Feasibility of Self-Operated Breast Ultrasound and AI-Generated Report

吳聲宏*, 施廷翰†, 廖幃萱‡, 黃松山§
*NTHU CS, 113062523; †NTHU EE, 113061606
‡NTHU CS, 110062227; §NTHU IBP, 111006237

***Abstract*—Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide, with early detection playing a critical role in improving survival rates. However, access to diagnostic imaging is often limited, especially in low-resource settings, where reliance on skilled personnel poses a major barrier to widespread screening. To address these challenges, this proposal presents a model for self-operated breast ultrasound system powered by artificial intelligence, designed to automate both image acquisition and diagnostic interpretation. The system leverages advanced image pre-processing, accurate lesion segmentation, and deep learning-based classification to provide reliable diagnostic support with minimal user involvement. Explainable AI techniques such as Grad-CAM are integrated to enhance model transparency and support clinical validation. Experimental evaluation demonstrates that the system achieves diagnostic accuracy comparable to expert radiologists, while maintaining consistency across varying image conditions. These results suggest that autonomous, AI-driven ultrasound could serve as a scalable solution for expanding access to breast cancer screening.**

I. Introduction

Early detection of breast cancer is critical to improving patient outcomes, yet access to diagnostic tools remains limited in many regions. Standard breast ultrasound involves trained staff for image acquisition as well as evaluation, meaning that significant-scale screening is not possible. With the advancement of artificial intelligence in medical imaging, there is growing potential to automate both the scanning and diagnostic processes.

This project investigates the feasibility of a self-operated breast ultrasound system enhanced by AI-generated diagnostic reports. Our approach begins with comprehensive image preprocessing to ensure consistent input quality across varied acquisition conditions. Data augmentation techniques such as horizontal flipping and intensity variation are applied to improve model generalization and robustness. By utilizing robust image preprocessing, precise lesion segmentation, and deep learning-based classification, our system aims to deliver accurate and interpretable results with minimal user intervention. Support from explainable AI techniques like Grad-CAM also helps ensure clinical transparency, with greater potential to increase trust among users and assist medical clinicians in validation.

II. Related Work

Latif et al[1]. decided to tackle the inherent problem of speckles present in ultrasound images that limits the accuracy of medical diagnosis. To address this, a two-stage approach has been proposed that uses convolutional neural networks (CNNs). First, a CNN model is used to remove speckles from ultrasound images to minimize the adverse effects of noise while preserving key features. Following that, another CNN model is proposed for the classification of ultrasound images into benign and malignant classes. The proposed models were tested on the Mendeley Breast Ultrasound dataset, and experimental results indicate that accuracy is significantly improved by using CNN for despeckling as a preprocessing step.

Jabeen et al.[2] introduce a framework for automated breast cancer classification from ultrasound images to address limitations such as similarity between lesions and irrelevant features. The method involves data augmentation and utilizes a DarkNet-53 model with transfer learning for feature extraction. Subsequently, reformed differential evolution and reformed gray wolf optimization algorithms are employed for feature selection, and the best features are fused using a novel probability-based serial approach before being classified using different machine learning algorithms. Although both methods achieve relatively high accuracy, they lack mechanisms to assist doctors in resolving conflicting diagnoses.

Latha et al.[3] proposed an approach which involves fine-tuning EfficientNet-B7 on the BUSI dataset, utilized data augmentation to address class imbalance and enhance model robustness. Furthermore, the integration of Explainable AI (XAI) techniques, specifically Grad-CAM, provides visual insights in order to identify the important features, enhancing interpretability and trust.

Qasrawi et al.[4] introduce a hybrid approach for breast cancer detection. To address challenges in image quality, the proposed method first employs Contrast Limited Adaptive Histogram Equalization for image enhancement. Subsequently, the enhanced images are classified using Ensemble Deep Random Vector Functional Link Neural Network. Furthermore, for lesion segmentation, the study initially employed YOLOv5 to highlight a general area of tumor, followed by MedSAM for a more precise segmantation. This method can create a more accurate sectionalization and helps ensure that MedSAM does not select an incorrect region. The interpretability of the model is also supported by Grad-CAM analysis, which highlights the image regions influencing the model's decisions.

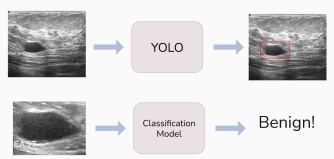


Fig. 1. Flowchart of our method

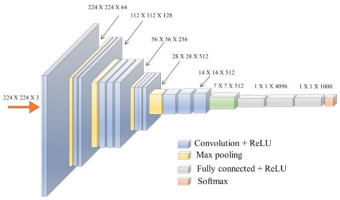


Fig. 2. Structure of VGG16

III. Methods

Our method consists of two main components. The first component utilizes a YOLO-based object detection model to identify tumors in breast ultrasound images. If a tumor is detected, the model localizes it by generating a bounding box and cropping the region of interest around the tumor. This ROI is then passed to the second component, a modified VGG16 convolutional neural network, which classifies the tumor as either benign or malignant. To enhance the interpretability of the classification results, we apply Grad-CAM (Gradient-weighted Class Activation Mapping) to the output of the VGG16 model, allowing us to visualize the regions of the input image that contributed most to the model's decision. The whole process is shown in Fig. 1.

A. YOLOv8

For tumor detection, we employed YOLOv8, the object detection model developed by Ultralytics. YOLOv8 adopts an anchor-free and modular design, offering improved accuracy and speed, which makes it well-suited for real-time tasks such as medical image analysis. The model was trained on the Breast Ultrasound Images[5] (BUSI) dataset, which includes images labeled as benign, malignant, or normal. To improve model robustness and generalization, we applied various data augmentation techniques during training. Additionally, normal images were included in the dataset to help the model learn to distinguish images with tumor or normal. To enhance image quality and emphasize structural features, we applied Contrast Limited Adaptive Histogram Equalization[6] (CLAHE) to increase the contrast of ultrasound images. We conducted experiments to assess the impact of these preprocessing steps on detection performance.

B. VGG16

For tumor classification, we utilized a pretrained VGG16[7] model. Since the original VGG16 architecture was designed for multi-class classification over 1000 ImageNet classes, we modified the final fully connected layer to a single linear layer followed by a sigmoid activation function, enabling binary classification between benign and malignant tumors. Since VGG16 expects input images of size 224×224×3, as shown in Fig. 2, and the cropped tumor pictures vary in size, we first padded the images to form square shapes while preserving aspect ratio. These padded images were then resized to match the required input dimensions. For finetuning part, we froze all convolutional layers and trained only the fully connected layers, allowing the model to adapt to the tumor classification while retaining the benefits of pretrained feature extraction.

C. Grad-CAM

To improve interpretability and provide visual explanations for the model's predictions, we integrated Gradient-weighted Class Activation Mapping[8] (Grad-CAM) with the VGG16 classifier. Grad-CAM generates heatmaps that highlight the regions in the input image that most strongly influence the model's decision. This helps validate that the model is focusing on relevant tumor areas, thereby enhancing transparency and trust in the diagnostic process.

D. Preprocessing Method

We employed a classical and powerful denoising technique called Block-Matching and 3D Filtering (BM3D) that used in an effort to reduce speckle and Gaussian noises affecting the ultrasonic images. Our implementation of the preprocessing pipeline involves performing just the first step within the BM3D algorithm, also called the basic estimate. The algorithm groups together similar image patches, which are then transformed and filtered in a collaborative way to suppress noise levels without sacrificing anatomical details.

1) Noise Simulation: Gaussian noise with a standard deviation σ = 20 was artificially induced onto the original grayscale ultrasound images to simulate and test the denoising methods' efficacy. The noisy image I_n was:

$$I_n = I + \mathcal{N}(0, \sigma)$$

where I is the original image and $\mathcal{N}(0, \sigma)$ is zero-mean Gaussian noise.

2) Block-wise DCT Transformation: Overlapping blocks of size 8×8 were extracted from the image. A 2D DCT was applied to each extracted block mainly for reasons of computational efficiency and energy compaction. The effect of such transformation is to represent the image data more sparsely and thus allow thresholding to be more effective during the collaborative divergence.

3) Block-matching and Grouping: A reference block was selected per stride of 5 pixels, and for each reference block, a search was conducted in a 30×30 window for similar blocks based on their DCT-domain Euclidean distances. The top 16 blocks that exhibited a distance below a certain threshold were grouped together along the third dimension to form a 3D array.

4) Collaborative 3D Filtering: These grouped blocks underwent the collaborative filter, including the following steps: - Performing 1D DCT along the third dimension of the 3D group,

- Performing 1D DCT along the third dimension of the 3D group,
- Performing hard-thresholding to suppress noise coefficients with a threshold proportional to the noise level,
- Performing an inverse 1D DCT to reconstruct the filtered group.

5) Aggregation and Reconstruction: Each filtered block was projected back onto its original position. Because blocks may overlap, these contributions were weighted with a Kaiser window, and the final pixel values were computed by weighted averaging. The output of this step corresponds to the basic estimate of the denoised image.

6) Performance Metric: The Peak Signal-to-Noise Ratio (PSNR) was computed between the original and denoised (basic estimate) images to select the restoration quality:

$$\mathrm{PSNR} = 20 \log \left(\frac{255}{\sqrt{\mathrm{MSE}}} \right)$$

where MSE stands for the mean squared error between the original and basic estimate images.

IV. Experiment Results

A. Dataset and Evaluation Metrics

We conducted experiments using the BUSI breast ultrasound dataset, applying YOLOv8 for lesion detection and MedSAM for segmentation. The segmentation performance was measured using the Intersection over Union (IoU), while object detection was evaluated using precision, recall, mean Average Precision at IoU threshold 0.5 (mAP@0.5), and mAP across thresholds from 0.5 to 0.95 (mAP@0.5:0.95).

B. Comparison of Preprocessing Strategies

We evaluated four different preprocessing strategies, including CLAHE (Contrast Limited Adaptive Histogram Equalization) and data augmentation. All models were trained with an image size of 256, a batch size of 16, and for 50 epochs. Table I shows the results.

Observation: Applying augmentation alone achieved the best precision and mAP@0.5:0.95, suggesting improved generalization. CLAHE alone did not improve performance significantly.

C. Effect of Including Normal Samples

To evaluate the benefit of incorporating normal (non-lesion) samples, we retrained all configurations with normal images added to the dataset. Results are shown in Table II.

Observation: Including normal samples led to a significant performance boost. The combination of CLAHE and augmentation resulted in the best mAP@0.5:0.95, indicating a more robust detection model for distinguishing for distinguishing between lesion and normal images.

D. Single-Site Evaluation

Single-Dataset Comparison: We evaluated classification performance on the BUSI dataset by applying YOLOv8 for lesion cropping, with and without an additional denoising step, to assess its impact on classification accuracy. The results are summarized in Table III.

Observation: Adding a denoising step before training significantly improved classification performance on the BUSI dataset. Specifically, it boosted accuracy from 0.938 to 0.963 and F1 score from 0.898 to 0.936. Both AUROC and specificity also increased slightly, indicating better discrimination and fewer false positives. This suggests that denoising enhances lesion feature clarity and model generalization on in-domain data.

Dataset-wise Evaluation: We compared classifier performance across different datasets—BUSI, BUS-UCLM[9], and QAMEBI[10]—with and without CLAHE preprocessing, to study how preprocessing and domain differences affect performance. The results are summarized in Table IV.

Observation: The effect of CLAHE was inconsistent across datasets. For BUS-UCLM, CLAHE reduced classification performance across all metrics — e.g., F1 dropped from 0.571 to 0.200. In contrast, applying CLAHE to QAMBEI led to slightly lower AUROC (from 0.851 to 0.646) but higher precision (from 0.692 to 0.857), suggesting that CLAHE might bias the model toward conservative predictions. These results indicate that CLAHE is not universally beneficial and its utility may depend on dataset characteristics.

1) Visualization of Model Performance: To better interpret the classifier's behavior, we visualize the confusion matrix for the BUSI dataset with augmentation and denoising (Figure 3), as well as the ROC curves for all three datasets—BUSI, BUS-UCLM, and QAMEBI (Figure 4). These plots illustrate the trade-off between true positive and false positive rates and provide insight into the class-wise prediction distribution.

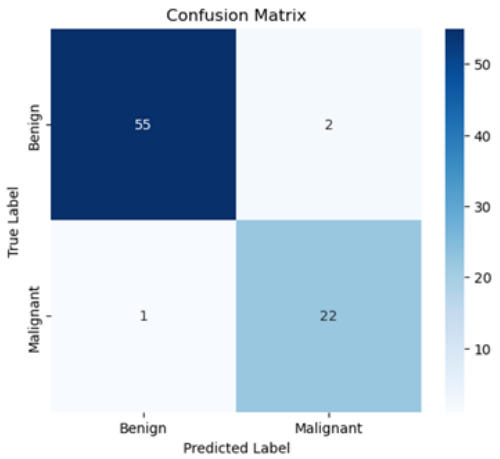


Fig. 3. Confusion matrix heatmap for BUSI classification (BUSI + Denoise + Aug).

TABLE I
PERFORMANCE OF YOLOV8 WITH DIFFERENT PREPROCESSING METHODS (WITHOUT NORMAL IMAGES).

Method	Precision	Recall	mAP@0.5	mAP@0.5:0.95
BUSI (raw)	0.618	0.667	0.703	0.371
BUSI + CLAHE	0.607	0.714	0.659	0.335
BUSI + Augmentation	**0.793**	0.619	**0.717**	0.393
BUSI + CLAHE + Aug	0.774	0.652	0.712	0.355

TABLE II
PERFORMANCE OF YOLOV8 WITH NORMAL IMAGES INCLUDED.

Method	Precision	Recall	mAP@0.5	mAP@0.5:0.95
BUSI (raw)	0.782	0.641	0.749	0.410
BUSI + CLAHE	0.607	0.714	0.659	0.335
BUSI + Augmentation	**0.814**	**0.787**	**0.832**	0.416
BUSI + CLAHE + Aug	0.807	0.718	**0.832**	**0.462**

TABLE III
BUSI VGG16 CLASSIFICATION PERFORMANCE UNDER DIFFERENT PREPROCESSING STRATEGIES.

Method	Accuracy	AUROC	F1	Precision	Recall	Specificity
BUSI + Aug	0.938	0.972	0.898	0.846	0.957	0.930
BUSI + Denoise + Aug	**0.963**	**0.975**	**0.936**	**0.917**	**0.957**	**0.965**

TABLE IV
CROSS-DATASET CLASSIFICATION PERFORMANCE USING VGG16 TRAINED ON DIFFERENT LESION CROPS. CLAHE PREPROCESSING SETTINGS ARE INDICATED.

Method	Accuracy	AUROC	F1	Precision	Recall	Specificity
BUS-UCLM (No CLAHE)	0.875	0.725	0.571	0.667	0.500	0.950
BUS-UCLM + CLAHE	0.652	0.647	0.200	0.250	0.167	0.824
QAMBEI (No CLAHE)	0.760	0.851	0.750	0.692	0.819	0.714
QAMBEI + CLAHE	0.563	0.646	0.632	0.857	0.500	0.750

E. Cross-Site Generalization

To test the robustness of our classifier across domains, we fine-tuned the model on two datasets and evaluated it on a third. Three such experiments were conducted, each leaving one dataset as unseen test data. The results are summarized in Table V.

Observation: When evaluating generalizability, the best performance was observed when training on BUSI + BUS-UCLM and testing on QAMBEI (AUROC = 0.850, F1 = 0.725). In contrast, testing on BUS-UCLM after training on BUSI + QAMBEI led to the weakest F1 (0.580), despite a relatively high specificity (0.910). This suggests that BUS-UCLM may have more domain-specific characteristics that generalize less easily. Overall, the model struggles more with recall than specificity across domains, indicating that malignant lesions are harder to detect consistently when testing across unseen datasets.

F. End-to-End Pipeline Evaluation

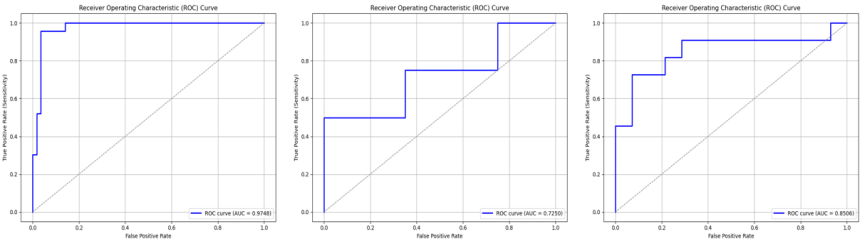
While the classification results presented earlier (Table III) reflect performance on successfully cropped lesion regions, they do not account for the entire BUSI dataset, as YOLOv8 failed to detect a bounding box in some images. To better reflect real-world deployment performance, we compute end-to-end pipeline-level metrics that integrate both the detection and classification stages.

From the original BUSI dataset, there are 437 benign and 210 malignant images. YOLOv8 successfully produced bounding boxes for 343 benign and 181 malignant cases, resulting in detection rates of 78.5% and 86.2% respectively. For the remaining 123 undetected images (94 benign and 29 malignant), no lesion region was available for classification. We assume that for these cases, the system defaults to predicting ”benign”—a reasonable fallback in low-resource or uncertain settings.

Based on the classification model's per-class performance (recall = 0.957 for malignant, specificity = 0.965 for benign), we combine detection coverage and classification accuracy to compute the overall system metrics shown in Table VI. These reflect true performance across the full BUSI dataset, present in Table VI.

G. Model Explainability via Grad-CAM

To enhance interpretability and build clinical trust, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the decision-making process of our VGG16



BUSI BUS-UCLM QAMBEI

Fig. 4. ROC curves comparing model performance across BUSI, BUS-UCLM, and QAMEBI datasets.

TABLE V
CROSS-SITE CLASSIFICATION RESULTS: MODELS FINE-TUNED ON TWO DATASETS AND TESTED ON THE REMAINING ONE.

Train Sites	Test sites	Accuracy	AUROC	F1	Precision	Recall	Specificity
BUSI + BUS-UCLM	QAMBEI	0.760	0.850	0.725	0.877	0.617	0.909
BUSI + QAMBEI	BUS-UCLM	0.652	0.820	0.580	0.645	0.526	0.910
BUS-UCLM + QAMBEI	BUSI	0.769	0.817	0.649	0.683	0.619	0.848

TABLE VI
PIPELINE-LEVEL PERFORMANCE ON THE ENTIRE BUSI DATASET, ACCOUNTING FOR BOTH YOLOV8 DETECTION SUCCESS AND VGG16 CLASSIFICATION.

Metric	Value
Accuracy	0.924
Recall (Malignant)	0.824
Specificity (Benign)	0.973
F1 Score (Malignant)	0.876

classifier. Grad-CAM generates heatmaps that highlight regions in the input image that contributed most to the model's classification decision. Figure 9 shows representative Grad-CAM results for both benign and malignant cases. These visualizations suggest that the model consistently attends to relevant tumor regions when making predictions. In malignant cases, high-activation areas often surround the lesion's irregular boundaries, while benign lesions tend to produce more localized and uniform attention. Multi-layer Grad-CAM (aggregating multiple convolutional layers) further reveals both texture and boundary-related features contributing to classification.

V. DISCUSSION

Across all three datasets (BUSI, BUS-UCLM, and QAMEBI), class imbalance is a recurring challenge. This imbalance skews classifier learning, often leading to high specificity but lower sensitivity (recall), especially for malignant lesions. For instance, as shown in Table IV, models trained on heavily imbalanced datasets (e.g., BUS-UCLM or QAMBEI) exhibit poor generalization when tested across domains, with some combinations producing recall or precision values as low as 0.17 or 0.25. Such disparity suggests that the model may become overly confident in predicting the dominant class. To mitigate this, future work should consider class-balancing strategies like oversampling, focal loss, or cost-sensitive training to improve model robustness and fairness in diagnosis.

Our cross-site experiments (Table V) show that domain shifts between datasets significantly affect model performance. For instance, recall dropped substantially when evaluating on BUS-UCLM after training on BUSI + QAMEBI. This highlights the need for domain adaptation techniques or data harmonization to build more robust clinical models.

Explainability is critical in medical AI applications. By visualizing model attention using Grad-CAM, we demonstrated that our classifier focuses on clinically meaningful lesion regions, supporting the reliability of the system's predictions.

As shown in Table VI, our end-to-end system achieves an accuracy of 92.4% on the full BUSI dataset, including cases where YOLOv8 failed to detect a lesion. While classifier-level recall was 95.7%, pipeline-level recall dropped to 82.4% due to missed detections. This suggests that improving detection robustness—especially for malignant cases—remains a key area for future work.

Finally, integrating clinical metadata (e.g., age, family history) may enhance prediction reliability. Moreover, since the BUSI dataset does not include multiple imaging views per patient, future systems should support multi-view analysis for improved robustness.

VI. Conclusion

This study demonstrates the potential of a self-operated breast ultrasound system powered by artificial intelligence to facilitate breast tumor detection. By combining YOLOv8-based lesion detection with a VGG16 classifier enhanced by Grad-CAM explainability, the system achieved high accuracy and interpretability. Preprocessing strategies, including denoising and augmentation, significantly improved model performance, while cross-dataset evaluations revealed the challenges of domain generalization. The end-to-end pipeline achieved an accuracy of 92.4% on the BUSI dataset. Our findings underscore the promise of autonomous AI-driven ultrasound as a scalable solution for breast cancer screening and diagnosis. Code is available here: https://github.com/Flora-Liao/MedAI_final_project

VII. Contribution

吳聲宏: 26% 施廷翰: 26% 廖幃萱: 29% 黃松山: 19%

References

[1] G. Latif, M. O. Butt, F. Yousif Al Anezi, and J. Alghazo, "Ultrasound image despeckling and detection of breast cancer using deep cnn," in *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pp. 1–5, 2020.

[2] K. Jabeen, M. A. Khan, M. Alhaisoni, U. Tariq, Y. Zhang, A. Hamza, A. Mickus, and R. Damaševičius, "Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion," *Sensors*, vol. 22, p. 807, Jan. 2022.

[3] M. Latha, P. S. Kumar, R. R. Chandrika, T. R. Mahesh, V. V. Kumar, and S. Guluwadi, "Revolutionizing breast ultrasound diagnostics with efficientnet-b7 and explainable ai," *BMC Medical Imaging*, vol. 24, no. 1, p. 230, 2024. Published: September 2, 2024.

[4] R. Qasrawi, O. Daraghmeh, S. Thwib, I. Qdaih, G. Issa, S. V. Polo, H. Owienah, D. A. Al-Halawa, and S. Atari, "Advancing breast cancer detection in ultrasound images using a novel hybrid ensemble deep learning model," *Intelligence-Based Medicine*, vol. 11, p. 100222, 2025.

[5] A. Shan, "Breast ultrasound images dataset (busi)." https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset, 2020. Accessed: 2025-04-13.

[6] A. Mishra, "Contrast limited adaptive histogram equalization (clahe) approach for enhancement of the microstructures of friction stir welded joints," 2021.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, p. 336–359, Oct. 2019.

[9] D. García-González, D. Naranjo-Hernández, J. Reina-Tosina, and E. J. Gómez, "Bus-uclm: A dataset for breast ultrasound image segmentation." https://data.mendeley.com/datasets/7fvgj4jsp7/1, 2020. Accessed: 2025-04-13.

[10] Q. Lab, "Breast ultrasound images database." https://qamebi.com/breast-ultrasound-images-database/, 2023. Accessed: 2025-04-13.

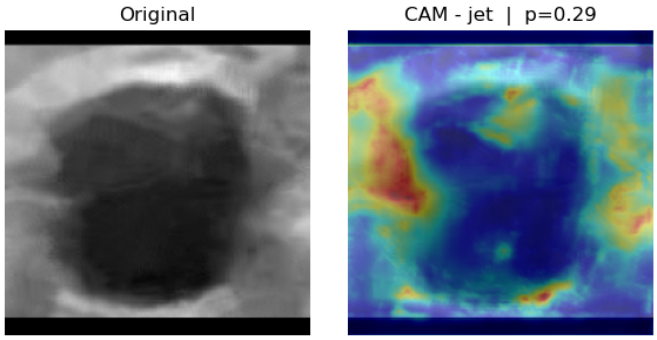


Fig. 5. Benign case 1 (Prediction: 0.29)

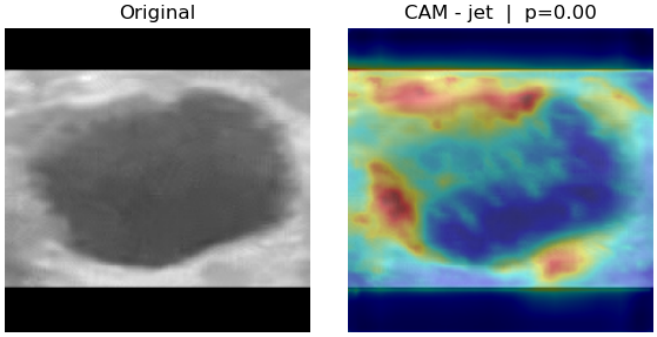


Fig. 6. Benign case 2 (Prediction: 0.00)

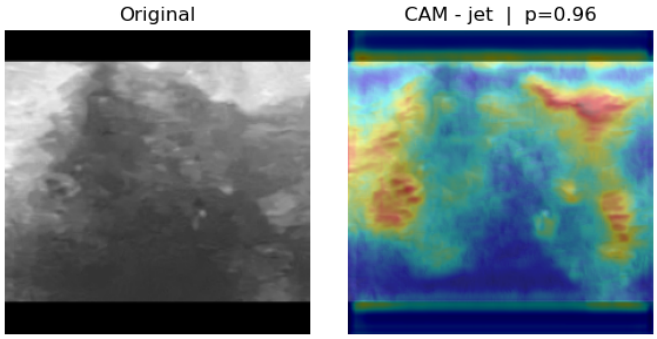


Fig. 7. Malignant case 1 (Prediction: 0.96)

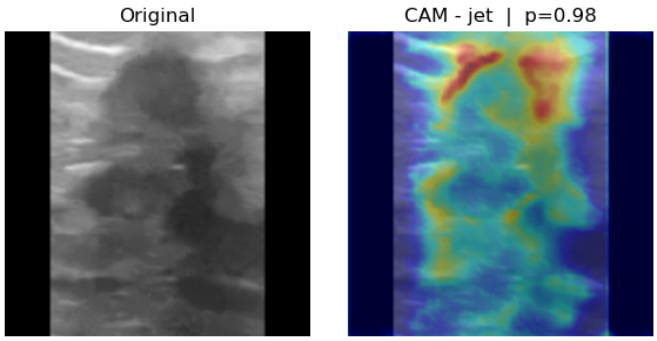


Fig. 8. Malignant case 2 (Prediction: 0.98)

Fig. 9. Grad-CAM heatmaps overlaid on breast ultrasound images. Red areas indicate regions the model focused on for prediction. Benign and malignant examples show distinct activation patterns.