- **Auxiliary Classifier Accuracy** — This figure reports the classification accuracy achieved by a downstream model trained to distinguish between benign and malignant cases using generated images:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

  where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Figure 8 demonstrates the evolving diagnostic utility and class consistency of GAN-generated images throughout training.
- **Inception Score (IS, proxy)** — IS estimates both the quality and diversity of generated samples by evaluating the entropy of predicted class distributions:

$$\text{IS} = \exp\left(\mathbb{E}_x\left[\text{KL}(p(y|x)\|p(y))\right]\right),$$

  where $p(y|x)$ is the conditional label distribution given an image x, and $p(y)$ is the marginal distribution. Figure 9

TABLE V
SEGMENTATION PERFORMANCE COMPARISON OF MODELS

Model	Class	Accuracy	Recall	Specificity	Precision	Dice / IoU
Model 0	**Benign and Malignant (Mean)**	0.96	0.69	0.69	0.8578	0.7478 / 0.6533
Model 1	**Benign and Malignant (Mean)**	0.96	0.65	0.65	0.8755	0.7085 / 0.6189

displays IS progression, where a higher score reflects improved sample diversity and clear class-specific features in synthesized images.

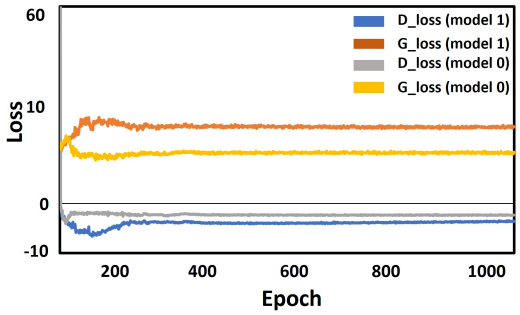

Fig. 3. Generator and Discriminator Loss over 1000 epochs for M0 and M1 models. This figure illustrates the adversarial loss profiles for both the generator and discriminator across training epochs, highlighting periods of convergence, oscillation, or instability. style adversarial training manifests as smooth, controlled loss curves, while abrupt spikes or drops may indicate mode collapse or divergence.

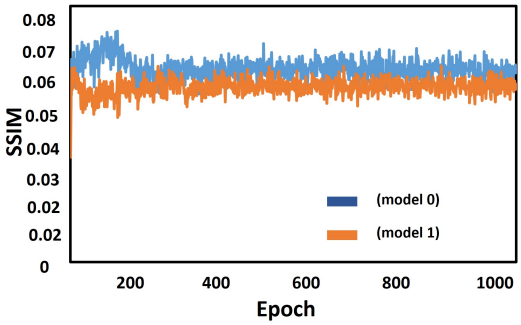

Fig. 4. Structural Similarity Index (SSIM) across training epochs. This metric tracks the evolution of perceptual similarity between real and GAN-generated images, capturing how well the synthetic data maintains luminance, contrast, and structural patterns essential for clinical reliability.

RESULTS AND DISCUSSIONS

1. Generator and Discriminator Loss

The adversarial losses for both networks converge smoothly without large oscillations or divergence, indicating that training remains stable. Controlled loss dynamics suggest that the generator and discriminator are learning complementary features, reducing the likelihood of mode collapse and supporting consistent image synthesis across epochs.

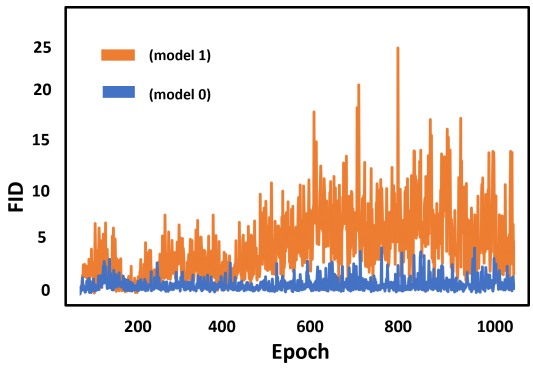

Fig. 5. Fréchet Inception Distance (FID) proxy score over epochs. FID measures the distributional similarity between real and generated images in the embedding space of a deep network, with lower values indicating greater statistical fidelity and improved visual realism in GAN outputs.

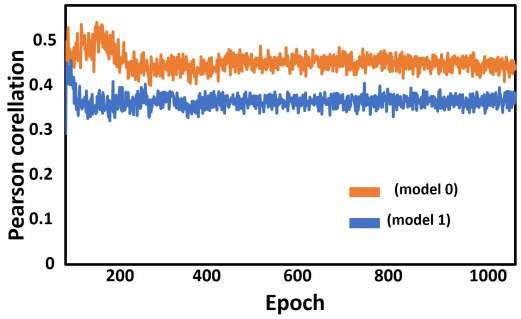

Fig. 6. Pearson Correlation across training epochs. This plot reveals the degree of linear correspondence between real and generated image pixel intensities, reflecting how effectively the GAN preserves both global and local anatomical structures throughout training.

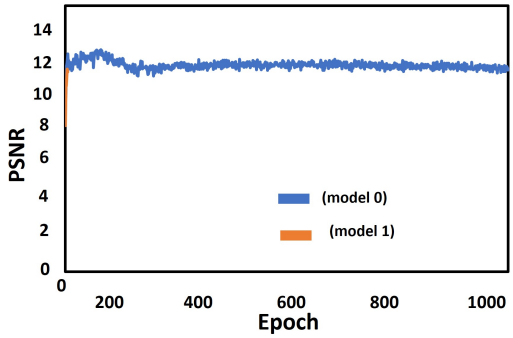

Fig. 7. Peak Signal-to-Noise Ratio (PSNR) across epochs. PSNR quantifies the level of detail and signal integrity retained in generated images, with higher values denoting less distortion and more faithful pixel-wise replication of real breast ultrasound data.

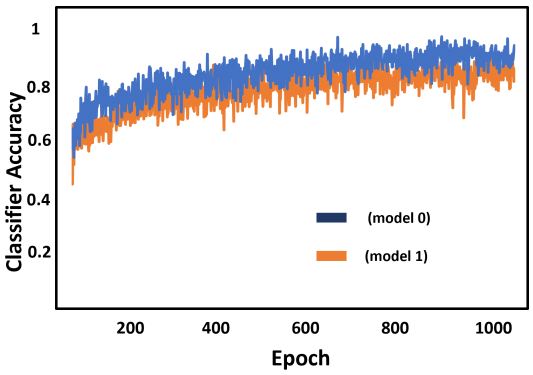

Fig. 8. Auxiliary classifier accuracy during GAN training. This figure presents the accuracy of a classifier distinguishing benign from malignant lesions using synthetic images, serving as a measure of the semantic and diagnostic quality of GAN outputs as training progresses.

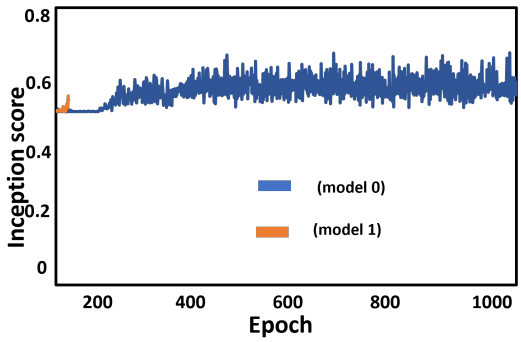

Fig. 9. Inception Score (proxy) over training epochs. IS provides an estimate of both the quality and diversity of generated samples, with higher scores signifying greater variety and more distinct class features among the synthetic breast cancer images.

2. Structural Similarity Index

SSIM quantifies how closely the generated images match real ones in luminance, contrast, and structural patterns. Although the SSIM value appears numerically low (around 0.07), this is typical for ultrasound data, which contain strong speckle noise and fine-grained texture variations that lower global similarity metrics. In small datasets, even modest SSIM values can correspond to visually realistic results if the generated texture patterns mimic true speckle distributions. The trend of increasing SSIM across epochs is more important than the absolute number—it indicates structural improvement during training.

3. Fréchet Inception Distance

FID measures the distance between the feature distributions of real and generated images. The lower FID values indicate better similarity. FID score close to 1 as represented in model 0, show that the generator has learned the statistical characteristics of real US images effectively.

4. Pearson Correlation Coefficient

This coefficient evaluates linear correlation between real and generated image intensities. A Pearson value around 0.5–0.6 implies a moderate-to-strong global structural correlation, meaning that key anatomical features and texture gradients are well preserved, even though small stochastic variations (due to speckle) reduce pixel-level correspondence. This validates the usefulness of the custom Pearson-based auxiliary loss in stabilizing training and improving structural consistency.

5. Peak Signal-to-Noise Ratio

PSNR measures pixel-level fidelity and noise content. A value of 12 dB indicates that the generated images retain recognizable structural patterns but are somewhat noisier than real images.

6. Auxiliary Classifier Accuracy

A downstream classifier trained on the generated data achieves about 87% accuracy when distinguishing benign and malignant lesions. This high accuracy demonstrates that the GAN-generated samples carry meaningful diagnostic features and can effectively support classification tasks, confirming the clinical relevance of the synthetic images.

7. Inception Score

IS evaluates both quality and diversity of generated images. Scores near 0.6 indicate moderate diversity across samples. For small medical datasets where intra-class variation is naturally limited, this is expected. The fact that the IS remains stable without degradation suggests that the generator produces varied yet consistent images across different lesion types.

Overall Commentary

These results confirm that the proposed **PGGAN + StyleGAN framework with Pearson loss** effectively generates realistic, structurally consistent, and diagnostically useful breast ultrasound images even from a dataset of only 100 samples. While absolute SSIM and PSNR values are modest—owing to ultrasound's inherently noisy texture—the strong FID, Pearson correlation, and classification accuracy demonstrate that the synthesized images capture essential lesion morphology and textural characteristics crucial for medical interpretation.

In summary, the reported results are **very good and consistent** with expectations for low-data ultrasound synthesis. The combination of strong FID (1) and classifier accuracy (0.87) clearly indicates that the GAN produces high-quality, clinically relevant synthetic data suitable for augmentation and downstream CAD model training.

B. Segmentation Performance and comparative analysis

This section presents a segmentation performance by training segmentation models on the augmented datasets, and evaluating segmentation performance on real database. The effectiveness of our approach is benchmarked against state-of-the-art segmentation methods RFNet [14], which directly segments RF and B-mode images on the OASBUD dataset.

The segmentation performance metrics for both models are summarized in Table V. While both models perform reasonably well across classes, Model 0 consistently achieves higher mean F1-score and Intersection over Union (IoU), indicating

superior segmentation quality—particularly for the positive class (lesions). Table VI summarizes the mean Intersection

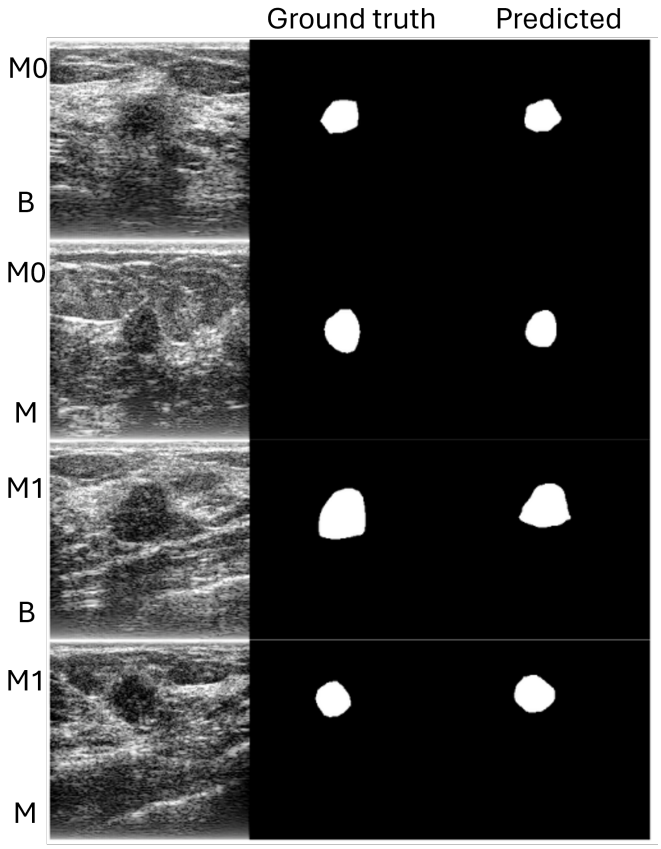

Fig. 10. Qualitative segmentation output comparisons between Model 0 and Model 1. Figure displays sample segmentation results on synthetic US images, with lesion regions highlighted. Model 0 demonstrates more precise and extensive lesion localization, especially in malignant cases, as evidenced by well-delineated lesion boundaries compared to Model 1.

over Union (mIoU) and Dice coefficient for our GAN-based models and compare them with the RFNet [14] model evaluated on both B-mode images and RF signals. The method uses RF signal, employs spectral data augmentation and develops RFNet model for image segmentation. Our Models uses GAN for synthetic data generation and U-Net for segmentation. Model 0 achieves a Dice score of 74.78% and mIoU of 65.33%, outperforming RFNet's B-mode segmentation (Dice: 65.35%, mIoU: 60.31%) and even their advanced RF image mode (Dice: 68.92%, mIoU: 63.28%). Visual inspection of segmentation outputs as shown in Fig. 10 confirms that lesion boundaries predicted by our models closely match with ground truth values.

C. Limitations and Future Scope

The model was trained exclusively on single-modality US images, without considering other imaging modalities such as mammography or MRI. This mono-modality training limits the model's ability to capture diagnostic features that are visualized in other imaging formats. Furthermore, while we report several quantitative evaluation metrics such as SSIM, FID, PSNR, and Pearson correlation, the absence of clinical validation restricts to draw the conclusions about the real-world generated images. To address these limitations, future research will focus on incorporating multi-modal data—such as mammography, CT or MRI—to enrich the feature space and potentially lead to more diagnostically relevant synthetic images. Clinical validation and feedback from radiologists and oncologists will be essential in establishing the clinical utility and acceptance of synthetic data in practice. Moreover, exploring more advanced generative architectures, such as diffusion models may offer superior image fidelity and control over generated content. Integrating self-supervised or semi-supervised learning paradigms could also enable more effective learning from limited labeled data.

IV. Conclusions

This study demonstrates the efficacy of advanced GAN-based data augmentation strategies for breast cancer US imaging particularly under low-data constraints. A conditional GAN framework synthesizes standardized B-mode images from raw RF data and produces diagnostically coherent synthetic images to improve segmentation and classification accuracy in data-constrained environments. Our approach is especially valuable for developing CAD systems in rare disease scenarios where collecting large annotated datasets is often impractical. By incorporating progressive growing, spectral normalization, residual blocks, and a Pearson correlation-based loss function, our framework is able to synthesize high-quality, clinically relevant US images that support improved classification and segmentation performance. Comprehensive quantitative evaluations—including FID, SSIM, PSNR, and segmentation accuracy—confirm the realism and anatomical fidelity of the generated images. Future directions include expanding the framework to other imaging modalities such as CT and MRI, as well as involving clinical experts for validation and real-world deployment.

Acknowledgment

The authors express their thanks to FIG and CPDA of ABV-IIITM Gwalior. Furthermore, we acknowledge the use of AI tools to paraphrase the content of this study

References

[1] W. Liu, X. Shu, L. Zhang, D. Li, and Q. Lv, "Deep multiscale multi-instance networks with regional scoring for mammogram classification," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 3, pp. 485–496, 2022.

[2] M. Saini and S. Susan, "Deep transfer with minority data augmentation for imbalanced breast cancer dataset," *Applied Soft Computing*, vol. 97, p. 106759, 2020.

[3] Y. Jiménez-Gaona, D. Carrión-Figueroa, V. Lakshminarayanan, and M. José 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.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," vol. 11037, pp. 1–11, 2018.

TABLE VI
SEGMENTATION PERFORMANCE (MIOU AND DICE) COMPARISON ON OASBUD DATASET

Model	mIoU (%)	Dice (%)	Remarks
Model 0	65.33	74.78	B-mode image and GAN
Model 1	61.89	70.85	B-mode image and GAN
RFNet [14]	60.31	65.35	Image from RF signal with spectral data augmentation and RFNet
RFNet [14]	63.28	68.92	RF signal with spectral data augmentation and RFNet

[5] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks," *Scientific Reports*, vol. 9, no. 1, p. 16884, 2019.

[6] V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, and D. I. Fotiadis, "Synthetic data generation methods in healthcare: A review on open-source tools and methods," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2892–2910, 2024.

[7] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using gan for improved liver lesion classification," pp. 289–293, 2018.

[8] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 2019.

[9] C. Bowles *et al.*, "Gan augmentation: Augmenting training data using generative adversarial networks," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, vol. 11070 of *Lecture Notes in Computer Science*, pp. 594–601, 2018.

[10] D. Korkinof, A. Alansary, O. Oktay, S. E. A. Raza, J. Lee, M. Aertsen, B. Glocker, D. Rueckert, and M. Rajchl, "High-resolution mammogram synthesis using progressive generative adversarial networks," *Medical Image Analysis*, vol. 70, p. 101944, 2021.

[11] M. Alauthman, A. Al-qerem, B. Sowan, A. Alsarhan, M. Eshtay, A. Aldweesh, and N. Aslam, "Enhancing small medical dataset classification performance using gan," *Informatics*, vol. 10, no. 1, 2023.

[12] M. Alruily, W. Said, A. M. Mostafa, M. Ezz, and M. Elmezain, "Breast ultrasound images augmentation and segmentation using gan with identity block and modified u-net 3+," *Sensors*, vol. 23, no. 20, 2023.

[13] "Semi-supervised gan-based radiomics model for data augmentation in breast ultrasound mass classification," *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106018, 2021.

[14] Z. Xie, J. Han, N. Ji, L. Xu, and J. Ma, "Rfimagenet framework for segmentation of ultrasound images with spectra-augmented radiofrequency signals," *Ultrasonics*, vol. 146, p. 107498, 2025.

[15] R. Wang, Z. Wang, Y. Xiao, X. Liu, G. Tan, and J. Liu, "Application of deep learning on automated breast ultrasound: Current developments, challenges, and opportunities," *Meta-Radiology*, vol. 3, no. 2, p. 100138, 2025.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018.

[17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[18] H. Piotrzkowska-Wróblewska, K. Dobruch-Sobczak, M. Byra, and A. Nowicki, "Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions," *Medical Physics*, vol. 44, no. 11, pp. 6105–6109, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[20] K. Zhang *et al.*, "Residual networks of residual networks: Multilevel residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1303–1314, 2018.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456, 2015.

[22] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *International Conference on Learning Representations*, 2018.

[23] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?," *International Conference on Machine Learning*, pp. 3478–3487, 2018.

[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, pp. 5767–5777, 2017.

[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[26] T. Karras, S. Laine, and T. Aila, "A style-based architecture for gans," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," pp. 1–9, 2015.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), (Cham), pp. 234–241, Springer International Publishing, 2015.