Reliable or Risky? Assessing Diffusion Models for Biomedical Data Generation

Abdalrahman Alblwi[*] **Qile Wang**[†] **Norbert Zolek**[‡]
Matthew Louis Mauriello[†] **Kenneth Barner**[¶]

[*]CEDAR – Knight Cancer Institute, Oregon Health & Science University
[†]Computer and Information Sciences Department, University of Delaware
[‡]Institute of Fundamental Technological Research, Polish Academy of Sciences
[¶]Department of Electrical and Computer Engineering, University of Delaware

Corresponding author: barner@udel.edu

Abstract

Biomedical image datasets are often scarce, expensive to annotate, and vary in quality due to differences in imaging hardware and techniques. Generative models, particularly diffusion models, have recently demonstrated strong potential to synthesize realistic medical images, offering a promising strategy for data augmentation. Yet, their application in clinical contexts requires careful validation, as trust, interpretability, and reliability are essential when medical decisions are at stake. This work introduces a human-in-the-loop framework for assessing the reliability and risks of diffusion models in generating breast ultrasound cancer images. Using a Denoising Diffusion Probabilistic Model (D-DDPM), we jointly generate ultrasound images and corresponding tumor masks from two benchmark datasets (BUS-BRA and UDIAT). The evaluation pipeline integrates quantitative image quality metrics (FID, IS, KID), radiologist interpretation, inter-rater agreement (Cohen's/Fleiss' Kappa, Krippendorff's Alpha), and alignment with large language model (LLM) outputs. Results show that while D-DDPM can produce images that are visually similar to real data and sometimes yield higher agreement among experts than original images, inter-rater reliability remains weak, particularly for malignant tumors. Radiologists consistently outperform LLMs in classification, though majority voting across experts improves diagnostic accuracy. These findings highlight both the promise and risks of diffusion models in medical imaging, including that synthetic ultrasound data can supplement limited datasets; however, robust expert validation remains indispensable to ensure clinical trustworthiness and safe integration.

1 Introduction

Cancer is a serious global health concern, with breast cancer being one of the most widespread and fatal types. In 2022, approximately 2.3 million women were diagnosed with breast cancer worldwide, and around 670,000 women lost their lives to the disease [1, 2]. Ultrasound, MRI, and Mammography are recommended biomedical imaging modalities for screening to reduce the risk of death from breast cancer. Ultrasound imaging, in particular, is preferred due to it being widely available, free of ionizing radiation, and low cost. However, making clinical decisions based on this image data is time-consuming and not without risk. Automatic tumor classification and segmentation are viewed as essential tasks for improving medical image analysis times and accuracy. In this process, identifying

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

the type and size of a tumor is crucial, as it offers vital information that assists clinicians in making precise assessments and informed treatment decisions.

Nevertheless, supervised learning-based models rely heavily on large datasets to effectively learn patterns and provide accurate predictions, particularly for segmentation and classification tasks. In the last decade, various data augmentation techniques such as Mixup [3], CutMix [4], AutoAugment [5], and conventional methods [6] (e.g., rotation, flipping, contrast adjustment, noise injection, etc.) have been proposed to address data scarcity and enhance the performance of supervised models. Meanwhile, Generative Adversarial Networks (GANs) present a promising alternative for augmenting ultrasound image datasets by generating realistic synthetic images, enhancing dataset diversity, and improving model performance. GANs leverage adversarial training between a generator and a discriminator networks to produce high-quality synthetic data that mimics real data distributions. Several studies have leveraged GANs and their variants to synthesize realistic ultrasound images, significantly increasing dataset sizes [7, 8, 9, 10]. This highlights the potential of generative models in expanding the size of ultrasound data. However, several limitations reduce the effectiveness of GAN-based models, including their inability to capture the full diversity of the data, as noted in [11].

Diffusion models have recently attracted attention for their ability to generate high-quality and diverse images, often exceeding GANs in both image quality and diversity [11]. For instance, Denoising Diffusion Probabilistic Models (DDPMs) are generative models that use a Markov chain of diffusion steps to progressively add random noise to data. The model operates by iteratively denoising a corrupted input to reconstruct the original data. With the advent of diffusion models, these techniques can synthesize new biomedical images that are often indistinguishable from real data to human experts [12, 13]. Alongside advances in image generation, large language models such as ChatGPT [14] and Gemini [15] are increasingly employed to extract structured insights from medical data [16]. For instance, Gemini, developed by Google DeepMind, supports multimodal input—text, images, video, and audio—and is trained on a wide range of content, including medical texts. Its ability to generate clinically relevant responses makes it a promising tool for assisting with medical image interpretation and information extraction.

Despite the impressive performance of generative models and large language models in medical data generation, such as synthesizing medical images and clinical reports, the safety-critical nature of clinical applications demands rigorous validation [17, 18]. Excluding medical experts from the evaluation loop risks safety concerns, ethical pitfalls, and errors in downstream analyses. To address this, we present a comprehensive expert-driven framework for assessing the reliability and risks of generative models in ultrasound cancer imaging, motivated by recent studies on the trustworthiness of medical data generation [19, 20, 21]. Our contributions are threefold:

- We measure the reliability, risks, and potential applications of generative models in medical imaging, including the use of diffusion models for annotation and large language models (LLMs) for diagnostic interpretation.
- We incorporate breast cancer radiologists in evaluating synthetic image quality, annotation fidelity, LLM interpretation reliability, and identification of outlier cases.
- We provide a systematic analysis that integrates quantitative metrics with expert feedback, highlighting both the promise and limitations of diffusion-based biomedical data generation.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 introduces our methodology, datasets, and experiments setup; Section 4 details the evaluation metrics; Section 5 presents the experimental results; and Sections 6 and 7, respectively, discusses key insights and implications for biomedical data generation and the conclusion of our work.

2 Related Work

To fully leverage the capabilities of the generative model for data augmentation, we combine the image x_i with its corresponding mask x_m and apply both the forward and reverse diffusion processes, as illustrated in Figure 1, [22, 23, 24]. This approach generates paired images and masks, significantly expanding the size and variety of biomedical datasets. The dimensions of the combined image I and mask M are defined as

$$X_{Comb} = \text{concat}(I, M) \in \mathbb{R}^{(C+1)\times H\times W}. \tag{1}$$

Assuming the images and masks are in grayscale, set the number of channels $C = 1$. The forward process for the merged image and its corresponding mask is then defined as follows:

$$q(x_{Comb,t} \mid x_{Comb,t-1}) := \mathcal{N}\left(x_{Comb,t}; \sqrt{1-\beta_t}x_{t-1},\ \beta_t\mathbf{I}\right), \tag{2}$$

where β_t is the noise schedule parameter at timestep t controlling the amount of noise added during diffusion. To generate images along with their corresponding masks during the sampling process, starting from Gaussian noise $x_{Comb,T} \sim N(0,1)$, the final equation can be reformulated as

$$x_{Comb,t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_{Comb,t} - \frac{1-\alpha_t}{\sqrt{1-\hat{\alpha}_t}}\epsilon_\theta(x_{Comb,t}, t)\right) + \sigma_t z, \tag{3}$$

In this setup, z denotes Gaussian noise sampled from $\mathcal{N}(0,1)$, σ_t is determined by the noise scheduler based on the DDPM approach [22, 24, 25], and $\epsilon_\theta(x_{Comb,t}, t)$ is the network-predicted noise at timestep t. Here, $\alpha_t = 1 - \beta_t$, and the cumulative product term $\hat{\alpha}_t$ is defined as $\hat{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, where β_t represents the predefined noise variance at timestep t.

Throughout the D-DDPM forward process, noise is progressively added to both the image and mask over T timesteps. During the reverse process, a modified U-Net based on residual U-shaped blocks [23] jointly denoises the concatenated image and mask, ensuring alignment throughout the generation process. Two samples are generated for inference: one for the image and its corresponding mask, using the trained D-DDPM model applied to both channels (image and mask). During inference, a noise sample $\mathbf{N} \in \mathbb{R}^{1\times2\times H\times W}$ is drawn from a standard normal distribution:

$$\mathbf{N}^{1,2,H,W} \sim \mathcal{N}(0,1).$$

Here, 2 denotes the channels (image and mask), while H and W represent the height and width of the spatial dimensions, respectively. The modified U-Net gradually denoises the sample $\mathbf{N}$ in each step to generate the final image and mask, ensuring their alignment remains consistent throughout the process.

3 Methodology

We designed a pipeline to incorporate expert radiologists' evaluations of synthetic images to gain an understanding of the feasibility of diffusion models in generating medical ultrasound images that contain tumors. The first stage of our pipeline involves generating synthetic ultrasound images of benign and malignant tumors independently. This separation ensures reliable, class-dependent image generation without mixing the distributions of the two tumor types, as shown in Figure 1, left. In this stage, we use a D-DDPM model to generate both the ultrasound images and the corresponding tumor masks for each class. In the second stage, as shown in Figure 1, on the right, the generated images and masks are used as inputs—along with descriptive prompts—for a zero-shot classification task using the Gemini large language model.

At the same time, we created a questionnaire for expert radiologists to review the quality of the generated images and masks. The questionnaire also asks them to rate their confidence in the synthetic data's anatomical accuracy and clinical usefulness. In addition, we included Gemini's responses in the questionnaire, allowing radiologists to assess whether the interpretation of the LLM matched their own image understanding. The medical experts involved in this study are three radiologists (R1, R2, and R3) affiliated with the Polish Academy of Sciences, each with extensive experience in ultrasound imaging and breast cancer analysis [26].

We designed a detailed questionnaire for the radiologists to assess the quality and clinical relevance of the generated images and masks. The questionnaire included specific questions such as: *"What is the class of this image?"*. The evaluation focused on classifying images into one of three categories: benign, malignant, or normal. Additionally, we measured the confidence level of each radiologist in relation to the diffusion-generated masks. To do this, we asked: *"How confident are you in the annotation quality of this image?"* Responses were recorded on a 0–100 confidence scale, with intervals of 20. While there are no healthy images in the dataset, we included the "normal" category to provide flexibility for radiologists in their assessment, particularly in cases where a tumor is difficult to detect through visual inspection. This approach ensures that medical experts can better assess the quality of the images and masks, even when the tumor is not clearly visible.

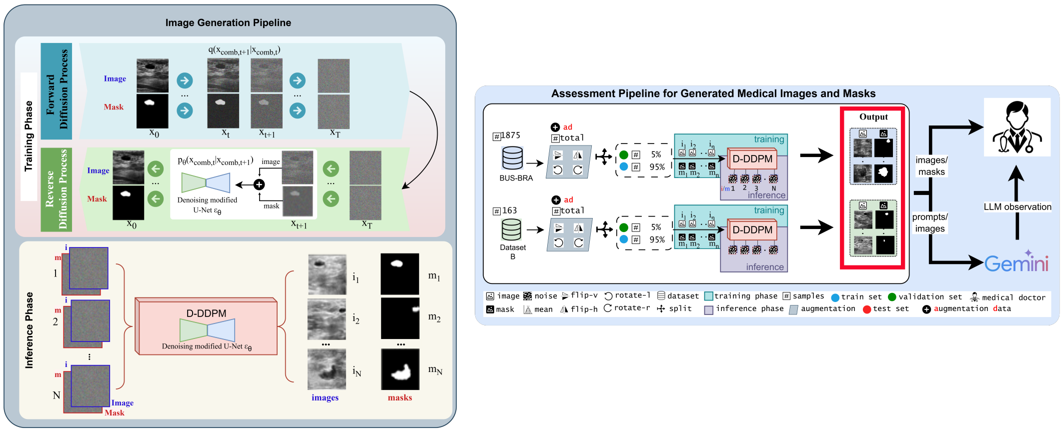


Figure 1: The proposed approach consists of two main stages. Stage 1 (left) generates ultrasound images along with their corresponding masks. For each image-mask pair, we apply a diffusion process to add noise and train a model to reverse this process. During inference, the model generates a two-channel output that includes an ultrasound image and its annotation mask. Stage 2 (right) takes batches of these synthetic images and masks, combines them with a specified prompt, and uses Gemini—a large language model—to generate textual summaries. Medical professionals then review the outputs, assessing the quality of the images and masks, their alignment, the realism of the generated images, and the accuracy of the LLM-generated summaries and interpretations. See Appendix 7 for further details.

We also incorporated a question regarding the large language model (LLM)'s observations to further explore the alignment between clinical perspectives and the generated content. Experts were asked to assess whether the LLM-generated description or classification aligned with their clinical interpretation. We included this step to evaluate the potential of zero-shot classification by LLMs in reflecting clinical reasoning within medical imaging.

Inspired by the work of Chen *et al.* [27], we adopt a set of clinical prompts to guide Gemini's evaluation of synthetic ultrasound images. These prompts cover key lesion characteristics commonly reported in radiology, such as tumor depth, anatomical region, lesion type, presence of calcification, and suspicion of malignancy. Although we do not include the full set of keys in the main paper, a complete list of prompts and LLM outputs is provided in the supplementary materials. Given the limitations in having medical experts manually evaluate all generated images, we incorporate Gemini's responses as an alternative source of clinical feedback. This helps us explore its potential for zero-shot classification and assess its reliability in supporting medical decision-making. As illustrated in Figure 1, expert radiologists reviewed Gemini's outputs for selected cases to evaluate clinical plausibility and alignment with expected findings.

3.1 Dataset

In this work, we utilize one private and one public dataset. **Dataset B** (Yap et al. [28]). This dataset consists of 163 ultrasound images, comprising 110 benign and 53 malignant cases. The images were collected at the UDIAT Diagnostic Centre, Parc Taulí Corporation, Sabadell, Spain. Tumors were imaged using the Siemens ACUSON Sequoia C512 system equipped with a 17L5 HD linear array transducer. The average image resolution is 760 × 570 pixels. **BUS-BRA** (Gómez-Flores et al. [29]): This dataset contains 1,875 ultrasound images from 1,064 female patients, including 1,268 benign and 607 malignant cases. The images were acquired using four ultrasound devices: GE Logiq 5, GE Logiq 7, Toshiba Aplio 300, and GE U-Systems. The data was sourced from the National Institute of Cancer in Rio de Janeiro, Brazil.

3.2 Experiments

The training data is split into a 95:5 ratio for training and validation. Due to each dataset's distinct statistical properties and distribution patterns, we train a separate Denoising Diffusion Probabilistic

Model (D-DDPM) for each one. This enables the model to learn dataset-specific image–mask generation characteristics better. For example, the BUS-BRA dataset exhibits high variability in tumor sizes, while Dataset B shows more consistent tumor proportions, averaging about 19% of the total image area. Training independently allows the model to capture these dataset-specific traits more effectively.

The model is optimized using mean squared error (MSE) to minimize the diffusion reconstruction loss, $\mathcal{L}_{\text{Diff}}$, encouraging the model to accurately predict the added noise during the denoising process. The training assumes that the covariance matrices of the forward and reverse diffusion steps—$q(x_{t-1} \mid x_t, x_0)$ and $p(x_{t-1} \mid x_t)$—are equivalent, and that the mean is predicted by the noise estimator $\hat{\epsilon}_\theta$.

The loss function used for training is given by:

$$\mathcal{L}_{\text{Diff}}(\theta) = \left\| \epsilon - \hat{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_t} x_{Comb} + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right\|^2 . \tag{4}$$

To enhance the robustness of the model, we apply basic data augmentation techniques, including horizontal and vertical flipping, as well as rotations by $-90°$, $+90°$, $-180°$, and $+180°$. All images are resized to 256×256 pixels and converted to grayscale for consistency and efficient computation.

4 Evaluation Metrics

4.1 Tumor Image Generation Assessment

To evaluate both the diversity and quality of the generated images, three commonly used metrics are applied, as outlined by [30] and [22]. These include the Fréchet Inception Distance (FID), Inception Score (IS), and Kernel Inception Distance (KID). FID quantifies the similarity between the feature distributions of real and generated images by utilizing the Inception v3 network, which has been pre-trained on the ImageNet dataset [30, 31]. Given two sets of images, one representing the reference (real) images X_{ref} and the other consisting of the generated images Y_{gen}, the FID is calculated as:

$$\begin{aligned} \text{FID}(X_{\text{ref}}, Y_{\text{gen}}) = & ||\mu_{\text{ref}} - \mu_{\text{gen}}||^2 \\ & + \text{Tr} \left(\Sigma_{\text{ref}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{ref}} \Sigma_{\text{gen}})^{1/2} \right) . \end{aligned} \tag{5}$$

where μ_{ref} and μ_{gen} are the mean feature vectors, and Σ_{ref} and Σ_{gen} are the covariance matrices of the reference and generated image distributions, respectively. A decrease in FID value reflects an increased similarity between the generated images and the real images. The IS quantifies both the diversity and quality of generated images by analyzing the uncertainty in class predictions from an Inception network. The IS is expressed as:

$$\text{IS}(Y_{\text{gen}}) = \exp \left(\mathbb{E}_{y_{gen} \sim p_{gen}} \left[D_{\text{KL}}(p(y|y_{gen}) \parallel p(y)) \right] \right) . \tag{6}$$

Here, y_{gen} denotes a generated image sample drawn from the model distribution p_{gen}. Where $p(y)$ is the marginal class distribution and $p(y|y_{gen})$ is the conditional class distribution for an image y_{gen}. Higher IS values correspond to improved quality and diversity in the generated images.

The KID measures the similarity between the feature distributions of real and generated images by utilizing the squared Maximum Mean Discrepancy (MMD) with a polynomial kernel. Calculated as:

$$\begin{aligned} \text{KID}(X_{\text{ref}}, Y_{\text{gen}}) = & \mathbb{E}_{x,x' \sim X_{\text{ref}}} k(x, x') + \\ & \mathbb{E}_{y,y' \sim Y_{\text{gen}}} k(y, y') - \\ & 2\mathbb{E}_{x \sim X_{\text{ref}}, y \sim Y_{\text{gen}}} k(x, y), \end{aligned} \tag{7}$$

where $k(x, y)$ represents a polynomial kernel function. Lower KID values reflect a higher degree of similarity between the real and generated distributions. These metrics, when considered collectively, provide a comprehensive assessment of the realism and diversity of the generated images, which is crucial for evaluating the performance of generative models.

4.2 Observer Interpretation and Classification Assessment

To evaluate the consistency of expert interpretations and assess classification performance on diffusion-generated outputs, we employ a combination of inter-rater agreement and evaluation metrics. For

Table 1: Quantitative comparison of feature vector distances between original and synthetic ultrasound images using FID, KID, and IS metrics.

Dataset	↑ IS	↓ KID	↓ FID
Dataset B	0.0019 ± 0.0193	0.1343 ± 0.0136	216.03
BUS-BRA	0.0005 ± 0.0117	0.0338 ± 0.0070	116.63

Table 2: Inter-Rater Reliability (IRR) in classifying ultrasound images (Normal, Benign, and Malignant). Abbreviations: APP = Average Pairwise Percent, FK = Fleiss' Kappa, CK = Cohen's Kappa (Pairwise), KA = Krippendorff's Alpha.

Category	**APP**	**FK**	**CK**	**Level**	**KA**	**Level**
All images (n=41)	67.48%	0.418	0.421	Weak	0.423	Poor
Original (n=10)	60%	0.373	0.391	Minimal	0.394	Poor
Generated (n=31)	69.89%	0.422	0.423	Weak	0.428	Poor
Benign (n=20)	66.67%	0.253	0.271	Minimal	0.265	Poor
Malignant (n=21)	68.25%	0.082	0.080	None	0.096	No

observer interpretation, metrics such as **Average Pairwise Agreement**, **Cohen's Kappa**, **Fleiss' Kappa**, and **Krippendorff's Alpha** are used to quantify the degree of consistency among medical experts when analyzing synthetic images and corresponding masks. Average Pairwise Agreement measures the proportion of agreement between all rater pairs; Cohen's Kappa accounts for agreement expected by chance between two raters; Fleiss' Kappa extends this to multiple raters; and Krippendorff's Alpha further generalizes to handle missing data and different measurement scales [32]. Therefore, these metrics account for chance agreement and variations in rater number and data completeness, offering a robust measure of interpretative reliability.

For the classification task, both medical experts and the LLM assign labels and clinical classes to the images and segmentation masks generated by the diffusion model. Performance is assessed using standard metrics: **Precision (P)**, **Recall (R)**, and the **F1 Score**, which together capture the accuracy and balance of correct and incorrect classifications. This framework enables a direct comparison between expert and LLM performance in interpreting diffusion-generated medical content.

5 Results

Table 1 presents the similarity scores between real and synthetic datasets using three widely adopted image quality metrics: IS, FID, and KID. These metrics collectively assess the realism and diversity of generated images and are standard benchmarks for evaluating generative models such as GANs and diffusion models. Our analysis focuses on quantifying how closely the synthetic ultrasound images align with their real counterparts, rather than on baseline model comparisons. For the BUS-BRA dataset, the model achieves a relatively strong FID of 116.63, indicating that the generated samples are visually similar to real images. Conversely, the lower IS observed for Dataset B reflects reduced diversity, likely influenced by the smaller tumor regions frequently present in this dataset. Still, the low KID (0.1343) and FID (216.03) values demonstrate that the model effectively captures key structural patterns such as tumor size and shape across datasets.

Table 2 presents the inter-rater agreement results among three radiologists tasked with classifying images into three categories: **Normal**, **Benign**, and **Malignant**. The agreement is reported across the following evaluation scenarios: The image dataset was analyzed across several subsets for detailed evaluation. The **All images** subset includes real and synthetic images. The **Original** subset contains only real ultrasound images, while the **Generated** subset is composed solely of synthetic images, covering both benign and malignant cases. The **Benign** subset combines real and synthetic benign cases, and the **Malignant** subset similarly combines real and synthetic malignant cases.

For both the **All images** and the **Generated** categories, the results from Fleiss' Kappa and the averaged Cohen's Kappa indicate weak inter-rater reliability, highlighting significant variation in the agreement among the radiologists. Radiologists showed slightly higher agreement for the generated images compared to the original images. Fleiss' Kappa scores for the generated images were 0.422, compared to 0.373 for the original images. Similarly, Cohen's Kappa for the generated images was 0.423, while for the original images, it was 0.391. These results indicate a marginal improvement in

Table 3: Classification results based on Recall (R) and F1 scores, covering the LLM model, individual radiologists (R1, R2, and R3), and majority voting. Green bold values highlight the best performance.

	n = 41 (All images)		**n = 10 (Original)**		**n = 31 (Generated)**	
Evaluators	**R**	**F1**	**R**	**F1**	**R**	**F1**
LLM	0.41	0.42	0.40	0.42	0.42	0.40
R1	0.78	0.81	0.60	0.68	**0.84**	**0.85**
R2	0.71	0.74	0.70	0.79	0.71	0.72
R3	0.75	0.77	0.70	0.73	0.77	0.77
Majority Vote	**0.83**	**0.84**	**0.80**	**0.84**	**0.84**	0.84

Table 4: Evaluation metrics for clinical descriptions: APP (Average Pairwise Percent), FK (Fleiss' Kappa), CK (Cohen's Kappa, pairwise), and KA (Krippendorff's Alpha). Radiologists evaluated the clinical descriptions generated by the LLM model, and their average confidence scores were also recorded for each individual.

	R1	**R2**	**R3**	**APP**	**FK**	**CK**	**KA**
All images (n=41)	2.675	2.725	2.600	53.33%	0.174	0.18	0.181
Original (n=10)	2.60	2.70	2.80	60%	0.341	0.374	0.363
Generated (n=31)	2.70	2.73	2.53	51.11%	0.104	0.111	0.111

inter-rater agreement for the generated images. The generated and original benign images showed Fleiss' Kappa scores of 0.253 and Cohen's Kappa scores of 0.271, indicating a minimal level of agreement.

In contrast, the malignant class showed almost no agreement, as reflected in significantly lower Kappa scores. This highlights a clearer consistency in classifying benign images compared to malignant ones. This also suggests that the generated and original malignant cancer images in ultrasound are quite diverse, both in nature and generative models, likely due to the irregular shapes and varying characteristics of malignancy; see Figure 6. Even with the options provided to the radiologists (normal, benign, and malignant), there are noticeable differences in how the radiologists interpreted the images, which could reflect the complexity and detail of the generated images. The overall results indicate that the reliability reaches a maximum of 35%, compared to the worst-case scenario where data reliability drops to less than 15%.

Table 3 presents radiologists' performance, LLM models, and majority voting on **All images**. The results show that radiologists consistently outperform the LLM models across all categories. Specifically, the best radiologist (R2) performance achieved a recall of 0.78 and an F1 score of 0.81, indicating that human expertise remains superior to machine learning models in this context. Interestingly, the majority voting approach significantly improved performance, surpassing the best radiologist's results by 0.83 in recall and 0.84 in F1 score. This corresponds to an improvement of approximately 6.4% and 3.7%, respectively, when multiple evaluators contributed to the medical image interpretation, resulting in higher overall classification accuracy. Majority voting achieved higher performance on generated ultrasound images than on the original data. For synthetic images, recall and F1 scores both reached 0.84, compared to 0.80 and 0.84, respectively, for real images. This result indicates that the diffusion model is capable of producing images of sufficient quality and consistency to be classified as accurately as, and in some cases more reliably than, original ultrasound scans.

Further improvement could be achieved with access to larger and more balanced datasets across categories, which would provide a more comprehensive basis for comparison and potentially enhance the robustness of the findings. In addition, the LLM performance was the lowest among all evaluators, which indicates that LLM models face challenges in interpreting complex medical images. This suggests that relying solely on LLMs for such tasks may not be effective, particularly when dealing with challenging datasets like ultrasound. The results highlight the importance of fine-tuning these models on domain-specific datasets to improve their classification accuracy.

Evaluating the descriptions generated by Gemini with radiologists' feedback is crucial to ensure clinical accuracy, identify potential biases or errors, improve the model's reliability, and confirm that the generated insights align with real-world diagnostic standards and medical expertise. Therefore, we evaluated the descriptions as shown in Figure 3. Table 4 presents the quality measurements

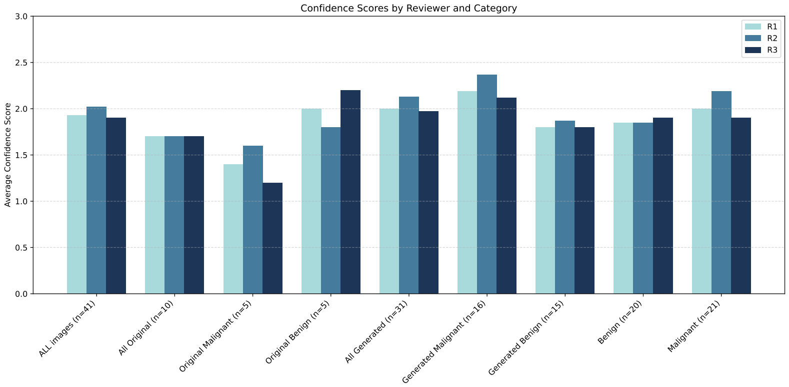


Figure 2: The average confidence scores range from 1 (very confident) to 5 (least confident). These scores represent the doctors' overall confidence in the annotations produced by the diffusion model for the corresponding synthetic ultrasound images.

Table 5: Evaluation of radiologists to identify unreal images with Precision (P), Recall (R), and F1 Score.

Radiologists	P	R	F1 Score
R1	0.267	1.0	0.421
R2	0.308	1.0	0.471
R3	0.273	0.75	0.400

of the descriptions provided by the LLM and the agreement between radiologists on the given descriptions. Interestingly, the descriptions of the generated and original (real) images show no significant difference, indicating the potential of leveraging LLMs to generate reliable clinical insights from medical images. Figure 3 shows an example of a high-agreement description generated by Gemini that radiologists found clinically acceptable. The radiologists were asked, *"Does the description accurately reflect the content and nature of this image?"* and all agreed on choosing "Mostly." However, they noted some minor discrepancies in the description. This highlights the value of involving large models trained on image-to-text problems, not only to support clinical decision-making but also as a tool to assess the quality of generated images. The results suggest that D-DDPM can generate acceptable samples for clinical interpretation, reinforcing the potential of combining synthetic data generation with advanced language models for enhanced diagnostic support.

For mask annotation using D-DDPM, each generated mask highlights the suspicious tumor within its corresponding ultrasound image. Because the model jointly produces both images and masks, assessing annotation quality with expert input is essential. To this end, radiologists were asked to rate their confidence in each image–mask pair on a scale from 1 (very confident) to 5 (not confident), with intermediate scores indicating decreasing levels of trust. Figure 2 summarizes these confidence ratings. Overall, scores ranged from high to moderate confidence, suggesting that D-DDPM produces annotations of generally acceptable quality. As expected, the original data yielded the highest confidence, reflecting their clearer tumor boundaries and reduced noise. In contrast, some synthetic images, particularly those depicting multiple or irregular tumors, received lower confidence scores, highlighting the challenges of maintaining annotation fidelity in more complex cases.

With the rapid advancement of generative models, distinguishing between real and synthetic images has become increasingly important, particularly in medical imaging, where patient privacy and diagnostic accuracy are of paramount importance. To evaluate this challenge, we designed a grid-based test (Figure 4) in which each row contains a single real image alongside multiple synthetic counterparts. The real image was randomly positioned within the row to prevent bias, and grids were constructed using samples from both Dataset B and BUS-BRA. This setup provides a controlled framework for assessing experts' ability to reliably differentiate authentic ultrasound scans from diffusion-generated images. The outcomes of this experiment, along with broader implications for clinical reliability and data integrity, are examined in the following Discussion section.

6 Discussion

This study employed the D-DDPM framework to generate synthetic ultrasound images conditioned on paired segmentation masks, thereby mitigating data scarcity in clinical imaging scenarios. To assess the clinical interpretability of these generated images, we involved expert radiologists who evaluated both real and synthetic data across various diagnostic categories. The resulting findings offer key insights into the diagnostic reliability of generative models in ultrasound imaging. Overall agreement among radiologists was weak, as reflected in the low Kappa scores of Fleiss and Cohen in all types of cases. This indicates considerable variability in expert interpretation, underscoring the inherent subjectivity of ultrasound-based diagnosis. Surprisingly, synthetic images generated by D-DDPM exhibited slightly higher agreement levels than real images, suggesting that well-trained generative models can approximate the quality and clinical utility of authentic data. This finding supports the use of synthetic data as a supplement to real datasets in training and evaluation pipelines.

When stratifying the results by diagnostic category, benign cases showed limited agreement among radiologists, while malignant cases demonstrated virtually no consensus. This stark contrast highlights the greater diagnostic ambiguity and complexity of malignant tumors. Factors such as irregular shapes, high noise levels, and textural complexity likely contribute to the difficulty in reaching consistent interpretations, regardless of whether the images are real or generated. The levels of agreement varied significantly between individual cases. In some instances, inter-rater agreement reached up to 35%, but in others, it fell below 15%. This variability highlights the challenges of subjective interpretation in ultrasound imaging, especially in diagnostically ambiguous scenarios. Such inconsistency raises concerns about clinical reliability and poses challenges for evaluating and benchmarking machine learning models trained on this data. These findings underscore the crucial need for enhanced standardization in training datasets and annotation protocols. Additionally, they highlight the value of integrating expert consensus mechanisms, such as majority voting or ensemble decision-making, into clinical workflows and evaluation frameworks. Such strategies may help reduce interpretative variability, enhance the overall classification accuracy, highlight the potential for more reliable and consistent clinical decision-making, and ensure more robust diagnostic outcomes.

While D-DDPM demonstrates strong potential in producing clinically interpretable synthetic ultrasound images, the notable variability in expert interpretation—particularly for malignant cases—highlights the need for improved data curation and radiologist support tools. A related limitation of generative models, such as diffusion models, is their tendency to generate images where malignant and benign classes are not clearly distinguished (e.g., generating benign-looking features from malignant cases or vice versa), leading to discrepancies from real data and potential instability if evaluation relies solely on visual assessment without expert feedback. Incorporating clinical guidance—through text prompts or reference images—may improve model training and enhance the accuracy and reliability of the generated outputs. These considerations should guide future development and integration of generative imaging models into diagnostic practice.

Although this study establishes the clinical realism and reliability of synthesized ultrasound images and masks through direct radiologist assessment, our evaluation was constrained by a small cohort of three radiologists reviewing 41 images and a single zero-shot LLM configuration, limiting the statistical robustness and scope of our conclusions. Future work should extend beyond perceptual evaluation by incorporating deep learning-based segmentation and classification models to rigorously quantify whether synthetic data improves downstream diagnostic performance on real clinical tasks, thereby establishing a complete validation framework that combines both physician judgment and computational benchmarks.

7 Conclusion

This work evaluated diffusion models for breast ultrasound cancer imaging through a human-in-the-loop framework that combined quantitative metrics, radiologist assessments, and LLM outputs. While D-DDPM produced realistic images and masks that sometimes achieved higher agreement than real data, inter-rater reliability remained low, particularly for malignant cases. Radiologists consistently outperformed LLMs, though majority voting improved diagnostic accuracy, underscoring the importance of expert consensus. These findings highlight both the potential and the risks of diffusion-based biomedical data generation: synthetic data can augment limited datasets, but expert validation is essential to ensure clinical trust and safe integration into medical research.

References

[1] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J.R. Gralow, F. Cardoso, S. Siesling, and I. Soerjomataram. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast*, 66:15–23, Dec 2022.

[2] Joanne Kim, Andrew Harper, Valerie McCormack, Hyuna Sung, Nehmat Houssami, Eileen Morgan, Miriam Mutebi, Gail Garvey, Isabelle Soerjomataram, and Miranda M Fidler-Benaoudia. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nature Medicine*, pages 1–9, 2025.

[3] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[4] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

[6] Teerath Kumar, Rob Brennan, Alessandra Mileo, and Malika Bendechache. Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access*, 2024.

[7] Jie Luo, Heqing Zhang, Yan Zhuang, Lin Han, Ke Chen, Zhan Hua, Cheng Li, and Jiangli Lin. 2s-busgan: A novel generative adversarial network for realistic breast ultrasound image with corresponding tumor contour based on small datasets. *Sensors*, 23(20):8614, 2023.

[8] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[9] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Fahmy Aly. Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. *Int. J. Adv. Comput. Sci. Appl*, 10(5):1–11, 2019.

[10] Zhaoshan Liu, Qiujie Lv, Chau Hung Lee, and Lei Shen. Gsda: Generative adversarial network-based semi-supervised data augmentation for ultrasound image classification. *Heliyon*, 9(9), 2023.

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[12] Kyuri Kim, Yoonho Na, Sung-Joon Ye, Jimin Lee, Sung Soo Ahn, Ji Eun Park, and Hwiyoung Kim. Controllable text-to-image synthesis for multi-modality mr images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7936–7945, 2024.

[13] Francisco Carrillo-Perez, Marija Pizurica, Yuanning Zheng, Tarak Nath Nandi, Ravi Madduri, Jeanne Shen, and Olivier Gevaert. Generation of synthetic whole-slide image tiles of tumours from rna-sequencing data via cascaded diffusion models. *Nature Biomedical Engineering*, 9(3): 320–332, 2025.

[14] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[15] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[16] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[17] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, 2022.

[18] Bayrem Kaabachi, Jérémie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, Bogdan Kulynych, Fabian Prasser, and Jean Louis Raisaro. A scoping review of privacy and utility metrics in medical synthetic data. *NPJ digital medicine*, 8(1):60, 2025.

[19] Luke William Sagers, Aashna Shah, Sonnet Xu, Roxana Daneshjou, and Arjun Kumar Manrai. Demo track: Directing generalist vision-language models to interpret medical images across populations. In *GenAI for Health: Potential, Trust and Policy Compliance*.

[20] Hejie Cui, Lingjun Mao, Xin Liang, Jieyu Zhang, Hui Ren, Quanzheng Li, Xiang Li, and Carl Yang. Biomedical visual instruction tuning with clinician preference alignment. *Advances in neural information processing systems*, 37:96449–96467, 2024.

[21] Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. *Research square*, pages rs–3, 2023.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[23] Abdalrahman Alblwi, Saleh Makkawy, and Kenneth E Barner. D-ddpm: Deep denoising diffusion probabilistic models for lesion segmentation and data generation in ultrasound imaging. *IEEE Access*, 2025.

[24] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pages 1336–1348. PMLR, 2022.

[25] Xin Yuan and Michael Maire. Factorized diffusion architectures for unsupervised image generation and segmentation. *arXiv preprint arXiv:2309.15726*, 2023.

[26] Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1):148, 2024.

[27] Yuxuan Chen, Haoyan Yang, Hengkai Pan, Fardeen Siddiqui, Antonio Verdone, Qingyang Zhang, Sumit Chopra, Chen Zhao, and Yiqiu Shen. Burextract-llama: An llm for clinical concept extraction in breast ultrasound reports. In *Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine*, pages 53–58, 2024.

[28] Moi Hoon Yap, Gerard Pons, Joan Marti, Sergi Ganau, Melcior Sentis, Reyer Zwiggelaar, Adrian K Davison, and Robert Marti. Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 22(4): 1218–1226, 2017.

[29] Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024.

[30] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024.

[31] Eyal Betzalel, Coby Penso, Aviv Navon, and Ethan Fetaya. A study on the evaluation of generative models. *arXiv preprint arXiv:2206.10935*, 2022.

[32] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596, 2008.

Appendix

How confident are you in the annotation quality of this image?

1 - Very confident	(80–100%)
2 - Somewhat confident	(60–79%)
3 - Neutral	(40–59%)
4 - Somewhat unsure	(20–39%)
5 - Not confident at all	(0–19%)

What is the class of this image?

1 - Normal	*// healthy*
2 - Benign	*// non-cancerous tumor*
3 - Malignant	*// cancerous tumor*

Does the description accurately reflect the content and nature of this image?

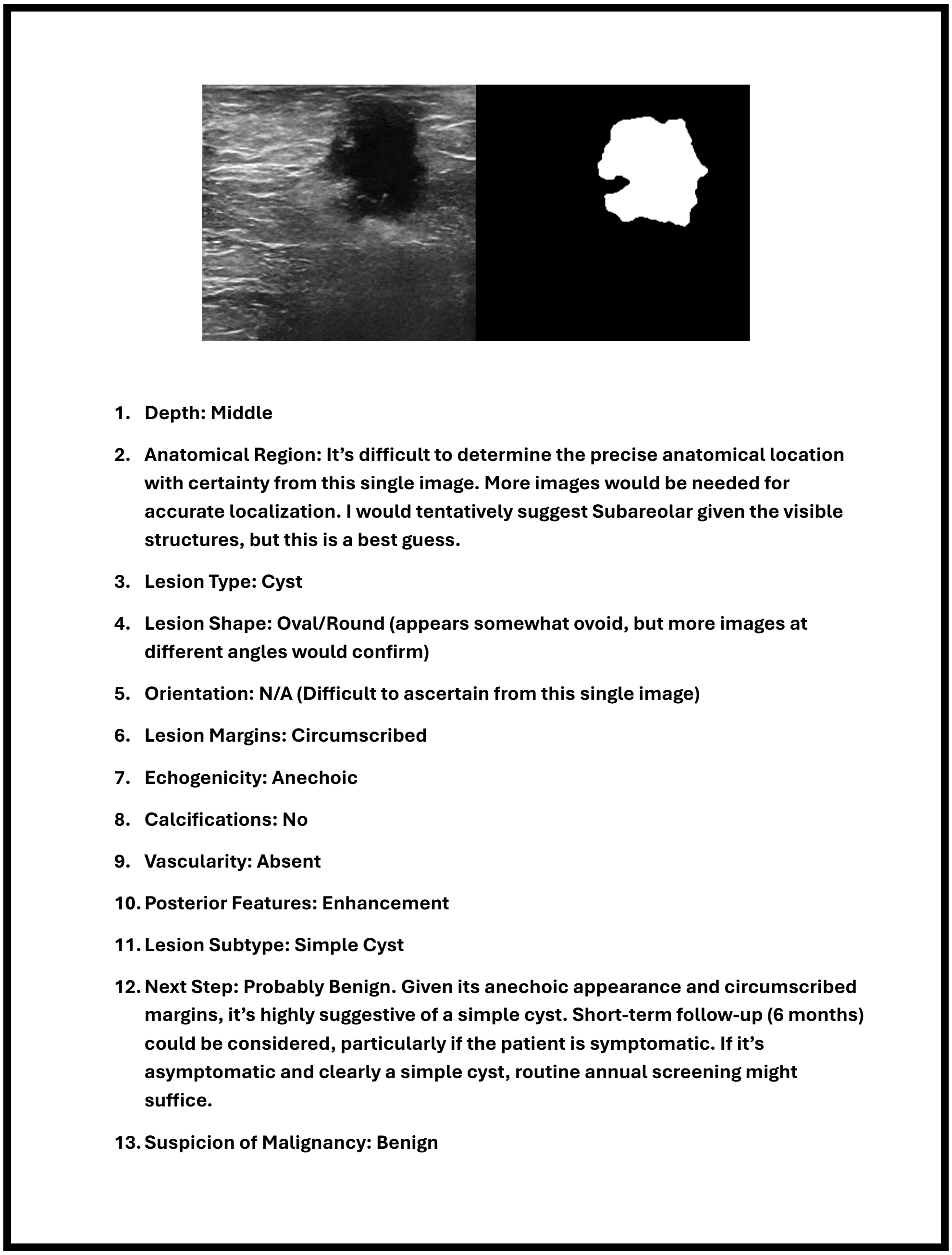
1. Yes, the description perfectly matches the image.
2. Mostly, but there are some minor discrepancies.
3. Neutral, the description and image are somewhat aligned but not entirely.
4. No, there are significant differences between the description and the image.
5. I'm not sure, I need further clarification.

Prompt used to generate interpretation from the LLM based on the synthetic ultrasound image

You are a helpful assistant for healthcare professionals. Based on attached ultrasound images, assign labels for each condition. Please return the values for the following keys and their corresponding options:

- **1. Depth**: Posterior, Middle, Anterior, N/A
- **2. Anatomical Region**: Retroareolar, Axillary Tail, Periareolar, Subareolar, Retropectoral, N/A
- **3. Lesion Type**: Nodule, Cyst, Mass, Lymph Node, Scar, Duct, Seroma, Post-Surgical Change, Post-Biopsy, N/A
- **4. Lesion Shape**: Oval, Round, Irregular, N/A
- **5. Orientation**: Parallel, Non-Parallel, Other, N/A
- **6. Lesion Margins**: Circumscribed, Obscured, Angular, Microlobulated, Spiculated, Lobulated, Irregular, Septated, N/A
- **7. Echogenicity**: Anechoic, Hyperechoic, Hypoechoic, Isoechoic, Heterogeneous, Solid, N/A
- **8. Calcifications**: Yes, No, N/A
- **9. Vascularity**: Absent, Present, N/A
- **10. Posterior Features**: Enhancement, Shadowing, N/A
- **11. Lesion Subtype**: Abnormal Lymph Node, Simple Cyst, Complicated Cyst, Cyst with Debris, Reactive Lymph Node, Fat Necrosis, Sebaceous Cyst, Lipoma, Cyst Cluster, Focally Ectatic Duct with Debris, N/A
- **12. Next Step**: 1 Year Screening Mammogram, MRI Follow Up, 6 Months Follow-Up, 12 Months Follow-Up, Fine Needle Aspiration, Ultrasound Guided Core Biopsy, Surgical Excision, N/A
- **13. Suspicion of Malignancy**: Low, Moderate, High, Benign, Probably Benign, Negative

Please return the corresponding values for each key from the list above, based on the ultrasound imaging results provided.



1. Depth: Middle
2. Anatomical Region: It’s difficult to determine the precise anatomical location with certainty from this single image. More images would be needed for accurate localization. I would tentatively suggest Subareolar given the visible structures, but this is a best guess.
3. Lesion Type: Cyst
4. Lesion Shape: Oval/Round (appears somewhat ovoid, but more images at different angles would confirm)
5. Orientation: N/A (Difficult to ascertain from this single image)
6. Lesion Margins: Circumscribed
7. Echogenicity: Anechoic
8. Calcifications: No
9. Vascularity: Absent
10. Posterior Features: Enhancement
11. Lesion Subtype: Simple Cyst
12. Next Step: Probably Benign. Given its anechoic appearance and circumscribed margins, it’s highly suggestive of a simple cyst. Short-term follow-up (6 months) could be considered, particularly if the patient is symptomatic. If it’s asymptomatic and clearly a simple cyst, routine annual screening might suffice.
13. Suspicion of Malignancy: Benign

Figure 3: A description generated by Gemini on an ultrasound image, highlighting key factors in breast cancer diagnostics.

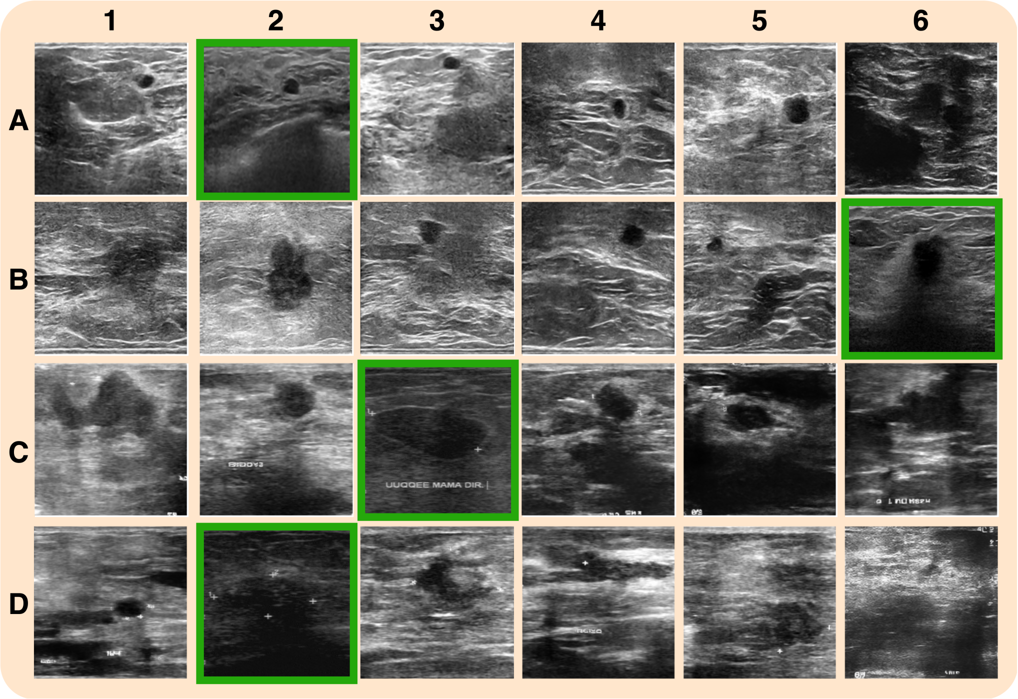


Figure 4: Examples of generated and original ultrasound images for synthetic image identification. Green outlines indicate original images from real patients.

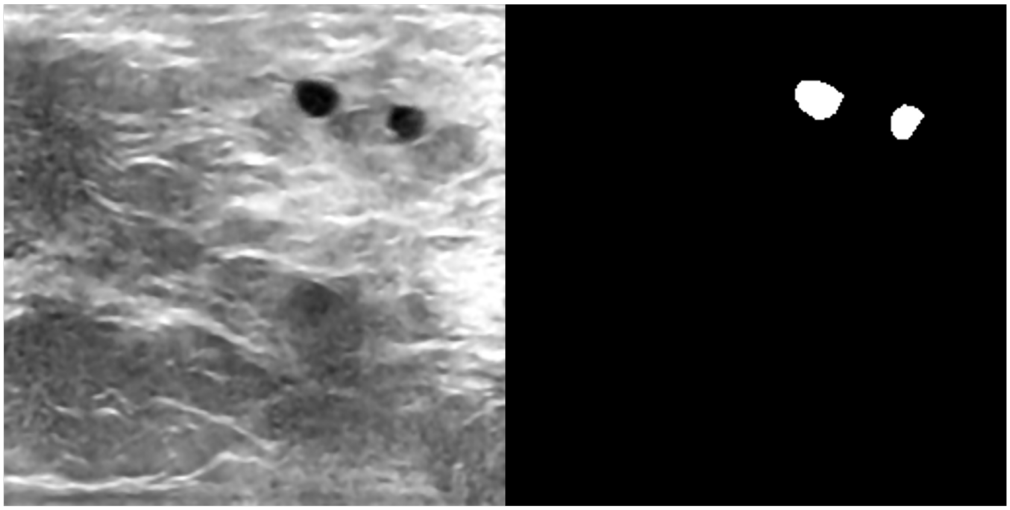


Figure 5: An example of a synthetic ultrasound image and its mask used for annotation confidence evaluation.

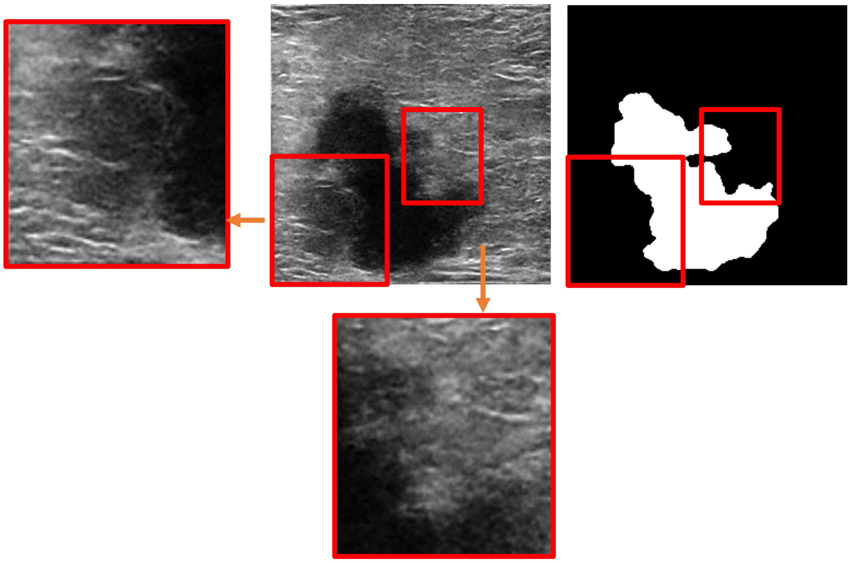


Figure 6: Example of a generated image with low confidence among radiologists. The red box highlights defects, notable noise, and inconsistencies between the annotation and the corresponding image, indicating areas of unreliability.

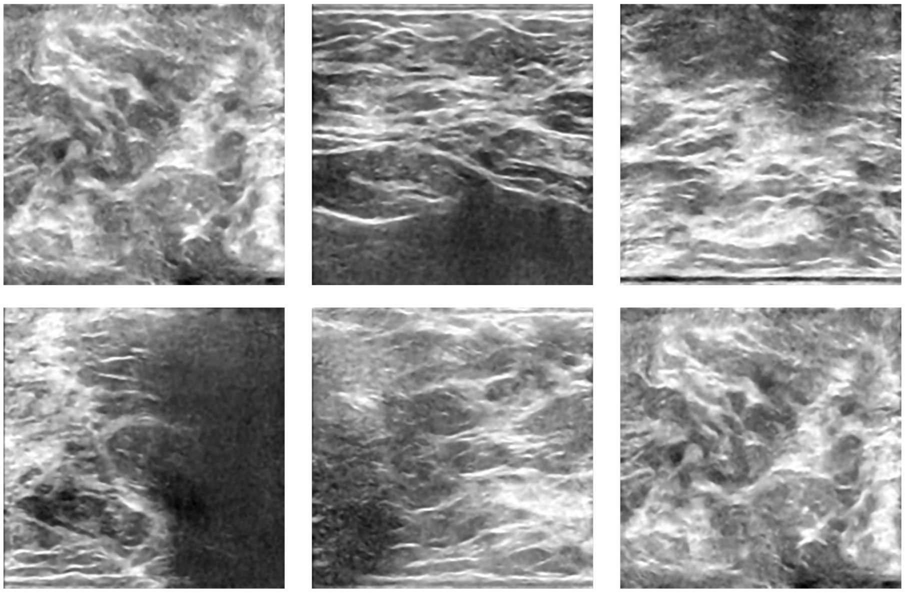


Figure 7: Examples of generated images from the benign class that visually resemble healthy tissue, with no obvious signs of tumor presence, refer to Table 6

Table 6: Annotations of Generated Benign Images with Healthy Appearance.

R1	R2	R3	Ground Truth B = Benign	Majority Vote
Normal	Benign	Normal	Benign	Normal
Benign	Normal	Normal	Benign	Normal
Malignant	Benign	Malignant	Benign	Malignant
Malignant	Malignant	Malignant	Benign	Malignant
Malignant	Malignant	Benign	Benign	Malignant
Normal	Benign	Normal	Benign	Normal

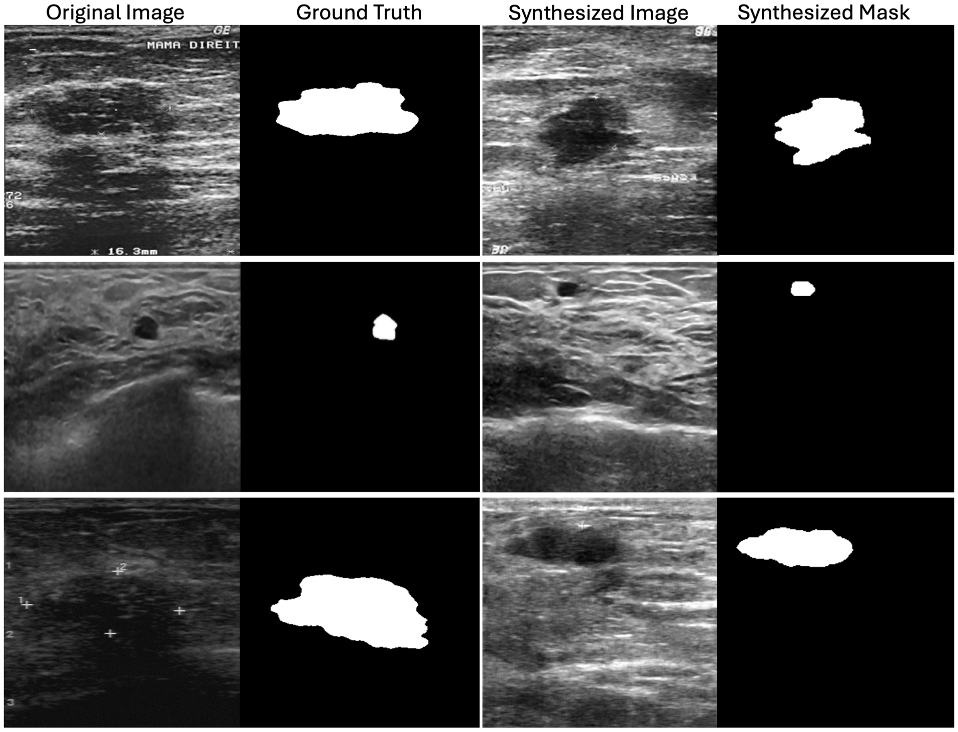


Figure 8: Comparison between real ultrasound images and generated samples with corresponding masks, illustrating variations in tumor size and shape that closely resemble real-world patterns.