Generation and Validation of Synthetic Mammography Images Using Diffusion and Blending Approaches

Fatima Alshihri[1] fsalshihri@humain.com
Aqeelah J. Makki[1] amakki@humain.com
Fay A. Al Shammari[1] falshammari@humain.com
Abdullah A. Alsheeha[1] aalsheeha@humain.com
Shahad M. Albati[1] salbati@humain.com
Pedro Jose Moreno Mengibar[1] pedro@humain.com
Mona F. Almehaid[1] mona@humain.com
Salman Albshan[2] Salbeshan@ksu.edu.sa
Hend Samir Ibrahim[3] Ahmedhe@rchsp.med.sa
Manal Ahmed ElRefaei[4] manalelrefaei@gmail.com
Fatma Eliraqi[5] fahali@moh.gov.sa
Yousef Alkathiri[6] Y.Alkathiri@rfhc.gov.sa
Ammar Almoalem[6] A.almoallem@rfhc.gov.sa
Hamad Al-Msalam[6] H.almsalam@rfhc.gov.sa
[1] *Humain Company, Saudi Arabia*
[2] *King Saud University, Saudi Arabia*
[3] *Royal Commission Medical Centre*
[4] *Alahrar Teaching Hospital, General Organization for Teaching Hospitals and Institutes*
[5] *The General Organization for Teaching Hospitals and Institutes, (National Institute of Neurology and Urology), Cairo*
[6] *King saud medical city*

Editors: Under Review for MIDL 2026

Abstract

Breast cancer is among the most prevalent diseases that affect women and remains a significant global health concern. Because medical imaging is highly confidential, publicly accessible mammography datasets are limited. This shortage has forced us to explore alternative ways of obtaining reliable training data. Generative models serve as an alternate data augmentation method to mitigate the data scarcity issue encountered in medical imaging. Diffusion models have garnered significant attention due to their novel generation methodology, the superior quality of the produced images, and their comparatively simpler training process. In this paper, we explore the use of three different techniques: Stable Diffusion 3.5, Poisson blending, and Stable Diffusion Inpainting for generating synthetic mammography images. Using these methods, we created thousands of images that span multiple categories of BI-RADS, breast laterality, and standard mammographic views (MLO and CC). To evaluate the generated images, we invited seven radiologists to review them using a dedicated assessment tool. Each radiologist was asked not only to decide whether an image was real or synthetic but also to assign a BI-RADS category as would be in routine practice. What stood out in the results was the amount of confusion: about 37% of the synthetic images were judged to be real, and approximately 30% of the authentic images were misidentified. The difficulty experienced by Radiologists in distinguishing between the two indicates that synthetic images are nearing the visual complexity of real mammograms. These discoveries indicate a distinct opportunity to utilize synthetic mammograms in research, particularly in contexts where access to extensive, diverse clinical datasets is constrained.

Keywords: Stable Diffusion 3.5, Inpainting, Poisson blending, mammography, Otsu

1. Introduction

Breast cancer remains one of the most common illnesses affecting women and continues to pose a major worldwide health challenge. Mammography plays a central role in modern medical practice and is considered the standard approach for early detection (Yoon and Kim, 2021). Despite recent advances in artificial intelligence, building reliable systems for BI-RADS classification and disease staging is still difficult. One of the main challenges stems from the limited number of publicly available mammography datasets. Strict privacy rules make it difficult to share clinical images, and the datasets that can be accessed are usually quite small, often coming from a single hospital and showing uneven representation across BI-RADS categories. Because of this, many AI models end up overfitting to the narrow data they are trained on and struggle to perform well when applied to new or more diverse patient populations. One promising way to ease these limitations is through the generation of realistic synthetic mammograms images that capture the appearance, anatomy, and fine textural details of true breast tissue. When produced with care, synthetic data can help broaden existing datasets, improve representation of under-sampled BI-RADS categories, and ultimately enhance the accuracy and robustness of AI systems used for lesion detection and clinical decision support (Shah et al., 2024).

A number of techniques have been used to enhance conventional data augmentation methods and to expand medical datasets, with generative adversarial networks (GANs) (Goodfellow et al., 2014) becoming the predominant technology for many years as a result of their high image quality and photorealistic qualities. While GAN-like architectures are capable of generating medical imaging synthesis, they are often hampered by unstable trainings, insufficient diversity in generation, and poor sample quality. (Kazerouni et al., 2023) (Müller-Franzes et al., 2022).

Diffusion models are a cutting-edge class of generative models that have proven to be exceptionally efficient in producing synthetic data, serving as a valuable complement to existing actual data and as a generative prior in biomedical inverse imaging problems. The conditional diffusion model stable diffusion (SD) (Ommer et al., 2022) uses text prompts for generation conditioning. In this paper, we set out to tackle two persistent challenges in medical imaging: the lack of sufficiently large datasets and the privacy constraints that limit data sharing. To do this, we generated synthetic mammography images using three different approaches: Stable Diffusion 3.5 (AI, 2024), Stable Diffusion Inpainting (Montoya-del Angel et al., 2024), and Poisson blending (Shen and Li, 2023). We then worked with seven radiologists, each with several years of clinical experience, to assess how realistic and diagnostically useful the images from each method appeared. Their feedback helped us understand the strengths and weaknesses of the three techniques. To encourage continued work in this area, we have also made the synthetic dataset publicly accessible on Hugging

Face https: https://huggingface.co/syntheticmammo. This work makes the following key contributions:

- A Large-Scale, Class-Controlled Synthetic Mammography Dataset. We created more than **48,000** BI-RADS–specific synthetic mammograms covering different views and lesion severities, resulting in one of the most comprehensive synthetic mammography collections currently available for research.

- A Novel BI-RADS–Conditioned Inpainting Framework. Our method trains three separate inpainting models one each for BI-RADS 3, 4, and 5. This allows the generated lesions to more closely reflect the expected appearance and severity associated with each classification.

- Anatomically Guided Lesion Placement Using Brightness–Texture Constraints. To make sure lesions appear only in areas that truly resemble breast tissue, we designed a placement approach that takes into account the natural intensity range of each breast, the local texture patterns around the insertion site, and anatomical masks that define the usable regions. By considering all three factors, the generated lesions settle into the image more naturally, avoiding the sharp edges or misplaced insertions that can break the illusion of realism.

- Blinded Radiologist Reader Study. Blinded Radiologist Reader Study. We conducted a blinded review with seven radiologists who were asked to decide whether images were real or synthetic and to assign BI-RADS categories. Their assessments provided valuable clinical insight into how convincing the generated images were and revealed perceptual differences among the various synthesis techniques.

2. Datasets

We utilized the publicly available Vindermamo dataset for the training of the diffusion models to generate synthetic mammographic images. The dataset comprises approximately 10,000 images that are annotated according to the BI-RADS classification system, ranging from categories 1 to 5.

3. Synthetic Mammogram Generation

To increase the number of training samples, we produced the synthetic mammography images via three complementing methodologies. Initially, we utilized Stable Diffusion 3.5-Large (AI, 2024) , which is based on a deep generative framework directed by descriptive prompts. It systematically enhances images from stochastic noise using a regulated noise schedule and tuned hyperparameters.

We utilized Poisson blending, a mathematical method that enables the smooth integration of visual regions. This method guarantees the maintenance of local illumination and brightness consistency, enabling smooth transitions between the original and inserted regions without noticeable boundaries. (Shen and Li, 2023). Ultimately, inpainting was utilized via mask-guided restoration, analogous to the method used in Stable Diffusion. Utilize inpainting to complete absent or obscured regions, guaranteeing contextual consistency and accurate anatomy in the restored image (Montoya-del Angel et al., 2024).

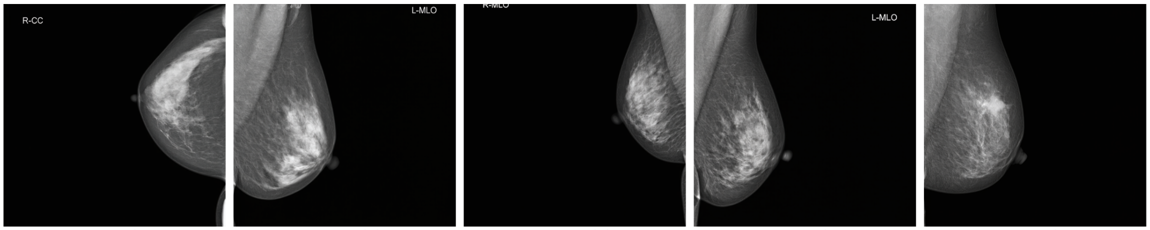
3.1. Stable Diffusion 3.5

The SD3.5 technique was developed using the VinDr-Mammo dataset with its standard train, validation, and test splits. Each image was assigned a short caption describing its laterality and view, breast density (A–D), and BI-RADS category, which served as the text conditioning input. Stable Diffusion 3.5-Large was then fine-tuned with LoRA (rank 64, dropout 0.1) applied to the UNet attention layers. Training was conducted on 8 NVIDIA H100 GPUs using `Accelerate` with a resolution of 1024×1024, a per-GPU batch size of 3, gradient accumulation of 8, and 2,000 total steps. The model was optimized with AdamW (learning rate 1×10^{-4}, cosine schedule, 100 warmup steps), using fp16 precision and gradient checkpointing. Approximately 20,000 images were used throughout training, which required 18–20 hours. Five checkpoints were saved for each BI-RADS category, and the same prompt format was applied during both training and inference. As shown in Figure 1, the model was trained using a structured BI-RADS prompt. Figure 2 shows five different BI-RADS categories generated using this technique.

Input Prompt (Text)

A Breast Mammogram
Laterality: Right
View Position: MLO (Medio-Lateral Oblique)
Breast Density: D
Photometric Interpretation:
MONOCHROME2
BI-RADS: 5

Figure 1: Input prompt used for Stable Diffusion 3.5 text-to-image generation.



BI-RADS 1 BI-RADS 2 BI-RADS 3 BI-RADS 4 BI-RADS 5

Figure 2: BI-RADS 1-5 generated by SD3.5

3.2. Poisson blending

We utilize BI-RADS 1 mammograms as background images and BI-RADS 5 clipped lesions as foreground elements. Device metadata, including manufacturer and model, is extracted from the dataset's CSV file to create an ID→(manufacturer, model) mapping. This mapping facilitates a manufacturer matching policy, wherein lesions are, by default, taken from the same device manufacturer as the base mammography to maintain visual uniformity. Utilization of the Poisson image editing approach to provide realistic and varied training examples. In contrast to traditional copy-paste techniques, which frequently create artificial boundaries in medical images. Consequently, to impede training performance, we employ Poisson image editing for image pasting to enable smooth boundary transitions (Pérez et al., 2003).

3.2.1. Anatomical Region Extraction

We derive a mask that isolates the main anatomical region of the image from the background. The basic image is initially subjected to histogram equalization to standardize illumination. Otsu thresholding is subsequently employed to achieve an initial binary segmentation (Otsu, 1979). Due to Otsu's potential misclassification of the darker background as the foreground, the algorithm assesses whether side accurately depicts the anatomical region; if necessary, the mask is flipped to guarantee that the brighter mode aligns with the actual tissue.

A morphological closing operation is utilized to eliminate minor gaps and refine borders, retaining just the largest connected component as the definitive anatomical region mask.

We calculate the 10th and 98th percentile intensities within this mask to define the brightness profile of each image. These values establish an image specific intensity window utilized subsequently to confirm that lesions are positioned in areas where the surrounding intensities reside within a believable and natural range for that image. Figure 3 shows the pipeline of the proposed technique.

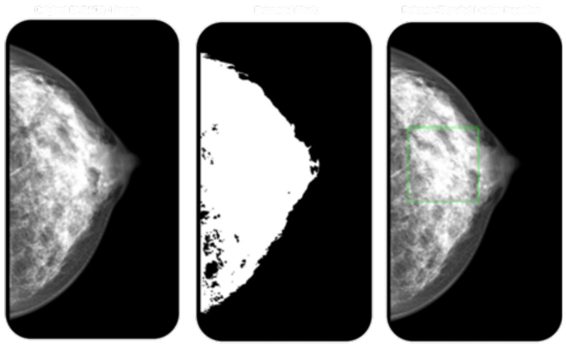


Figure 3: The pipeline of the Poisson blending technique

3.3. Stable Diffusion Inpainting

The overall workflow is inspired by the MAM-E(Montoya-del Angel et al., 2024) technique, integerating several aspects of their approach into this pipeline. Similar to MAM-E, the generation process used an allowed region within the breast and applied rejection sampling to ensure that lesions were placed only in valid locations. The general inpainting framework and the idea of sampling lesion sizes from empirically informed ranges were also preserved. However, several key differences distinguish this pipeline from the original MAM-E method.

In MAM-E the inpainting model was design to generate a non-specific lesion and did not incorporate any BI-RADS class throughout training or inference. In contrast, this pipeline trained three separate checkpoints, one each for BI-RADS 3, 4, and 5, allowing the generation to be explicitly class-specific and more aligned with clinical lesion patterns.

The placement strategy was also refined. While MAM-E defined an allowed region, this implementation extended it by excluding a fixed 24-pixel margin near the breast boundary, reducing the risk of unrealistic edge insertions. In terms of lesion dimensions, MAM-E sampled sizes based on broad empirical ranges. Here, bounding box width and height were drawn from empirically derived distributions after removing outliers, producing more realistic and consistent lesion shapes.

3.3.1. Dataset Preparation and Preprocessing

The VinDr-Mammo training split has been utilized, which contains mammography images with bounding-box annotations. Only BI-RADS 3-5 cases were used for fine-tuning, as the inpainting method requires lesion bounding boxes, which are not available for BI-RADS 2. After filtering, the training set consisted of approximately 800 BI-RADS 3 images, 800 BI-RADS 4 images, and 200 BI-RADS 5 images.

For synthetic lesion insertion, we used all 13k BI-RADS 1 (healthy) images from VinDr-Mammo. All images were resized to 512×512 to match the resolution expected by the base stabilityai/stable-diffusion-2-inpainting model and converted to 3-channel RGB, as required by the Stable Diffusion architecture. Pixel intensities were normalized to [-1, 1]. All training masks were binary.

Although fine-tuning was performed at 512×512, inference was carried out at 1024×1024, leveraging the fully convolutional UNet and VAE, which support arbitrary resolutions at test time, a common practice in latent diffusion pipelines.

During training, lesion masks were created directly from the VinDr bounding-box annotations. The masks were resized with the same interpolation technique applied to their associated images, reserving consistent alignment.

3.3.2. Lesion Mask Construction and Inpainting Workflow

During generation, the breast region was first isolated. Otsu thresholding was used to separate the breast from the background, and the largest connected component was retained as the main breast area. A small erosion step was then applied, producing a clean and reliable binary breast mask. Next, a safe inpainting region was defined. A fixed 24-pixel margin was removed from the breast boundary to prevent synthetic lesions from appearing too close to the edges, where artifacts were more likely. This step resulted in a conservative "allowed region" within the breast.

Inside this allowed region, random lesion bounding boxes were generated using rejection sampling. The width and height of each box were sampled uniformly from ranges derived from real bounding-box sizes in the training set, after removing outliers. This ensures that the generated lesion sizes are aligned with realistic distributions.

Finally, three synthetic versions of each healthy mammogram were created—one for each BI-RADS category (3, 4, and 5). As a result, every healthy image produced three distinct synthetic samples. Figure 4 shows the pipline of Stable Diffusion Inpainting technique.

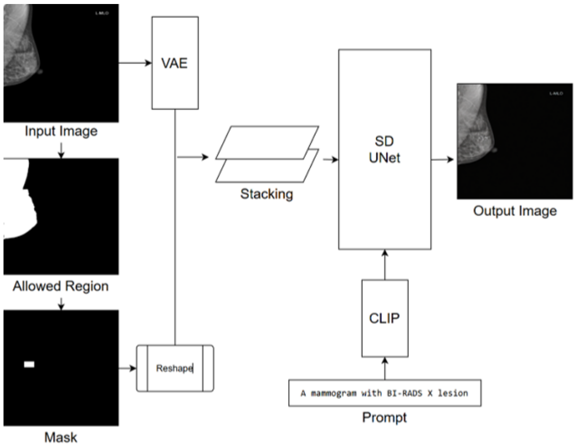


Figure 4: Overview of the Stable Diffusion Inpainting Pipeline

3.3.3. Fine-Tuning Pipeline

The stabilityai/stable-diffusion-2-inpainting model was fine-tuned using bf16 precision. Fine-tuning was carried out separately for BI-RADS 3, 4, and 5, resulting in three independent checkpoints. This separation allowed each model to specialize in generating lesions that matched the characteristics of its specific BI-RADS category.

Training was performed on 8 NVIDIA H100 GPUs, with each GPU processing a batch size of 4, resulting in a total batch size of 32. The AdamW optimizer was used with a learning rate of 5×10^{-6}, a weight decay of 0.01, and a constant learning-rate schedule. The training objective followed the standard noise-prediction MSE loss. The text encoder remained frozen throughout training, and the prompt was fixed as: "A mammogram with BI-RADS X lesion."

Each model was fine-tuned for 10,000 training steps, which corresponded to approximately 50 epochs on a single-GPU equivalent. Checkpoints were saved every 1,000 steps. The full training process for each BI-RADS-specific model required about two hours. Training was implemented using a custom PyTorch loop that built on Hugging Face diffusers, transformers (for CLIP components), and torchvision for preprocessing. Multi-GPU training was managed through Hugging Face Accelerate, which provided DDP-style distributed training. Deterministic seeds were assigned to each image to ensure the results could be reproduced reliably. To do this, the image filename was converted into a hash and then combined with a base seed 700 for creating the masks and 1337 for the sampling step. Using this approach kept the outputs consistent every time the process was repeated.

For inference, the DPM-Solver++ multi-step scheduler was used with a guidance scale of 7 and 24 sampling steps. The prompt matched the one used during training: "A mammogram with BI-RADS X lesion." All outputs were generated at a resolution of 1024×1024. The inpainting mask was defined by a randomly selected bounding box placed within the segmented breast region. Figure 5 shows example BI-RADS outputs generated by this method.

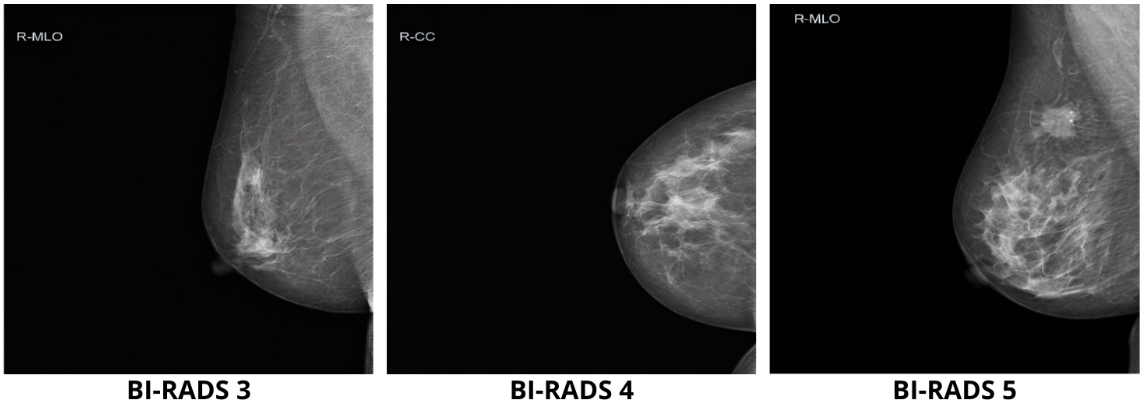


Figure 5: Samples of the BI-RADS images generated by Stable Diffusion Inpainting

4. Evaluation

To evaluate the three synthesis techniques, an assessment tool was developed and distributed to radiologists whose clinical experience ranged from five to twenty years. The tool included a total of 2196 mammogram images: 1062 real images from the VinDr-Mammo dataset and 1,134 synthetic images produced by the three methods (347 from inpainting, 366 from Poisson blending, and 421 from the SD3.5 model). The images were presented in a fully randomized order. Each radiologist received a unique sequence containing a mixture of real and synthetic cases. They were informed that the set could contain any combination: entirely real images, entirely synthetic images, or any proportion in between.

4.1. Results and discussion

Table 1 summarizes the radiologists' performance in classifying real and synthetic images. The results of the evaluation show that radiologists could recognize a considerable number of both real and synthetic mammograms, although their accuracy was not the same across the two types. Real images were identified more reliably: 70.7% were classified correctly, while 29.3% were mistaken for synthetic. This suggests that the visual cues present in genuine mammograms remain more familiar and intuitive to interpret. Synthetic images, on the other hand, proved more challenging. Only 63% were correctly labeled, and 37% were misclassified, indicating that even highly realistic synthetic studies can still contain small irregularities or occasionally a certain uniformity that may unsettle readers' expectations. The fact that radiologists made errors in both directions highlights how closely the synthetic images resembled actual breast anatomy. In several cases, even highly experienced readers paused and reconsidered their decisions, which reflects how much detail these generative models are now able to capture. At the same time, the results remind us that although synthetic mammography holds real promise for education, training, and performance evaluation, the technology is still maturing and has room to grow. Continued refinement is needed to ensure that these generated images portray the full breadth of subtle variations and complexities that appear in everyday clinical practice, rather than only approximating them.

Table 1: Radiologists' Classification Performance on Real and Synthetic Images

Category	**Count**	**Percentage**
Total Images Evaluated	2,196	100%
Real Images	1,062	48.4%
Synthetic Images	1,134	51.6%
Correctly Identified Real	751	70.7%
Misclassified Real	311	29.3%
Correctly Identified Synthetic	716	63%
Misclassified Synthetic	418	37%

Table 2: Radiologist Identification Accuracy for Synthetic Images per Technique

Technique	**Total Images**	**Correctly Identified**	**Accuracy (%)**
SD3.5	421	261	61.99%
Inpainting	347	203	58.50%
Poisson	366	252	68.85%

The findings in Table 2 make it clear that the synthetic mammograms looked remarkably real, often enough to mislead even experienced radiologists. Each technique was convincing in its own way: SD3.5 led to mistakes in about 38% of its images, Poisson blending in roughly 31%, and inpainting in about 42%. The results show that the inpainting method produced the most convincing images overall. Out of all three techniques, it was the one that most often caused radiologists to mistake synthetic images for real ones. From a clinical point of view, these findings tell two different stories. On the positive side, they show just how far synthetic mammography has come. Many of the generated images looked realistic enough that even experienced breast radiologists struggled to tell the difference. Their overall accuracy was roughly two-thirds, meaning that about one in every three images real or synthetic was misclassified. As a result of his confusion, the synthetic images resemble real mammograms in many ways. In particular, they may be useful for teaching, practice exams, and expanding datasets, especially when representing rare findings or specific categories of breast density (Bart and Hegdé, 2018; Al-Dhabyani et al., 2019; Saffari et al., 2020; Garrucho et al., 2023a). That said, the ability of an image to fool a reader does not necessarily make it clinically interchangeable with a real one. In our study, radiologists were not always consistent when assigning BI-RADS categories, even in cases where they struggled to distinguish between real and synthetic images. This reminds us that visual realism alone is not enough to judge whether these images are ready for true diagnostic use (Liu et al., 2023) (Xing et al., 2023). Future work should focus on whether synthetic images can genuinely support diagnostic tasks such as detecting lesions, describing their features, assigning BI-RADS scores, and estimating breast density before they can be considered a dependable substitute in clinical settings (Mariscotti et al., 2017),(Garrucho et al., 2023),(Alshafeiy et al., 2017).

4.2. Radiologist Survey Results and Perceptual Analysis

The survey in Appendix A explored how practicing radiologists distinguished real mammograms from AI-generated synthetic ones. Seven radiologists participated in the post examination survey, representing a mix of training, clinical backgrounds, and years of experience. Their feedback offers a detailed look at how specialists view conditions and how reading habits affect performance when synthetic images are added to the diagnostic environment. The radiologists had between 2 and 21 years of practice. Viewing time per image ranged from just a few seconds to close to 20 seconds, suggesting different reading styles and levels of scrutiny. Time-of-day effects were noticeable as well, with several completing the task late in the evening. These elements lighting, screen quality, fatigue, and environment likely influenced both perceptual sharpness and decision confidence.

Most radiologists stated the task was challenging but not impossible. The synthetic visuals were convincing while rendering differentiation harder than expected. Their confidence varied since some synthetic investigations mirrored real mammographic anatomy so closely that it raised question. Their comments demonstrate how far modern generative models have developed in capturing the subtle visual signals and textures physicians need to read mammograms. Participants indicated they made decisions utilizing many clues. Background texture, parenchymal pattern, lesion appearance, and natural artifacts were important. Occasionally, manufactured lesions did not fit clinical patterns in terms of their forms and borders.

It was noted by several participants that the synthetic cases appeared more realistic than expected, a strong indication of the model's overall success. A majority of radiologists supported the use of synthetic images to teach, train, prepare for exams, and develop AI datasets. However, confidence dropped when direct clinical applications were considered. In general, participants viewed synthetic mammography as a useful supplementary resource but were cautioned against using it in real clinical situations.

5. Conclusion

In this paper we investigated the use of different techniques to generate mammography, and we asked professional radiologists to evaluate these images. The results indicate that these techniques generate images that look real, and the radiologists were often unable to distinguish between the two. These discoveries indicate a distinct opportunity to utilize synthetic mammograms in research, particularly in contexts where access to extensive, diverse clinical datasets is constrained.

References

Stability AI. Stable diffusion 3.5 large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, 2024. Model card. Accessed: 2025-01-15.

TI Alshafeiy, A Wadih, BT Nicholson, CM Rochman, HR Peppard, JT Patrie, and JA Harvey. Comparison between digital and synthetic 2d mammograms in breast density interpretation. *American Journal of Roentgenology*, 209(1):W36–W41, 2017.

Lidia Garrucho, K Kushibar, R Osuala, O Diaz, A Catanese, J Del Riego, M Bobowicz, F Strand, L Igual, and K Lekadir. High-resolution synthesis of high-density breast mammograms: Application to improved fairness in deep learning based mass detection. *Frontiers in Oncology*, 12:1044496, 2023.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, 88:102846, 2023.

Z Liu, S Wolfe, Z Yu, R Laforest, JC Mhlanga, TJ Fraum, M Itani, F Dehdashti, BA Siegel, and AK Jha. Observer-study-based approaches to quantitatively evaluate the realism of synthetic medical images. *Physics in Medicine & Biology*, 68(7):074001, 2023.

Giuseppina Mariscotti, Marco Durando, Nehmat Houssami, Michele Fasciano, Alessandro Tagliafico, Daniela Bosco, Carolina Casella, Carlo Bogetti, Lorenzo Bergamasco, Paolo Fonio, and G Gandini. Comparison of synthetic mammography, reconstructed from digital breast tomosynthesis, and digital mammography: evaluation of lesion conspicuity and birads assessment categories. *Breast Cancer Research and Treatment*, 166(3):765–773, 2017.

Ricardo Montoya-del Angel, Karla Sam-Millan, Joan C Vilanova, and Robert Martí. Mam-e: Mammographic synthetic image generation with diffusion models. *Sensors*, 24(7):2076, 2024.

Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Sven Nebelung, Jakob Nikolas Kather, et al. Diffusion probabilistic models beat gans on medical images (2022), 2022.

Dominik Lorenz Patrick Esser Björn Ommer, Robin Rombach, and Andreas Blattmann. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv*, 2112, 2022.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. ACM, 2003.

Dilawar Shah, Mohammad Asmat Ullah Khan, Mohammad Abrar, Farhan Amin, Bader Fahad Alkhamees, and Hussain AlSalman. Enhancing the quality and authenticity of synthetic mammogram images for improved breast cancer detection. *IEEE Access*, 12:12189–12198, 2024.

Wei-Hsiang Shen and Meng-Lin Li. Copy-paste image augmentation with poisson image editing for ultrasound instance segmentation learning. In *2023 IEEE International Ultrasonics Symposium (IUS)*, pages 1–3. IEEE, 2023.

Xin Xing, Yu Nan, Felix Felder, Sean Walsh, and Guang Yang. The beauty or the beast: Which aspect of synthetic medical images deserves our focus? In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 523–528. IEEE, 2023.

Jeong Hee Yoon and Haeryoung Kim. Ct characterization of aggressive macrotrabecular-massive hepatocellular carcinoma: a step forward to personalized medicine, 2021.

Appendix A. Radiologist Survey

Dear Colleague,

You recently participated in a study evaluating radiologists' ability to distinguish between real mammograms and images synthesized by an AI model. We kindly ask you to complete this short survey about your background, reading environment, and your perceptions of the images you reviewed.

1. **Gender**
   □ Male   □ Female

2. **Current professional position (choose one)**
   □ Resident / Registrar
   □ Fellow (Breast Imaging)
   □ General Radiologist
   □ Breast Imaging Radiologist / Consultant
   □ Other (please specify): ________________

3. **Years of independent radiology practice (post-qualification)**
   In years: ______

4. **Primary practice setting (tick all that apply)**
   □ Government / public hospital
   □ University / academic hospital
   □ Private hospital
   □ Private imaging centre / clinic
   □ Other (please specify): ________________

5. **Approximate number of mammography examinations you interpret per week**
   □ ¡ 20   □ 20–49   □ 50–99   □ 100–199   □ 200

6. **Would you describe your clinical breast practice as mainly:**
□ Screening-focused
□ Diagnostic / work-up focused
□ Mixed (screening and diagnostic)

7. **Modalities you routinely interpret (tick all that apply)**
□ 2D digital mammography
□ Digital breast tomosynthesis (DBT)
□ Contrast-enhanced mammography (CEM)
□ Breast ultrasound
□ Breast MRI
□ Other (please specify): ______________________________

8. **Familiarity with AI in breast imaging (1 = Not familiar, 5 = Very familiar)**
□ 1 □ 2 □ 3 □ 4 □ 5

9. **Have you used AI tools in routine breast imaging?**
□ No, never
□ Yes – AI CAD for lesion detection
□ Yes – AI for breast density assessment
□ Yes – AI for image quality/positioning feedback
□ Yes – Other (please specify): ______________________________

10. **Before this study, had you seen AI-generated mammograms?**
□ Yes □ No □ Unsure

11. **Where were you when you completed the image-review session?**
□ Hospital reading room – dedicated breast workstation
□ Hospital reading room – general radiology workstation
□ Office computer
□ Home computer / laptop
□ Tablet device
□ Other: ______________________________

12. **Time of day you completed most of the session**
□ Early morning (06:00–09:59)
□ Late morning (10:00–12:59)
□ Early afternoon (13:00–15:59)
□ Late afternoon (16:00–18:59)
□ Evening (19:00–22:59)
□ Night (23:00–05:59)
□ Multiple sessions

13. **Were you working a normal clinical shift that day?**
□ Yes, during working hours
□ Yes, but after working hours
□ No, it was my day off

14. **Average time spent per image**
□ ¡ 3 sec   □ 3–5 sec   □ 6–10 sec   □ 11–20 sec   □ ¿ 20 sec

15. **Difficulty distinguishing real vs synthetic (1 = Very easy, 5 = Very difficult)**
□ 1   □ 2   □ 3   □ 4   □ 5

16. **How realistic did the synthetic images appear? (1 = Not realistic, 5 = Indistinguishable)**
□ 1   □ 2   □ 3   □ 4   □ 5

17. **Confidence in real/synthetic decisions (1 = Not confident, 5 = Very confident)**
□ 1   □ 2   □ 3   □ 4   □ 5

18. **Did you use specific visual clues?**
□ Yes   □ No
If yes (tick all):
□ Background pattern / noise
□ Skin line or breast contour
□ Pectoral muscle appearance
□ Parenchymal texture
□ Calcifications or masses
□ Artefacts / unnatural features
□ Other: ________________________

19. **Did your decision strategy change over time?**
□ Yes   □ No   □ Unsure
If yes, describe:
____
____

20. **Rate the overall image quality compared to clinical images**
□ Much worse
□ Slightly worse
□ Similar
□ Slightly better
□ Much better

21. **Most challenging aspect (select main factor)**
□ Synthetic images looked very similar to real
□ Parenchymal pattern/density hard to judge
□ Lesion appearance not always typical
□ Variation in quality, contrast, noise
□ Fatigue or concentration
□ Difficulty with interface/monitor
□ Other: ________________________

22. **Key features that helped your decisions**

□ Background texture
□ Skin line / contour
□ Pectoral muscle
□ Parenchymal density
□ Masses / distortions
□ Calcifications
□ Artefacts or unnatural features
□ Overall clinical feel
□ Other: ____________________

23. **How did you assign BI-RADS for synthetic images?**

□ Applied same clinical criteria as real
□ Influenced by possibility of synthetic image
□ BI-RADS felt random/not meaningful
□ Did not assign BI-RADS separately
□ Other: ____________________

24. **Could synthetic mammograms be safely used for:**

	Yes	No	Unsure
a) Resident/fellow teaching	□	□	□
b) Exams / OSCE-style cases	□	□	□
c) AI research datasets	□	□	□
d) Reader studies	□	□	□
e) Routine clinical diagnosis	□	□	□