

ORIGINAL RESEARCH ARTICLE

Uncertainty estimation of breast cancer classification using deep learning methods with explainable artificial intelligence

Arthi Rengaraj¹, Manoj Kumar Devendran^{1*}, Mahesh Kumar Nadesan¹, Praveenkumar Babu¹, and Vijayan Thiruvengadam¹

Department of Electronics and Communication Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu, India

Abstract

Globally, breast cancer-related mortality among women requires early and precise diagnostic support to improve patient outcomes. The existing research work discusses using Machine learning and deep learning models for precise diagnosis with a combination of classified outcomes using various datasets. The proposed work focuses on uncertainty estimation of breast cancer using a deep learning framework with Explainable Artificial Intelligence (XAI) with the Curated Breast Imaging Subset of Digital Database for Screening Mammography Dataset. The novelty of the work is that it classifies breast cancer as benign, benign-without-callback, and malignant using a convolutional neural network (CNN) and a vision transformer (ViT), uncertainty estimation using Monte Carlo dropout, and gradient-weighted class activation mapping for visual localization of discriminative regions and SHapley Additive exPlanations for pixel-level feature importance to increase transparency as well as reliability. The consistency of the XAI maps produced by CNN and ViT has been examined quantitatively using the structural similarity index measure. The results show that ViT has more dispersed attention patterns, while CNN focuses more on isolated lesion sites. As a result, the proposed work supports clinical decision-making for precise breast cancer diagnosis, improves model interpretability, and supports uncertainty forecasting.

***Corresponding author:**Manoj Kumar D
(manojkud1@srmist.edu.in)

Citation: Rengaraj A, Devendran MK, Nadesan MK, Babu P, Thiruvengadam V. Uncertainty estimation of breast cancer classification using deep learning methods with explainable artificial intelligence. *Artif Intell Health*. doi: 10.36922/AIH026280083

Received: July 7, 2026**Revised:** August 8, 2026**Accepted:** September 3, 2026**Published online:** September 21, 2026**Copyright:** © 2026 Author(s).

This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Breast cancer; Explainable artificial intelligence; Convolutional neural network; Vision transformer; SHapley Additive exPlanations; Structural similarity index measure

1. Introduction

Breast cancer is one of the most common and deadly diseases in women throughout the world. Timely detection and precise diagnosis are vital to enhance treatment outcomes and survival rates. Mammography is a widely accepted method of imaging for the early detection and diagnosis of breast problems. However, manual interpretation of mammography images can be time-consuming, prone to inter-observer variability, and heavily reliant on radiologists' expertise. Recent advances in deep learning have enabled automated feature extraction and strong classification performance, significantly improving medical image analysis. Convolutional neural networks (CNNs) are deep

learning models that have achieved strong results in image classification tasks because they can effectively extract local spatial information. More recently, the vision transformer (ViT) has been proposed as an alternative design that learns global contextual correlations by self-attention mechanisms. ViTs have shown promising results in large-scale computer vision applications, but their efficacy on tiny medical datasets remains unclear.

While deep learning algorithms have shown promising performance, the lack of interpretability, referred to as the “black-box” problem, is a major challenge in medical artificial intelligence (AI) systems. Clinicians need transparency and confidence in model predictions to build trust and reliability in critical healthcare applications. Explainable AI (XAI) techniques offer ways to interpret model behavior and the logic behind classification decisions. While methods like gradient-weighted class activation mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) can visually identify important regions in the image and the contribution of features, respectively, uncertainty estimation methods like Monte Carlo dropout (MCD) can provide confidence in the prediction. The proposed work focuses on an explainable deep learning architecture for breast cancer classification using the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) dataset. A CNN and a ViT are trained and evaluated using a dataset of mammography images from three classes: benign, benign-without-callback, and malignant. The uncertainty estimations are done using MCD, Grad-CAM for visualizing discriminative regions, SHAP for feature importance, and the structural similarity index (SSIM) for consistency to enhance interpretability and assess the various attention mechanisms of CNNs and ViTs. In addition to effectively classifying breast cancer, the proposed architecture aims to improve transparency and confidence in deep learning-based medical detection systems.

Artificial intelligence can revolutionize many facets of the healthcare industry, such as administration, diagnosis, medical prediction, and decision-making. Thus, “DXAIB” eases diagnosis identification and provides clear explanations of the rationale behind a given prediction, enhancing transparency and trust in the AI system’s decision-making process. Compared to existing state-of-the-art approaches¹, the “DXAIB” approach outperformed in both explainability and prediction results. The investigation showed that deep learning techniques outperformed machine learning in terms of accuracy, increasing it by almost 2%. The application of XAI techniques in breast cancer diagnosis has been discussed,

and the benefits and disadvantages have been explored.² Among the models considered, ResNet50 had the highest accuracy (99.24%), with exceptional area under the curve (AUC), precision, and *F*-score. These conclusions help medical professionals diagnose patients with confidence and knowledge and increase model transparency, aiding anomaly identification by using XAI to highlight significant regions in mammograms. The techniques improve the diagnosis by balancing performance and interpretability, supporting radiologists in high-prevalence areas, and improving patient outcomes through useful AI-based forecasts.³ Random forest scored better than the other algorithms, with an *F1*-score of 84%. By combining the benefits of multiple models, the stacked ensemble model produced an *F1*-score performance of 83%. XAI methods, such as SHAP, local interpretable model-agnostic explanations (LIME), Explain Me I’m Five, Anchor, and Quantum Lattice, were used to interpret the findings. Interpretable algorithms can improve clinical decision-making, lower diagnostic errors, and predict breast cancer in the medical industry.⁴ The model uses XAI techniques like class activation mapping (CAM) and Saliency Maps, along with their integrated visualization, to increase model transparency and clinical trust. These techniques highlight the most relevant areas that contribute to cancer diagnosis, giving pathologists and physicians valuable insights. The method has certain shortcomings even though the model performs well in controlled environments. We did not develop a practical clinical device in the current study. Future studies should focus on combining these techniques and using the model in clinical settings to more accurately diagnose breast cancer.⁵ In comparison, baseline models such as Visual Geometry Group (VGG)19, SE-ResNet152, and ResNet152V2 yielded lower accuracy of 87.54%, 87.50%, and 85.24%, respectively, while transfer learning algorithms achieved a high accuracy of 87.37%. To increase therapeutic trust, this study utilized SHAP and Grad-CAM. These strategies highlighted locations in mammograms relevant to diagnosis. This improved prediction transparency. Because LBNet shows strong potential as an accurate, effective, and interpretable choice for breast cancer screening, further study could examine its applicability to multi-view mammography and real-time clinical deployment.⁶ To enhance the model’s performance, these characteristics are fed into the classification U-Net model that is merged with gated recurrent unit networks. Lastly, the classification result is enlarged, and the model’s predictions are visually explained using XAI. Using the 100X BreakHis dataset, the proposed Patho-Net model⁷ classifies breast cancer with 98.90% accuracy. The XAI-integrated MammXAI dashboard substantially enhances the multimodal framework for breast cancer classification

presented in this study. By providing healthcare practitioners with clear visualizations and interpretations of the algorithm's decision-making processes through techniques such as Grad-CAM, integrated gradients, and other XAI methods, this research aims to accelerate breast image assessment, reduce diagnostic errors, and improve overall clinical efficiency.⁸ To determine the correctness of quantified uncertainties, we present and review new algorithms for generating synthetic out-of-distribution data. The results demonstrated that ensemble-based uncertainty quantification methods often yielded more accurate uncertainty estimates and performed better in terms of sensitivity. In contrast, stochastic methods achieved improved calibration; in particular, Bayesian Neural Networks showed strong defense against adversarial attacks, which may enhance classifiers' capacity for generalization. Overall, making model predictions more transparent and providing reliable confidence measures and uncertainty estimates promote safer clinical implementation.⁹ The model has seen some notable successes, such as CNNs achieving 98.5% accuracy in mammography and ViTs achieving 96% accuracy in histopathological analysis. XAI techniques such as SHAP, LIME, and Grad-CAM are also highlighted to ensure transparency, accountability, and trust in clinical decision-making.¹⁰ The Dempster-Shafer-based classifier offered a good balance between interpretability and accuracy with well-defined decision criteria and respectable performance metrics. XAI approaches improve interpretability without appreciably sacrificing performance when applied to predictive models.

The study highlights the importance of balancing explainability and accuracy in clinical AI applications and encourages broader adoption of explainable models¹¹ to maintain ethical responsibility, transparency, and trust in healthcare. The recommended approach uses the Temporal Convolutional Network algorithm with the Synthetic Minority Over-sampling Technique, achieving the best accuracy for breast cancer prediction. Using the public dataset, the experimental results for accuracy, sensitivity, specificity, precision, F1-score, and AUC were 99.44%, 100.0%, 99.01%, 98.75%, 99.37%, and 99.89%, respectively. On the private dataset, the experimental findings for accuracy, sensitivity, specificity, precision, F1-score, and AUC were 97.33%, 93.33%, 100%, 100%, 96.55%, and 99.48%, respectively. In contrast, the Temporal Convolutional Network algorithm used the pooled datasets to reach 100% on every performance criterion.¹² Experimental results show that the TECNN model, which consists of AlexNet, VGG16, and EfficientNetB0, attained the highest accuracy of 98.75% on the AISSLab-v2 dataset. Applying XAI approaches greatly enhanced the

model's interpretability, enabling doctors to understand and evaluate its decision-making process. Clinical examination highlighted the model's practical relevance in a diagnostic setting, showing that the XAI outputs correlated well with expert judgments.¹³ Furthermore, tree-based ensemble machine learning models were the main explanation provided by the SHAP model. The most frequent explanation is that SHAP is popular and helpful for explaining any model prediction since it is model-agnostic. It also works well with performant models, such as tree-based models, and is comparatively simple to implement. XAI improves care quality and outcomes by making AI-enabled health systems and medical equipment more transparent, interpretable, equitable, and reliable.¹⁴ The model outperformed the state-of-the-art techniques (average C-index = 0.650, ranging from 0.619 to 0.677) when tested on seven public datasets of breast cancer (average C-index = 0.691). Significantly, UISNet identified 20 genes linked to breast cancer; 11 have been identified in earlier research, and the remaining genes were newly identified in this investigation.¹⁵ The ability of three suggested models to measure classification reliability using new performance measures used to evaluate predicted uncertainty is compared. Our study's quantitative and qualitative analyses suggest that models exhibit significant uncertainty in misclassifications, which is crucial for estimating the likelihood of medical diagnosis hazards.¹⁶ The results show that the proposed model achieves 99.95% accuracy on the CBIS-DDSM mammography dataset, 98.40% accuracy on the Biomedical Ultrasound Multimodal dataset, and 91.34% accuracy on the Breast Ultrasound Image (BUSI) dataset, respectively. Finally, because we observed high uncertainty in the proposed model's misclassifications, we used uncertainty values to identify appropriate subsets for further investigation to improve performance.¹⁷ Moreover, the proposed method can provide prediction explainability by automatically highlighting the suspicious area on the image using class activation mapping. We assessed the robustness of cancer prediction and localization by comparing results from two different class activation mapping algorithms. We evaluated the proposed approach on a dataset of 9,016 images and achieved 93.5% accuracy, demonstrating the network's effectiveness for breast cancer detection and localization.¹⁸ The model was evaluated on two benchmark datasets: the BUSI dataset, having 1,578 pictures, and Breakhis-400x (B-400x), consisting of 1,820 pictures. To make model training and evaluation more efficient, we systematically split the datasets into training (70%), test (20%), and validation (10%) sets. This resulted in 93.97% accuracy for the B-400x dataset (benign and malignant classes) and 89.87% accuracy for the BUSI dataset (benign, malignant, and normal classes) for breast

cancer detection.¹⁹ The aim is to find the model with the highest accuracy, then use three popular XAI techniques—LIME, SHAP, and Saliency Maps—to identify the best-performing model. The study shows that LIME is the most accurate and clinically relevant XAI method for breast cancer classification from downsampled histopathology images, with VGG19 performing best at 92.59% accuracy. These results, confirmed by medical experts, improve the interpretability and reliability of deep learning models for clinical use and encourage their responsible application in medical practice.²⁰

Histopathology images have been used to estimate uncertainty²¹ in breast cancer using Bayesian CNNs and transfer-learning CNNs. These uncertainty projects have been performed in a low-dimensional space through a nonlinear dimensionality reduction technique. Patient observations with different pathologies and images from different distributions²² are used to train models to estimate uncertainty using the feature density method. This mathematical model uses supervised learning to estimate breast cancer classification using a histogram or probability density function. Gaussian Processes and Deep Gaussian Processes have been used for crowdsourcing²³ to classify the breast tissue as an annotator for noisy regions. The results aimed to classify the breast tissue using scalable techniques. Deep neural networks and uncertainty quantification have been used with a heteroscedastic binary cross-entropy model²⁴ for uncertainty estimation. The results guided radiologists in assessing the tissue for decision-making.

Sample-based methods such as MCD and deep learning methods have been used to classify and estimate uncertainty as predictive entropy or predictive variance.²⁵ In our study, we analyzed various methods for classifying disease predictions based on the disease, sensor, and architecture used. Three models addressed high-level uncertainty in classifying correct predictions²⁶, such as uncertainty quantification, MCD, and ensemble MCD, for breast cancer based on their cognitive features. The findings accurately predicted high uncertainty and misclassifications. Artificial neural networks have been used to discriminate²⁷ whether breast cancer is malignant or benign using mammograms, using the Wisconsin State Cancer Reporting System to assess prediction accuracy. Automated segmentation using MCD and ensemble methods has addressed epistemic uncertainty in breast tumor characterization.²⁸ These methods are evaluated and calibrated for uncertainty estimation to support clinicians.

Combining Bayesian uncertainty with selective prediction²⁹ to classify breast ultrasound images improves reliability. The uncertainty estimations of

breast cancer have been discussed using the UCTH Breast Cancer Clinical Dataset with SHAP.³⁰ This study uses a mix of lightweight gradient boosters, random forests, and gradient boosters for classification. The Bayesian network model³¹ was used to classify cancer type; although the classification was accurate, it lacked confidence in classifying subtypes from transcriptomic data. The model supported informed decision-making for disease-type classification. Axillary lymph node metastasis for breast tissue classification³² has been done using demographic, radiological, histopathological, and molecular characteristics. These comparisons exhibited risk awareness and uncertainty predictions to enable transparent interpretation. Margin detection in cancer surgery³³ has also been suggested to predict uncertainty using a Bayesian neural network and two baseline models. This study showed the variance in tissue signatures. The family history method was adopted to predict the risk of breast and ovarian cancer and predicted the uncertainty level in three-tier pedigrees under certain conditions.³⁴ This approach integrates both clinical and uncertainty estimations. The deep learning-based diagnostic model³⁵ estimates the presence of tumor by pathological subtypes of renal cell carcinoma based on computed tomography images for the uncertainty estimations. The classification of imprecise data for breast cancer has been done by hierarchical methods³⁶ for multivariate interval-valued data towards predicting benign and malignant. Uncertainty estimates using adjuvant HER2-targeted therapy have been estimated using deep learning and MCD to predict breast cancer tissue.³⁷ The Gaussian mixture models³⁸ have been used for probable breast cancer classification towards mRNA expression data for 300 clinical samples of breast cancer and healthy tissue. The Dutch Breast Cancer Screening Program³⁹ used mammogram images to rule out breast cancer tissue and uncertainty estimation and misclassifications using a commercial lesion detection algorithm. The deep learning predictions used integrated model methods and a model-agnostic approach to detect mispredictions in underrepresented cancer subtypes.⁴⁰

2. Data and methods

An explainable deep learning framework for breast cancer to predict uncertainty estimation is shown in [Figure 1](#). The proposed methodology integrated deep learning models such as CNN and ViT with XAI for uncertainty estimation to quantify the model's confidence in its predictions, Grad-CAM for saliency mapping, and SHAP for feature importance, using mammogram images from the CBIS-DDSM Dataset. The dataset mammogram images were preprocessed and split for training, validation, and testing. We used the processed mammogram images

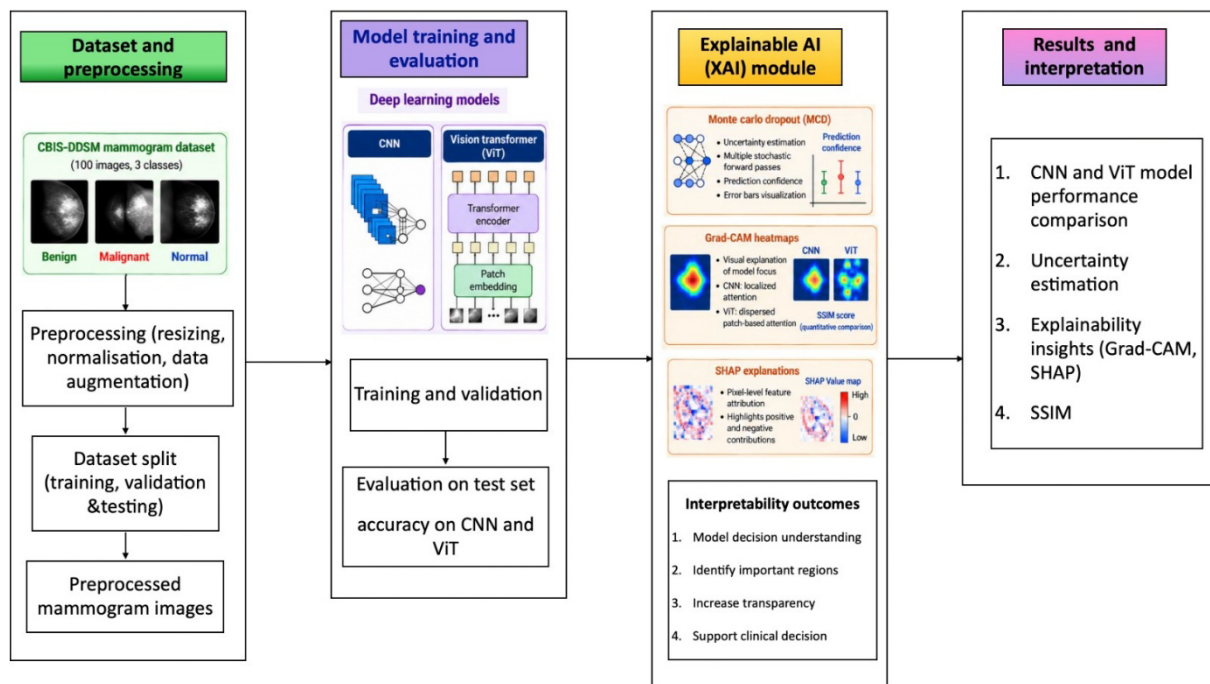


Figure 1. Architecture diagram of uncertainty estimations using deep learning models

to train and evaluate the CNN and ViT models. The CNN used preprocessed mammogram images as input, applied convolutional layers to extract features, and used pooling to downsample and reduce dimensionality. The downsampled features were fed into fully connected layers to classify them as benign or malignant. CNN model used the training data and validated its performance on the validation set. This involved setting up callbacks for early stopping and model checkpointing. ViTs processed mammogram images by splitting them into fixed-size patches, linearly embedding each patch into a specific hidden dimension as tokens, adding position embeddings to retain spatial awareness, and feeding the resulting sequence of vectors to a standard transformer encoder. This encoder used layer normalization and residual connections, mapping the feature representation to class probabilities. These specific architectures were selected to examine the differences between the local hierarchical features and global contextual relationships of CNNs and ViTs, respectively. CNNs focused on continuous, localized regions, whereas ViTs distributed their focus across patches. These architectures also highlighted the need to handle large datasets and estimate uncertainty. The tested images were classified for accuracy, precision, recall, support, and F1 score. The trained and tested images were then fed to the XAI module to estimate uncertainty using MCD, which performs multiple forward passes for

each input image (50 times in our case), with dropout layers randomly dropping out neurons in each pass. This produced a distribution of predictions for each input, rather than a single prediction. This technique leveraged dropout layers, which were typically used during training to prevent overfitting, by keeping them active during inference (prediction time). This yielded the mean prediction (the model's most likely output) and the prediction variance or standard deviation, which measured how much individual predictions vary and quantified the model's uncertainty. A high standard deviation indicates high uncertainty, meaning the model's predictions were inconsistent across the different dropout passes. MCD provides a crucial layer of transparency, moving beyond what the model predicts to how confident it is. This is vital in critical applications like medical imaging. Grad-CAM offers a localization map that highlights the discriminative regions the model uses for a given class by leveraging the gradients of the target class's prediction score with respect to the feature maps of a final convolutional layer.

The CNNs emphasized localized spatial features, producing concentrated and contiguous heatmaps, whereas ViTs leveraged global self-attention, resulting in more distributed and patch-oriented activation patterns. SHAP primarily focused on the tissue region rather than the background by seeking clinical attention to support

the classes. Positive SHAP contributions for the benign and benign-without-callback classes were concentrated within the central breast tissue, indicating that texture and density characteristics in these regions support a non-malignant diagnosis. In contrast, the malignant class exhibited predominantly weak or negative SHAP contributions, suggesting the absence of discriminative malignant features. The interpreted result performances were compared for both CNN and ViT, providing valuable insight into how deep learning architectures interpret medical images for precise diagnostics.

3. Results and discussion

The simulated results were obtained for three categories, namely deep learning models (CNN and ViT), uncertainty estimations using MCD, and XAI using Grad-CAM and SHAP, as discussed in the following sections.

3.1. Deep learning models

The mammogram images were taken from the CBIS-DDSM dataset, and the deep learning models such as CNN

and ViT were trained, validated, and tested to obtain the performance metrics such as precision, recall, F1 score, support, and accuracy. The hyperparameters chosen were the Adam optimizer with a learning rate of 0.001, a batch size of 5, 20 epochs, and a dropout probability of 0.1. The preprocessing images are 100, with resizing dimensions of (224, 224), layer normalization, and data augmentation specifics such as random flip and random rotation (horizontal and vertical).

Figure 2 shows per-class precision for benign, malignant, and benign-without-callback for both the CNN and ViT. We observed that ViT's precision for cancer cells in the benign-without-callback model is zero, indicating that the dataset includes both benign and malignant cases. The CNN model showed better classification performance compared to ViT in terms of precision.

Figure 3 shows recall by class (benign, malignant, and benign-without-callback) for both CNN and ViT. It was found that the recall for the cancer cells for the ViT model (benign-without-callback) is zero, suggesting that

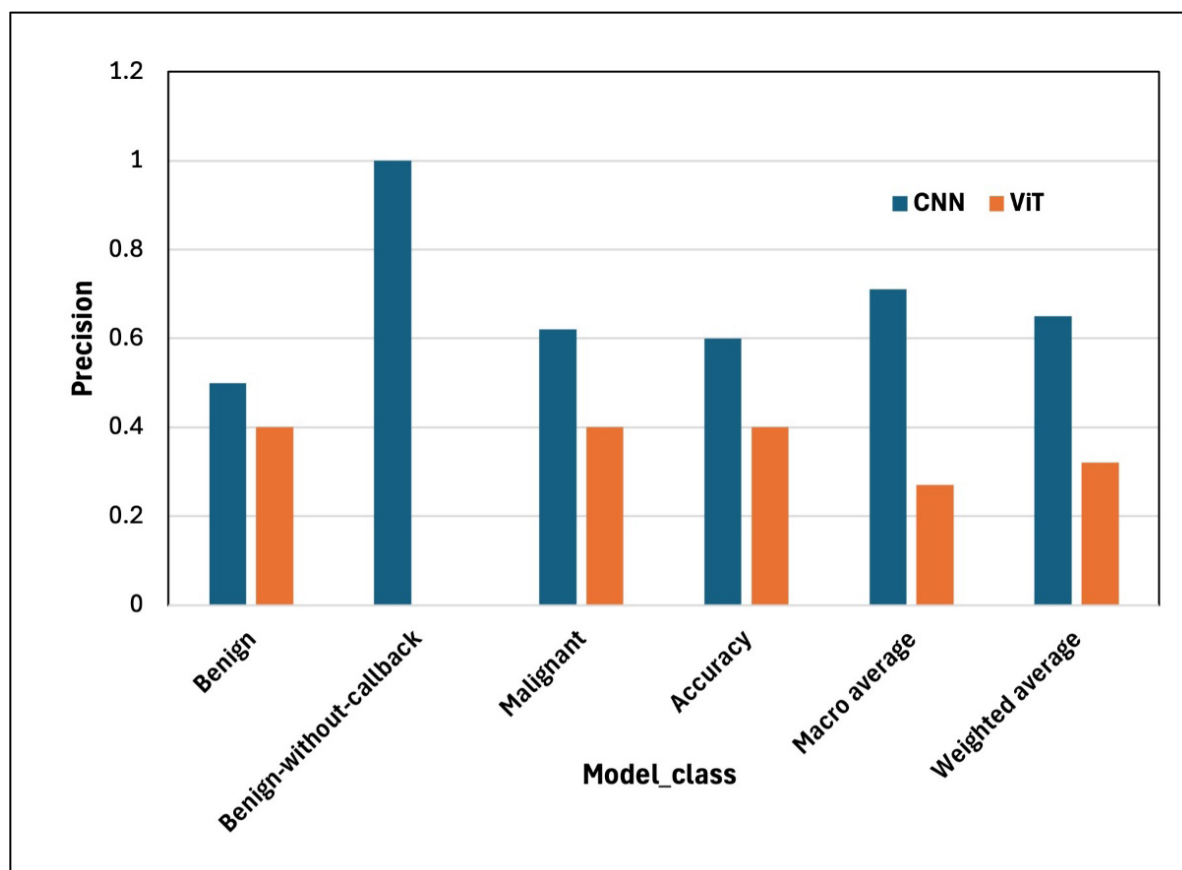


Figure 2. Precision by class (convolutional neural network [CNN] vs. vision transformer [ViT])

the dataset included both benign and malignant cases. Recall showed mixed results for both the CNN and ViT models in classifying the model class. Zero precision and/or recall for the benign-without-callback class indicates the model's inability to correctly classify instances of that class under the current experimental conditions, and this outcome is likely influenced by the limited dataset size, class imbalance, and restricted model generalization rather than by the presence of benign and malignant cases in the dataset.

Figure 4 shows the support by class, such as benign, malignant, and benign-without-callback, for both the CNN and ViT. The support was nearly identical for both models when classifying cancer cells. The dataset showed some class imbalance, with the benign-without-callback class containing only four samples compared to eight samples each for the *benign* and *malignant* classes in the test set. Although the imbalance was relatively small in absolute terms, it became substantial given the limited overall dataset size and may have contributed to the observed performance pattern of perfect precision but low recall for the minority class. Specifically, the classifier appears to make highly conservative predictions for this class, correctly identifying only one of the four available samples while avoiding false-positive predictions.

Figure 5 shows the F1 score by class (benign, malignant, and benign-without-callback) for both the CNN and ViT. The CNN achieved a higher F1 score than the ViT model.

As shown in Figure 6, the CNN achieved an overall accuracy of 60% on the unseen test set. While better than random chance (33.3% for the three classes), it was not a highly performant model, likely due to overfitting observed during training. Overall, for benign, the model achieved a precision of 0.50 and a recall of 0.62. This means that when it predicted “benign,” it was correct 50% of the time, and it identified 62% of the actual “benign” cases. The benign-without-callback class had a perfect precision (1.00), meaning every time the model predicted “benign-without-callback,” it was correct. However, its recall was only 0.50, indicating it missed half of the actual “benign_WITHOUT_CALLBACK” cases. The “malignant” class showed a precision and recall of 0.62. This is a crucial class for breast cancer detection, and the moderate performance suggests room for improvement. Despite learning quickly on the training data, the CNN showed signs of overfitting on the chosen dataset. Its test-set performance was moderate, with varying success across classes. Uncertainty analysis using MCD could further clarify the model's confidence in these predictions. Overall, the CNN serves as a baseline, but Grad-CAM interpretability is the key to understanding its decision-making.

3.2. Uncertainty estimations

The uncertainty estimations using MCD predictions for CNN and ViT models are discussed as follows:

3.2.1. Monte Carlo Dropout predictions for convolutional neural network model

Figures 7–11 show the uncertainty estimates for CNNs for five random samples across the three classes, with various combinations of model classifications.

The difference in prediction uncertainty between the CNN and ViT models, as indicated by the mean prediction standard deviation, has substantial clinical implications: Figures 7–11 show the CNN's MCD predictions, which have lower uncertainty (0.006 mean), meaning the model is more confident in its predictions. In a clinical setting for breast cancer classification, this means that when the CNN predicts a certain outcome (e.g., benign or malignant), it does so with less variability across multiple MCD runs. High-confidence predictions can be beneficial, as they may reduce the need for further diagnostic tests if the model's confidence aligns with the clinical picture. However, overconfidence, especially if the accuracy is not high, can be dangerous, leading to misdiagnoses if the model is confidently wrong.

3.2.2. Monte Carlo Dropout predictions for vision transformer model

The following images show the ViT model's uncertainty estimates for five random samples across the three classes, with various combinations of model classifications.

Figures 12–16 show that the MCD predictions for the ViT had a higher uncertainty of 0.339, meaning the ViT model implied greater variability and less confidence in its predictions. Clinically, this means that if a ViT model is to be used, its predictions will come with a larger margin of error or a higher degree of doubt. This increased uncertainty would require more careful human review and potentially additional diagnostic procedures to confirm findings, even when a prediction is made. While it might prevent over-reliance on a potentially incorrect confident prediction, it also adds to the diagnostic burden. In conclusion, the CNN's lower uncertainty and higher accuracy in this dataset suggest it could be more reliably integrated into a diagnostic workflow for initial screening, provided its accuracy is clinically acceptable. In this case, the low accuracy and high uncertainty of ViT demonstrate that it is less appropriate for direct clinical use; rather, it indicates a high likelihood of unreliable predictions. This characterization and quantification are essential for trust in AI-based tools.

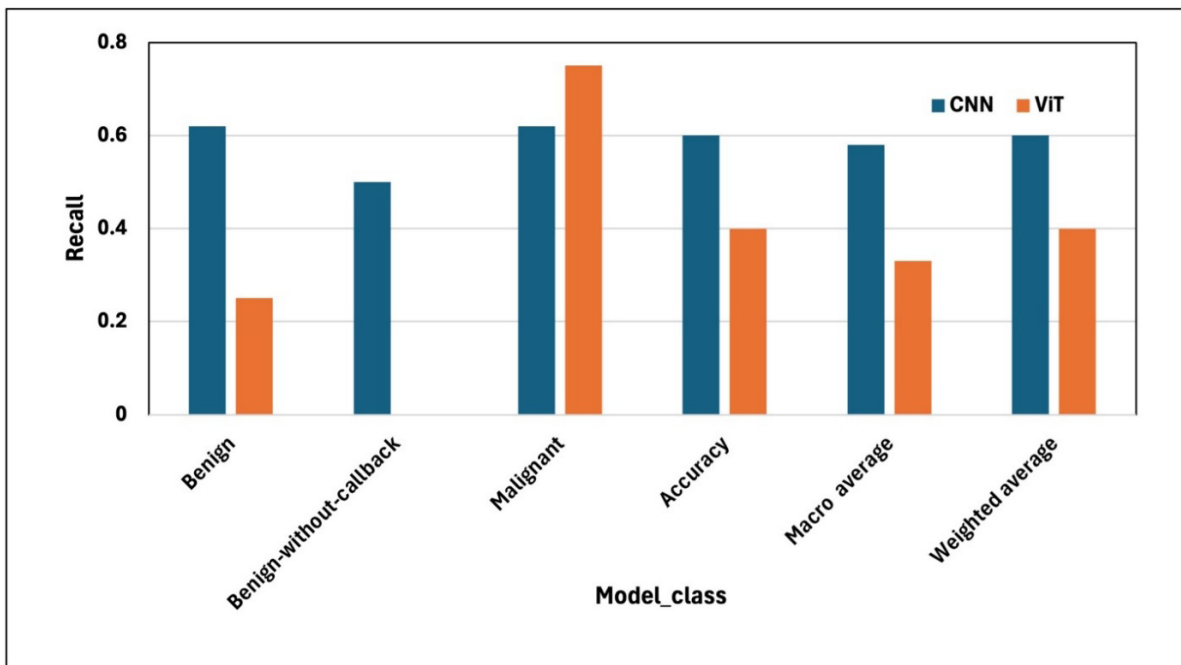


Figure 3. Recall by class (convolutional neural network [CNN] vs. vision transformer [ViT])

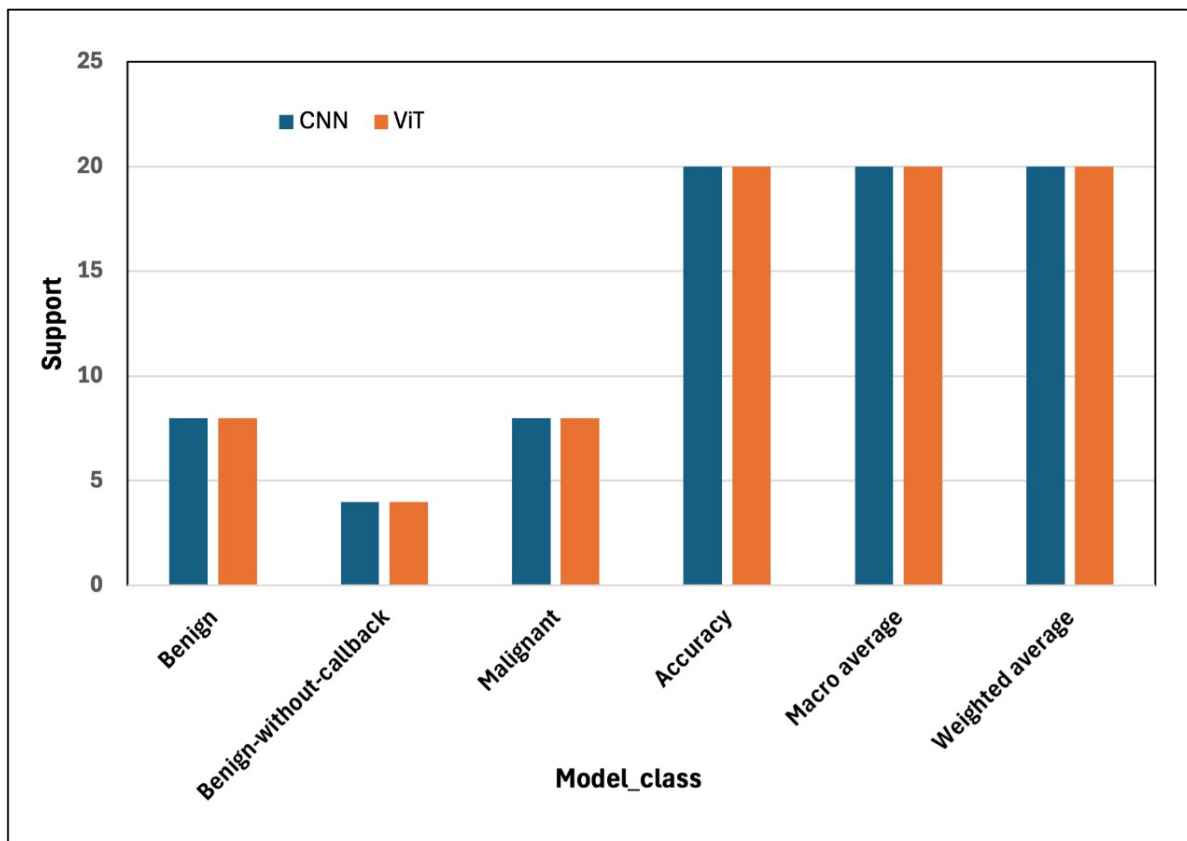


Figure 4. Support by class (convolutional neural network [CNN] vs. vision transformer [ViT])

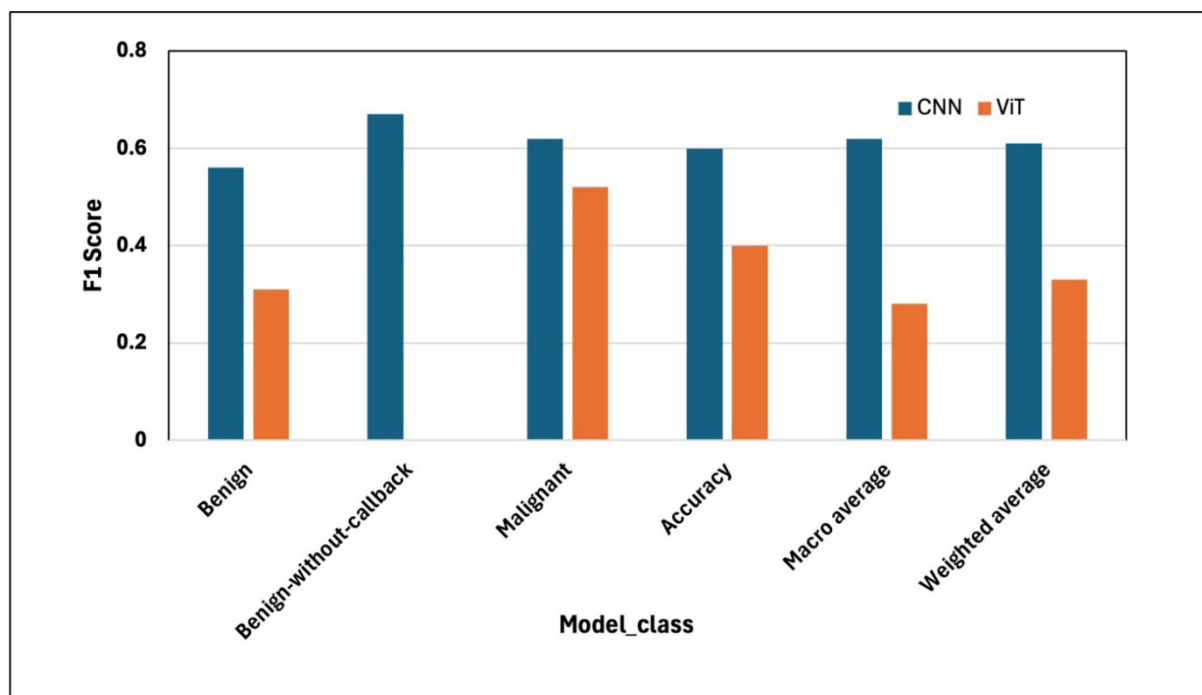


Figure 5. F1 Score by class (convolutional neural network [CNN] vs. vision transformer [ViT])

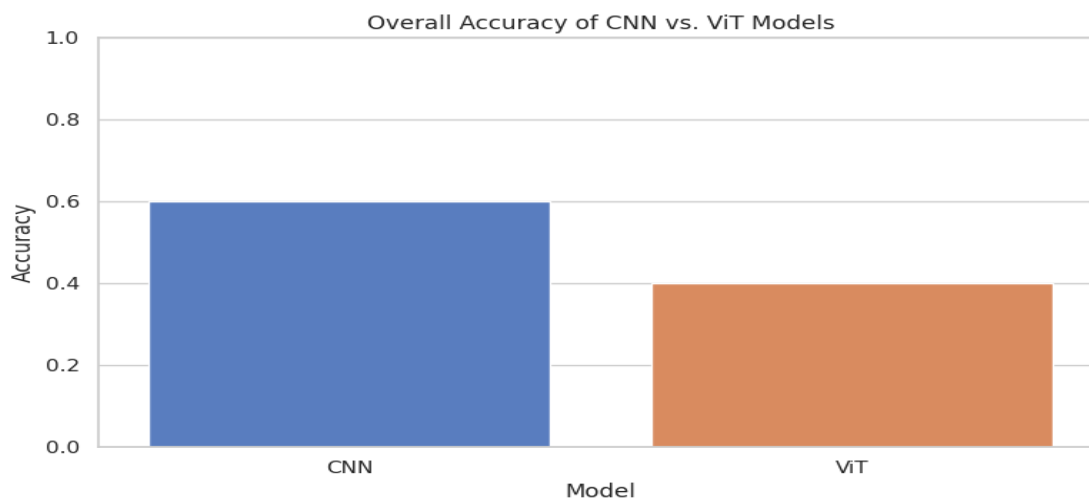


Figure 6. Accuracy of convolutional neural network (CNN) vs. vision transformer (ViT)

3.3. Explainable artificial intelligence techniques

This section discusses Grad-CAM for saliency maps and SHAP for feature importance for both CNN and ViT models for breast cancer classification. Grad-CAM was originally developed for CNNs and cannot be directly applied to ViTs because ViTs lack convolutional layers in their primary feature-extraction path. For ViTs, we

adapted this technique by targeting the output of the last transformer block's layer normalization layer before the Flatten operation. The output represented processed patch embeddings and could be reshaped into a pseudo-spatial grid, allowing us to compute gradients and generate a heatmap indicating the importance of different image regions (patches). We clarified that the ViT visualization

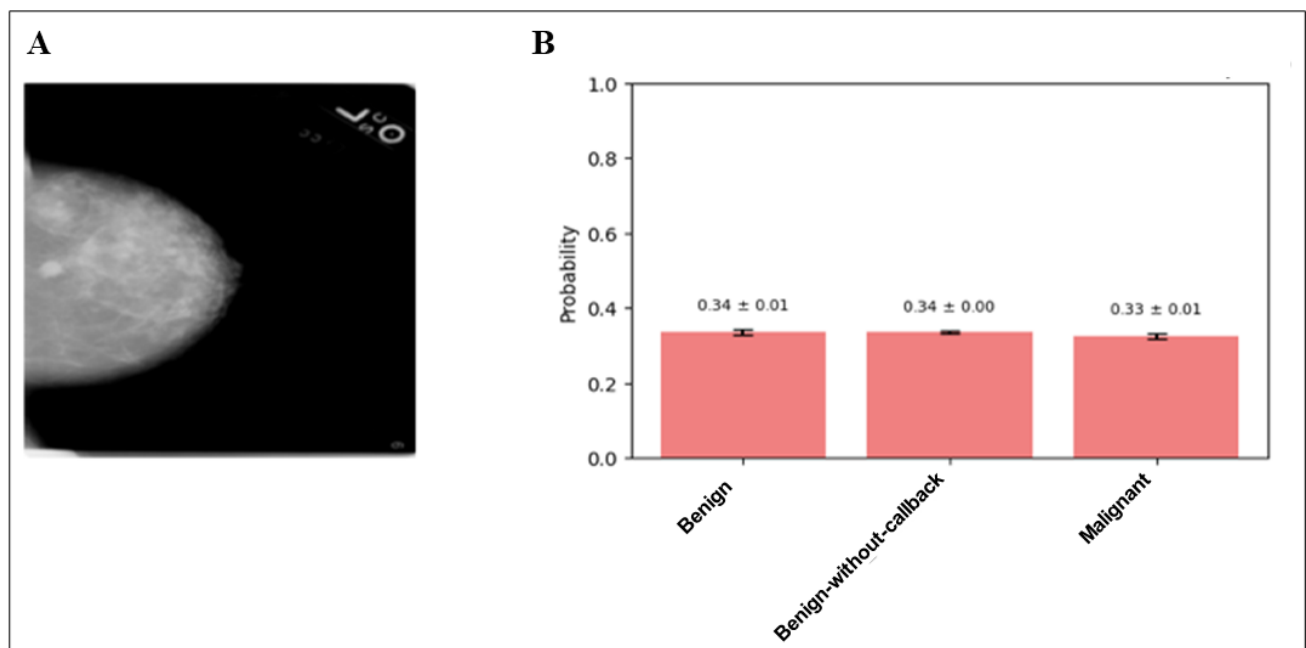


Figure 7. Monte Carlo Dropout predictions for a convolutional neural network (CNN) for sample 0. (A) Original image (true: Malignant; predicted: Benign). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

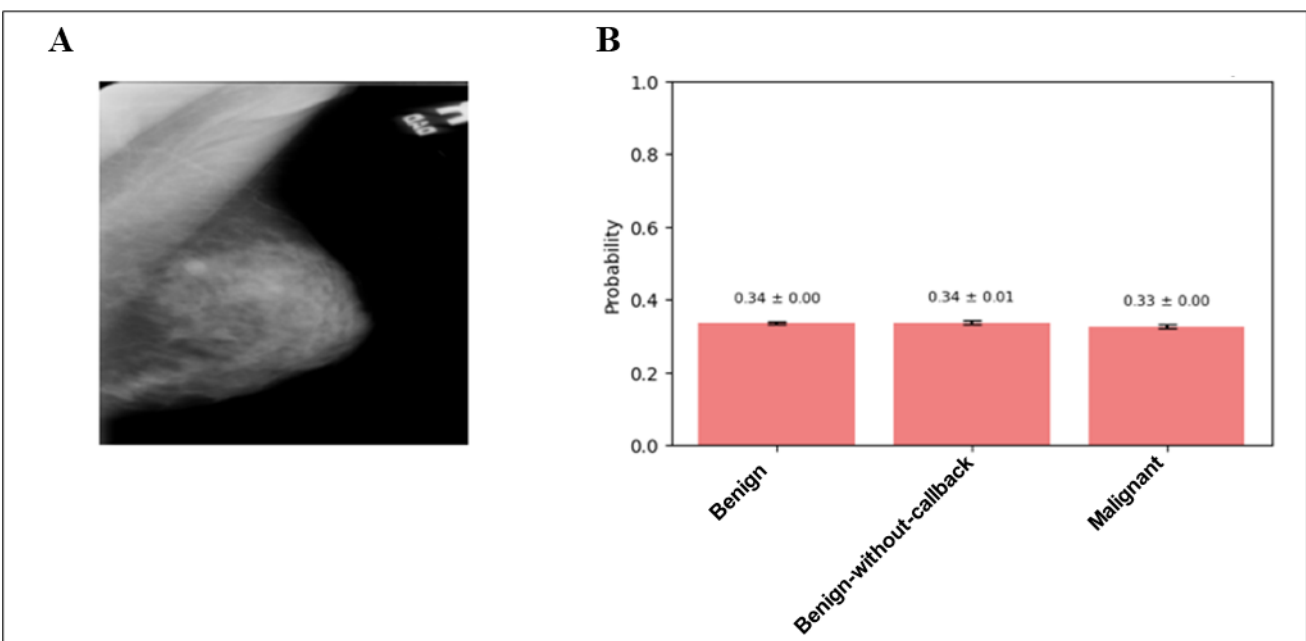


Figure 8. Monte Carlo Dropout predictions for a convolutional neural network (CNN) for sample 1. (A) Original image (true: Malignant; predicted: Benign-without-callback). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

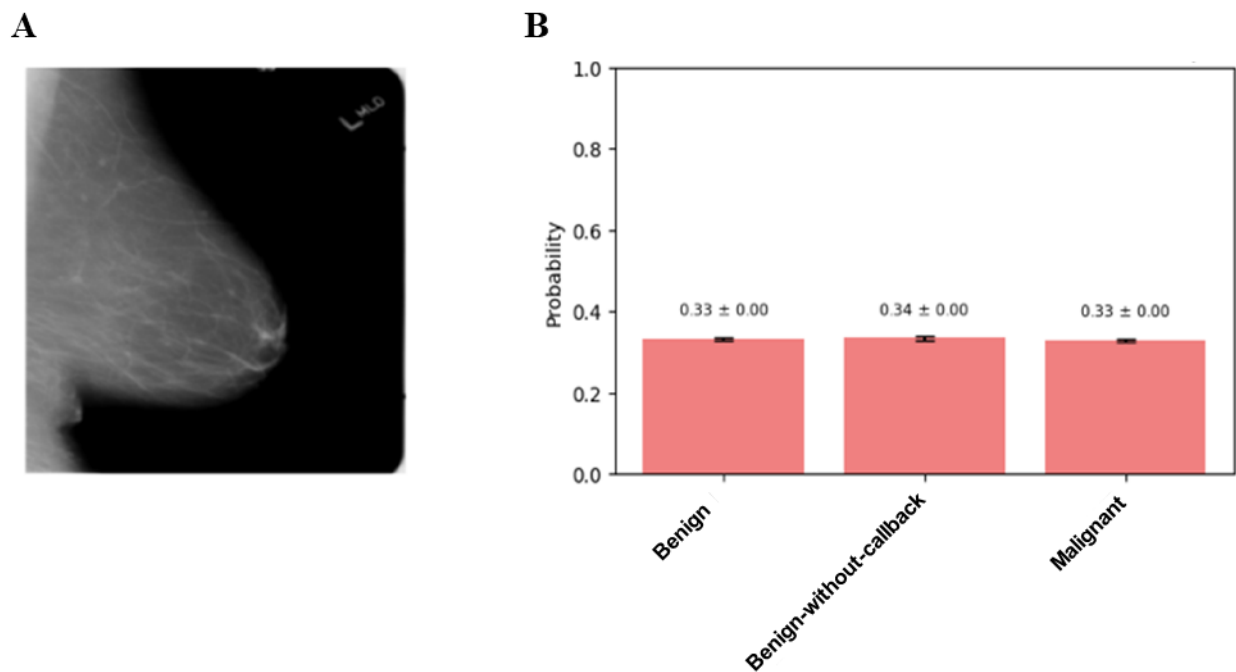


Figure 9. Monte Carlo Dropout predictions for a convolutional neural network (CNN) for sample 2. (A) Original image (true: Benign-without-callback; predicted: Benign-without-callback). (B) Monte Carlo Dropout predictions. Results presented are mean $\pm$ standard deviation.

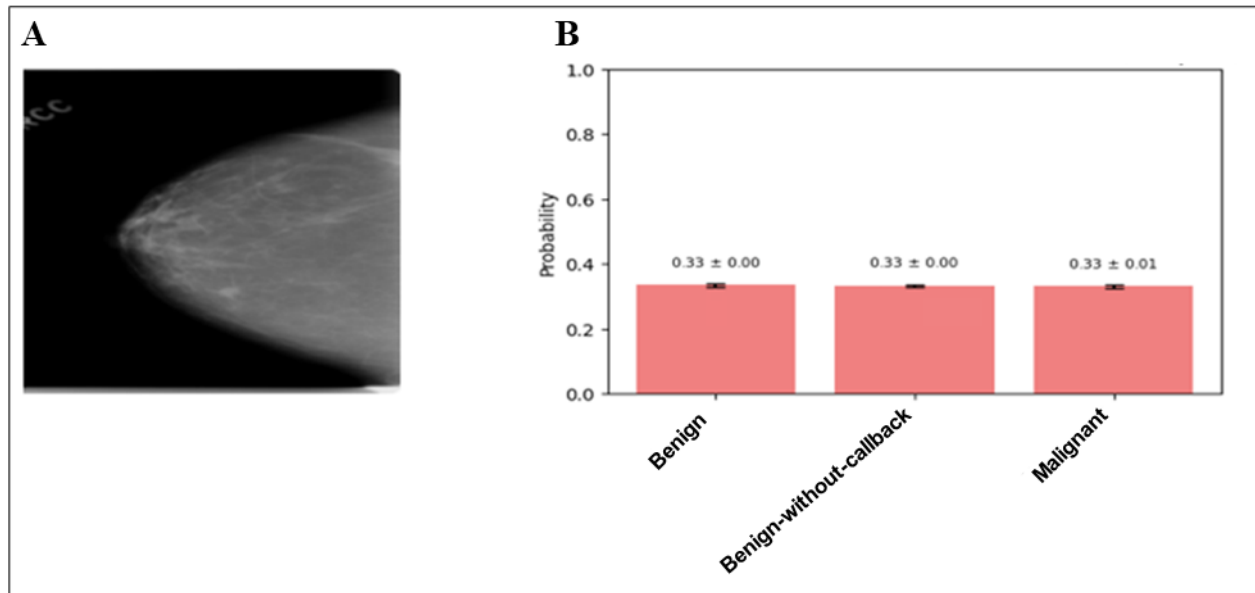


Figure 10. Monte Carlo Dropout predictions for a convolutional neural network (CNN) for sample 3. (A) Original image (true: Malignant; predicted: Benign). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

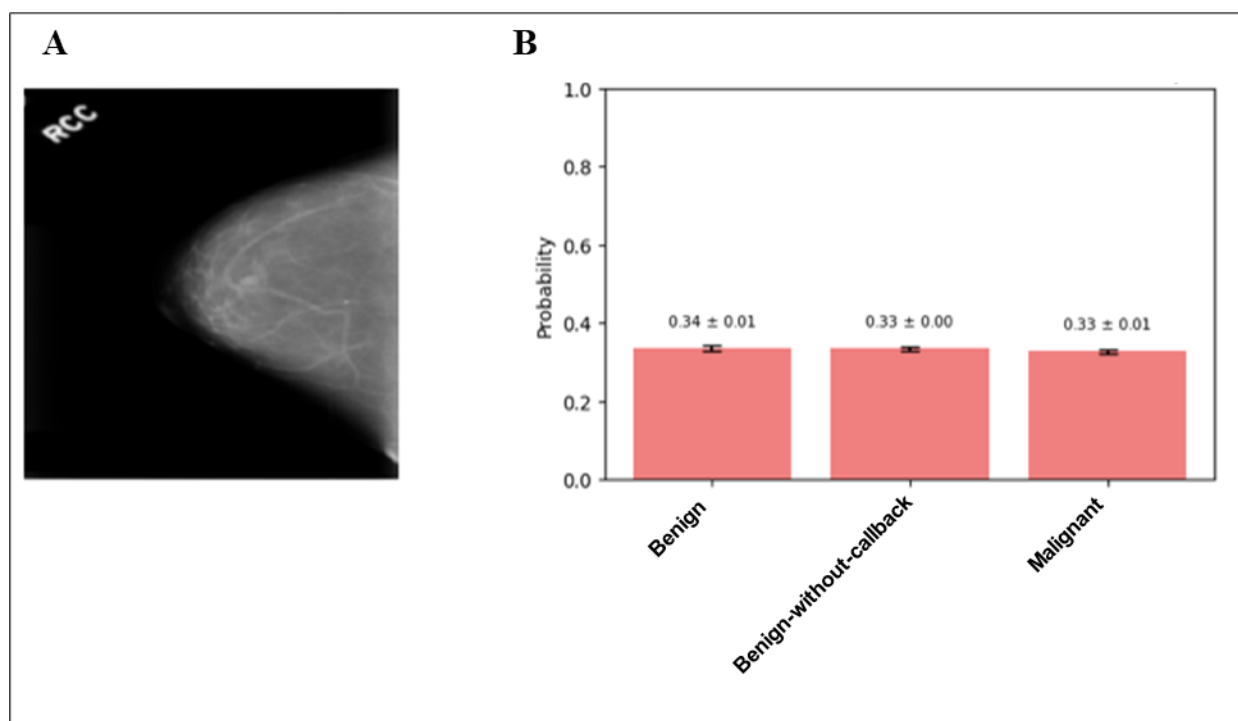


Figure 11. Monte Carlo Dropout predictions for a convolutional neural network (CNN) for sample 4. (A) Original image (true: Benign; predicted: Benign). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

followed a transformer-compatible adaptation of Grad-CAM rather than the standard CNN implementation. These additions improve the transparency, reproducibility, and methodological rigor of the proposed comparative analysis.

3.3.1. Explainable artificial intelligence gradient-weighted class activation mapping for convolutional neural network model

Unlike simpler methods, Grad-CAM provides a localization map (heatmap) that highlights the discriminative regions used by the model for a given class. It does this by leveraging the gradients of the target class's prediction score with respect to the feature maps of a final convolutional layer. Figures 17–19 show the CNN Grad-CAM for all three classes: benign, malignant, and benign-without-callback.

Figure 17 shows the heatmap generated for the CNN's predicted class (benign-without-callback); it shows which visual features the CNN identified as characteristic of benign-without-callback. For instance, it might highlight certain textures, densities, or edges within the mammogram image.

Figure 18 shows the heatmap for the true class (malignant); it shows which features the model would have

looked for to correctly identify the image as malignant. If this heatmap is diffuse or weak in relevant areas, it suggests the model failed to find the discriminative features for this class.

Figure 19 shows that the CNN incorrectly predicted benign-without-callback as malignant, so these Grad-CAMs are particularly valuable for understanding this misclassification. The results show that CNNs learn image representations through hierarchical convolutional operations. Lower layers of CNNs learn basic visual features, such as edges, textures, and gradients, while higher layers gradually learn more complex semantic features, such as object parts and the whole object. CNN heatmaps often appear as small areas that highlight disease-specific structures, making it easier for clinicians to identify the image regions that drive the prediction.

3.3.2. Explainable artificial intelligence gradient-weighted class activation mapping for vision transformer model

Figures 20–22 show the Grad-CAM for all three classes: benign, malignant, and benign-without-callback. Figure 20 shows a suspicious dense region in the mammogram image that should be recognized as malignant. However,

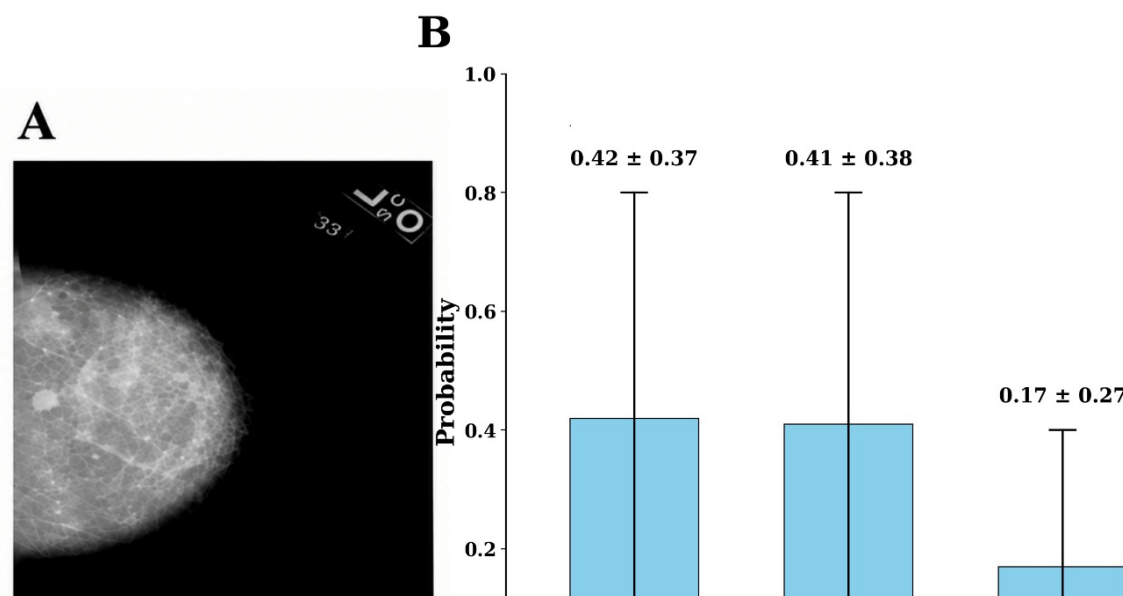


Figure 12. Monte Carlo Dropout predictions for vision transformer (ViT) sample 0. (A) Original image (true: Malignant; predicted: Benign). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

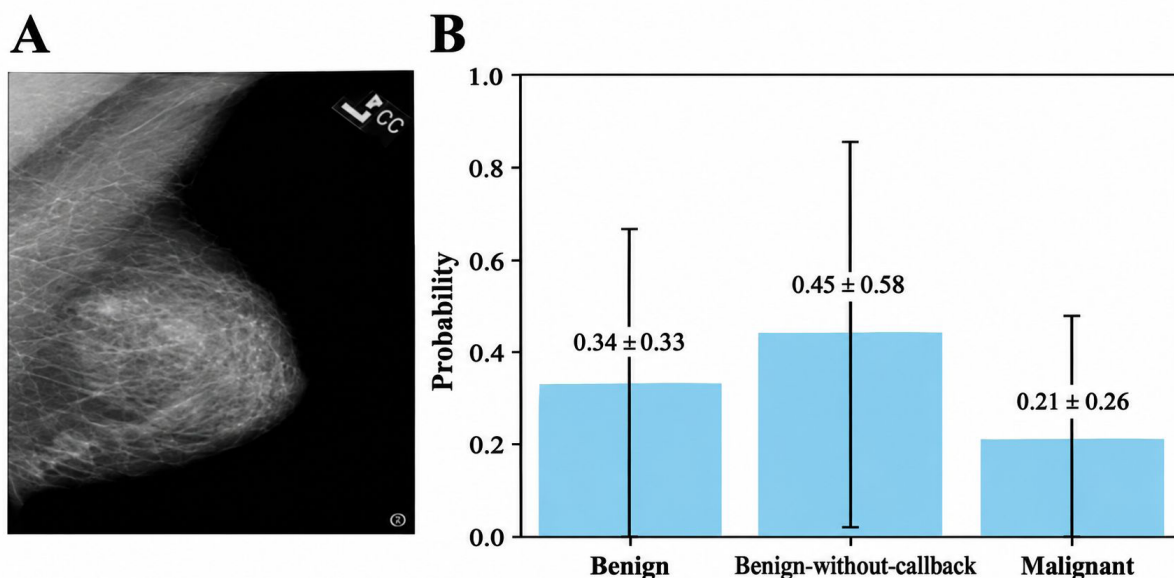


Figure 13. Monte Carlo Dropout predictions for vision transformer (ViT) sample 1. (A) Original image (true: Malignant; predicted: Benign-without-callback). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

the model predicted the image as benign, suggesting the learned features were insufficient to distinguish malignant characteristics. This indicates the ViT model relies on global contextual information rather than the suspicious lesion alone. Therefore, the model might focus on non-

discriminative tissue structures, resulting in an incorrect benign prediction. From a clinical point of view, the Grad-CAM explanation provides a general assessment of the breast tissue rather than a specific assessment of the suspicious lesion that led to the ViT model's decision.

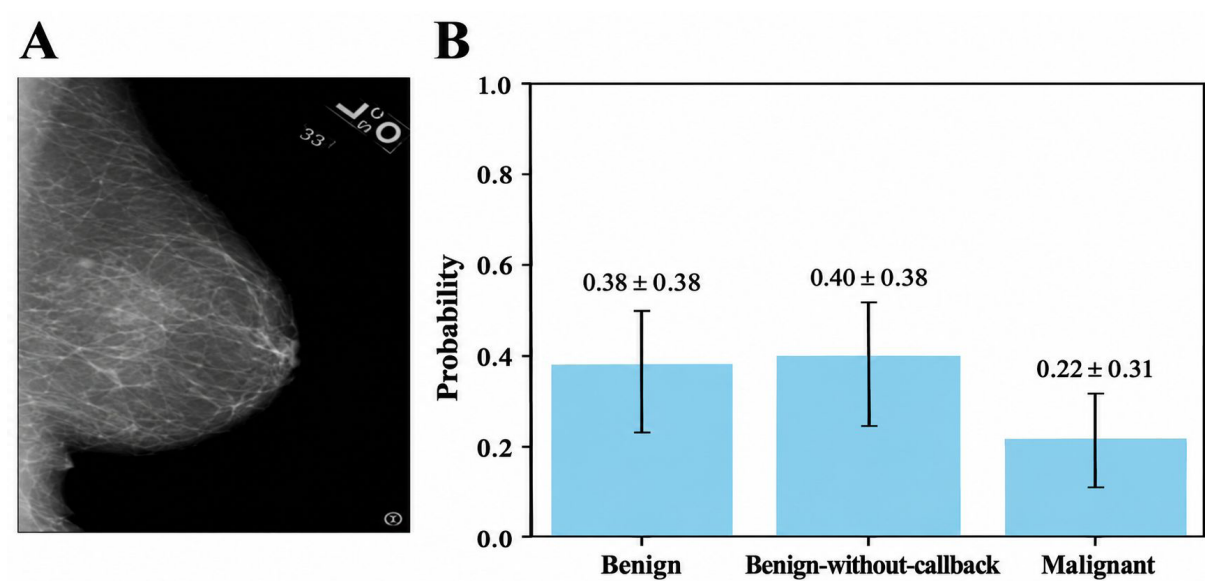


Figure 14. Monte Carlo Dropout predictions for vision transformer (ViT) sample 2. (A) Original image (true: Benign-without-callback; predicted: Benign-without-callback). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

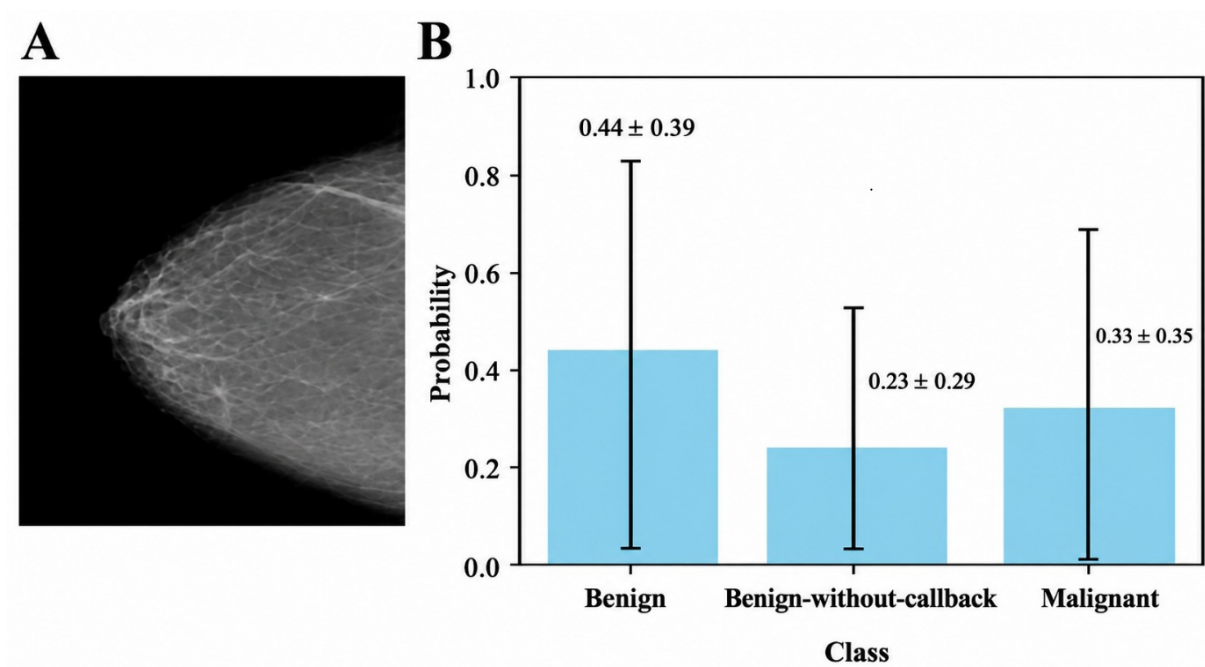


Figure 15. Monte Carlo Dropout predictions for vision transformer (ViT) sample 3. (A) Original image (true: Malignant; predicted: Benign). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

Since malignant tumors are often diagnosed by localized characteristics such as irregular shape, spiculated edges, and atypical density patterns, the distributed attention pattern shows that the model failed to capture the most

diagnostically relevant traits.

From Figure 21, the Grad-CAM visualization showed that the ViT primarily focused on diffuse breast tissue regions rather than the entire boundary of the malignant

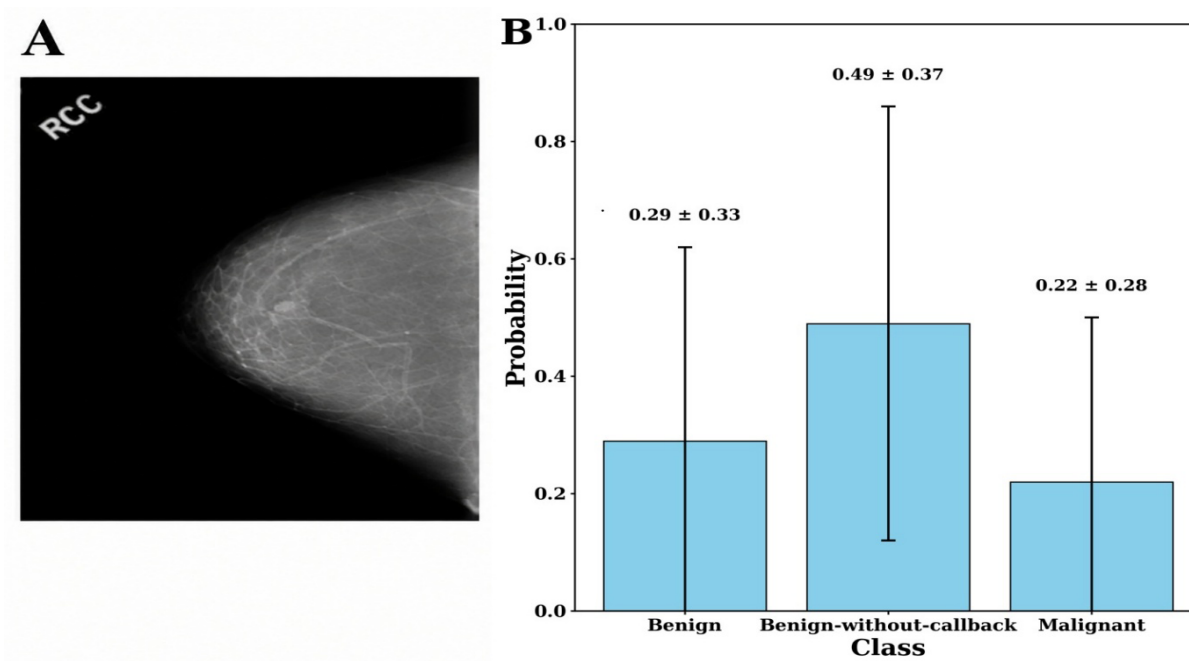


Figure 16. Monte Carlo Dropout predictions for vision transformer (ViT) sample 4. (A) Original image (true: Benign; predicted: Benign-without-callback). (B) Monte Carlo Dropout predictions. Results presented are the mean $\pm$ standard deviation.

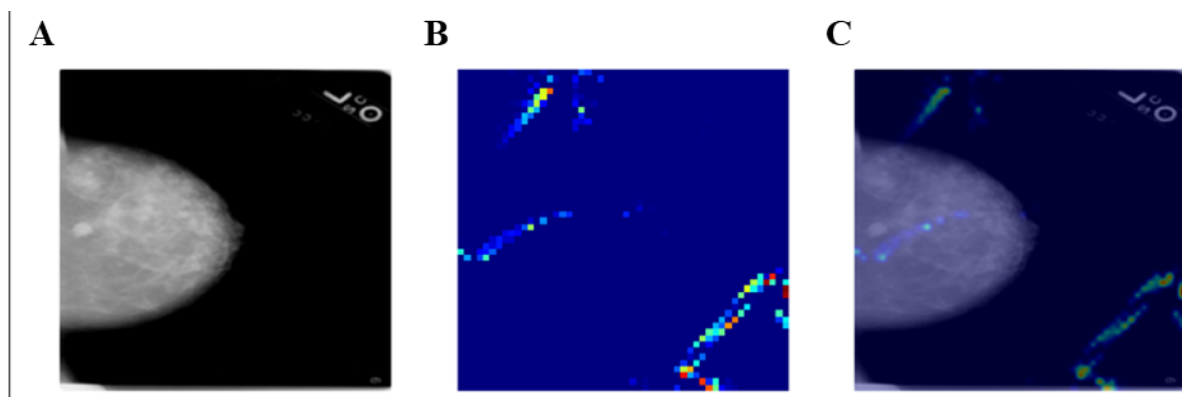


Figure 17. Generating convolutional neural network (CNN) gradient-weighted class activation mapping (Grad-CAM). (A) Original image (true: malignant; predicted: benign-without-callback). (B) CNN Grad-CAM heatmap (raw) for class: benign. (C) Grad-CAM explanation.

lesion. Although clinically relevant characteristics like speculated margins and architectural distortion received little attention, high-activation regions appeared within homogeneous tissue areas. As a result, the model falsely classified the benign mammogram as benign-without-callback. This false-negative prediction suggests the transformer may have learned global texture information but did not sufficiently represent local malignant features. This finding highlights the need for hybrid architectures

that combine transformer-based global context learning with CNN-based local feature extraction to improve breast cancer detection accuracy.

In Figure 22, the Grad-CAM visualization showed that the ViT had scattered, sparse attention throughout the mammogram. Several high-intensity activations were concentrated in the background and peripheral areas, even though a small activation region overlapped the malignant lesion. The model did not capture all aspects of the

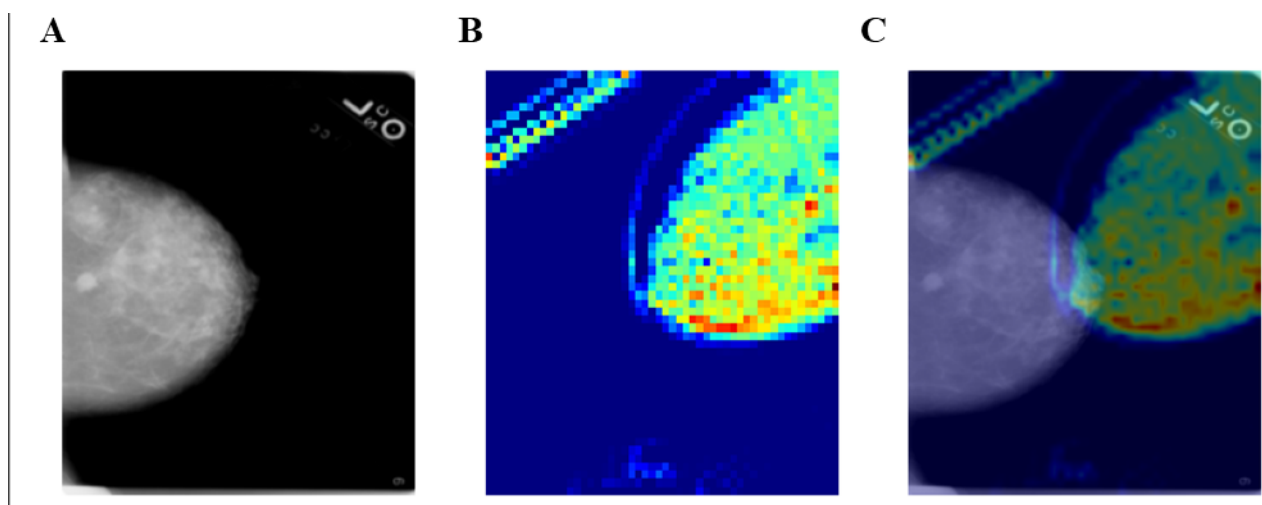


Figure 18. Generating convolutional neural network (CNN) gradient-weighted class activation mapping (Grad-CAM). (A) Original image (true: malignant; predicted: benign-without-callback). (B) CNN Grad-CAM heatmap (raw) for class: benign-without-callback. (C) Grad-CAM explanation.

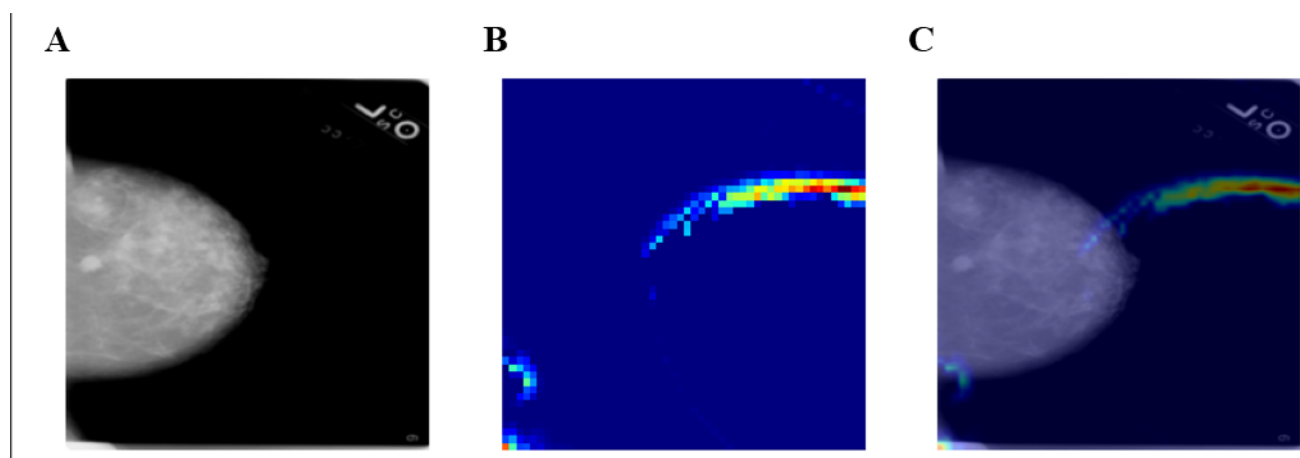


Figure 19. Generating convolutional neural network (CNN) gradient-weighted class activation mapping (Grad-CAM). (A) Original image (true: malignant; predicted: benign-without-callback). (B) CNN Grad-CAM heatmap (raw) for class: malignant. (C) Grad-CAM explanation.

lesion, including diagnostically important characteristics such as irregular margins and heterogeneous tissue structures. As a result, the model produced a false-negative prediction because the extracted feature representation was insufficient for accurate malignancy recognition. This behavior supports integrating CNN-based local feature extraction with attention-guided learning mechanisms to improve lesion localization and classification accuracy, and it highlights the limitations of patch-based transformer attention in mammographic analysis.

Unlike CNNs, ViTs divide an image into fixed-size patches and process them as tokens using self-attention.

The model can capture global contextual information and long-range dependencies because each patch can attend to every other patch in the image. ViTs may prioritize multiple spatially distinct patches that contribute to the classification decision rather than a single continuous region. Table 1 shows the observed differences in Grad-CAM focus between CNN and ViT models.

In summary, CNNs' Grad-CAMs often show a clear, localized “where” in the image, reflecting their local processing. Due to global attention, ViTs' Grad-CAMs may indicate a more dispersed or abstract set of “what patches” contribute, potentially highlighting more extensive or

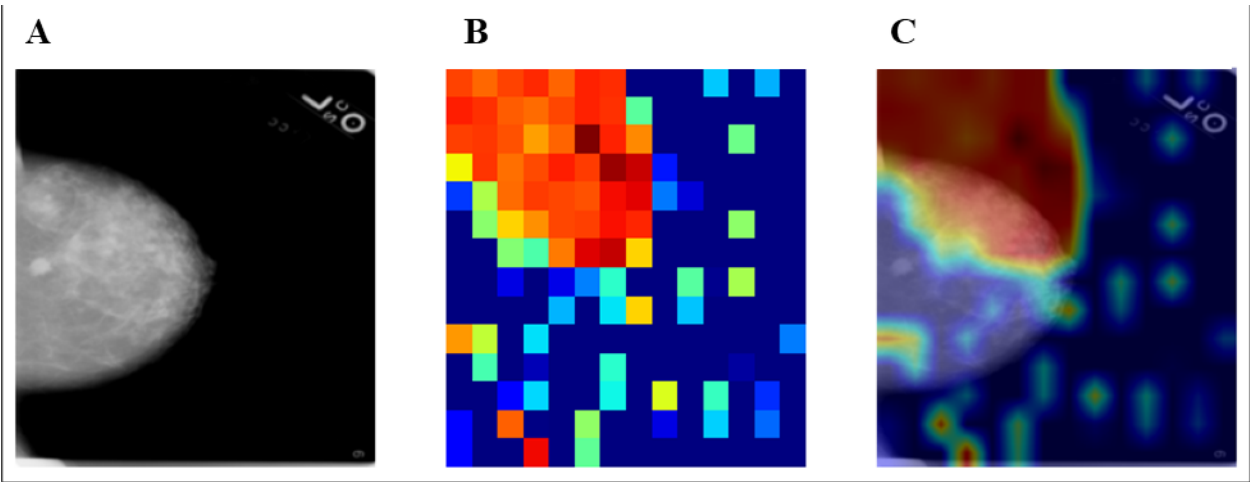


Figure 20. Generating vision transformer (ViT) gradient-weighted class activation mapping (Grad-CAM). (A) Original image (true: malignant; predicted: benign-without-callback). (B) ViT Grad-CAM heatmap (raw) for class: benign. (C) Grad-CAM explanation.

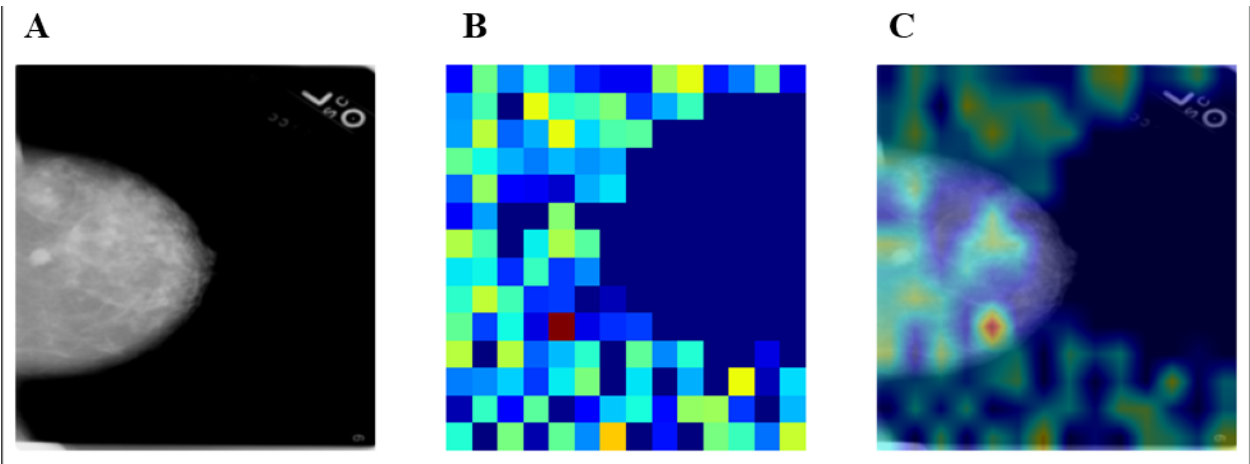


Figure 21. Generating vision transformer (ViT) gradient-weighted class activation mapping (Grad-CAM). (A) Original image (true: malignant; predicted: benign-without-callback). (B) ViT Grad-CAM heatmap (raw) for class: benign-without-callback. (C) Grad-CAM explanation.

Table 1. Observed differences in Grad-CAM focus between CNN and ViT

Aspect	CNN	ViT
Feature learning	Local hierarchical features	Global contextual relationships
Attention region	Continuous and localized	Distributed and patch-based
Spatial interpretability	High	Moderate
Context awareness	Limited local context	Strong global context
Grad-CAM appearance	Smooth blobs	Patchy or dispersed regions

Abbreviations: CNN: Convolutional neural network; Grad-CAM: Gradient-weighted class activation mapping; ViT: Vision transformer.

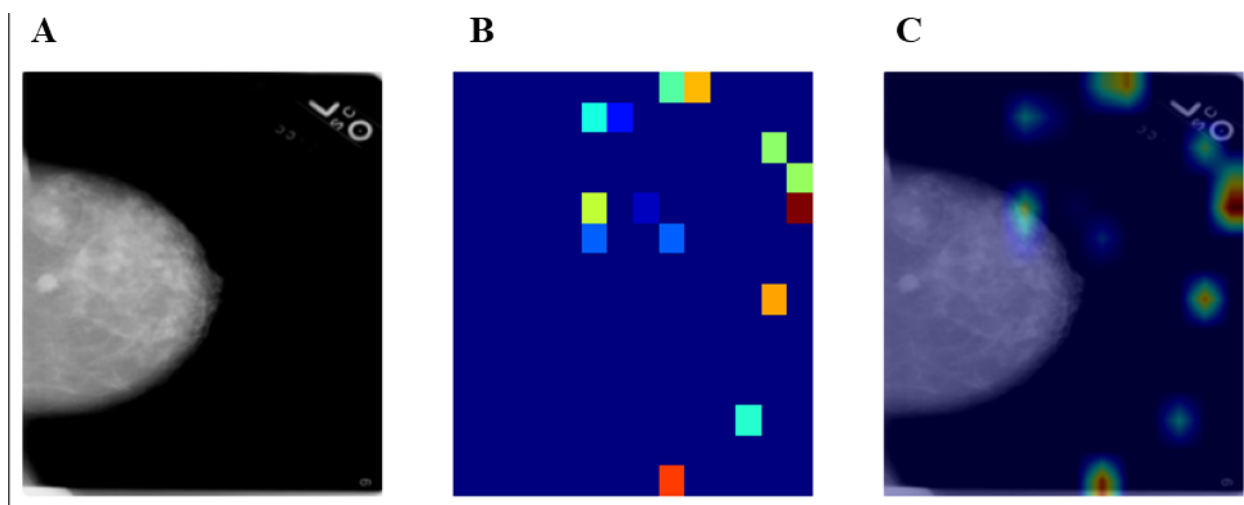


Figure 22. Generating vision transformer (ViT) gradient-weighted class activation mapping (Grad-CAM). (A) Original image (true: malignant; predicted: benign-without-callback). (B) ViT Grad-CAM heatmap (raw) for class: malignant. (C) Grad-CAM explanation.

disjointed areas of significance.

3.3.3. Explainable artificial intelligence—SHapley Additive exPlanations for convolutional neural networks

As shown in Figure 23, the red/pink regions (positive SHAP values) indicate an increase in the probability of the corresponding class and highlight image regions supporting the model's prediction. The color bar at the bottom represents SHAP values. Negative SHAP values, which are represented by blue regions, indicate a decrease in the probability of the corresponding class and contrast with the prediction. The classification decision is largely influenced by the gray regions. The numerical values reported for the CNN ($\approx 3 \times 10^{-6}$) and ViT ($\approx 1 \times 10^{-5}$) should not be directly compared as indicators of stronger or weaker feature importance. Instead, the analysis focused on the relative ranking and spatial distribution of important features within each model, which is the intended purpose of SHAP.

Figure 23 shows a SHAP visualization of a CNN model that classified breast mammograms into three output classes: benign, benign-without-callback, and malignant. SHAP highlighted the image regions that positively or negatively influenced each class prediction. The small magnitude of the SHAP values (around 3×10^{-6}) suggests that the model spreads its focus across multiple image patches rather than heavily focusing on a single region. In the SHAP visualization, the model mainly focused on clinically relevant breast tissue areas and largely ignored the background. The highest concentration of positive SHAP

contributions for the benign and benign-without-callback classes is within the central breast tissue, suggesting that texture and density characteristics in these regions support a non-malignant diagnosis. In contrast, the malignant class mostly showed weak or negative SHAP contributions, indicating a lack of discernible malignant characteristics. This framework interpretation and reliability confirm that the model's prediction enhances the localized patch-wise attribution pattern.

3.3.3. Explainable artificial intelligence—SHapley Additive exPlanations for vision transformer

Figure 24 shows wider SHAP values, indicating greater feature contributions and improved model interpretability. The SHAP visualization shows that the decision was driven by the primary contribution of clinically relevant breast-tissue characteristics. For the benign class, strong positive attributions in many breast-tissue patches suggest that the learned features align with benign morphological patterns. In contrast, the malignant class contributed both positively and negatively. This means some features were suspicious, but not enough to dominate the final prediction. The benign-without-callback class received intermediate support, reflecting partially normal tissue. Table 2 discusses the SHAP-based model interpretation with clinical relevance. Table 3 shows the overall model decision analysis.

Lastly, the SSIM measures similarity between two images based on three components: brightness, contrast, and structure. The SSIM score ranges from -1 to 1 , where 1 indicates perfect similarity, 0 indicates no structural

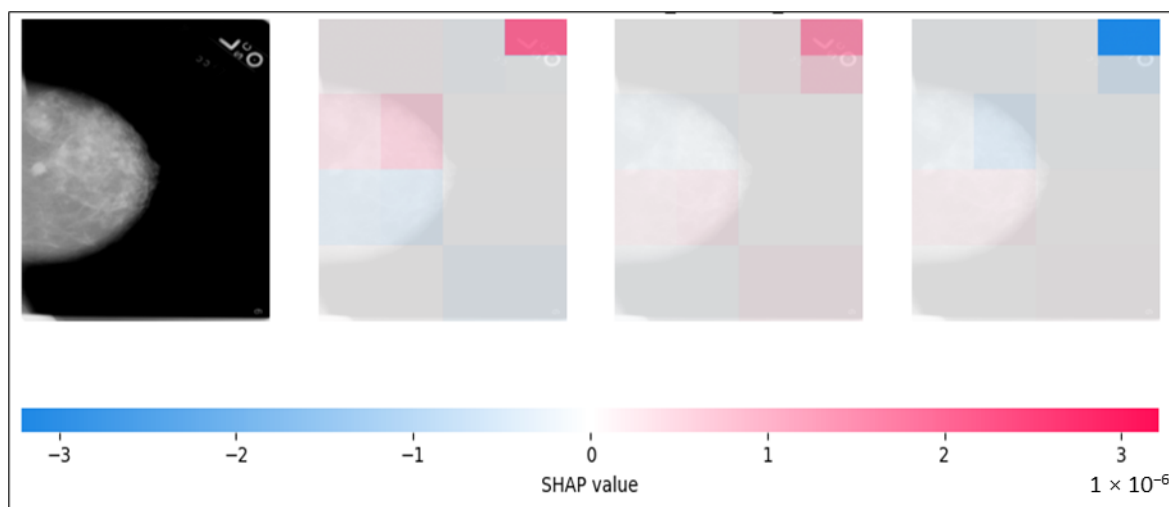


Figure 23. Generating SHapley Additive exPlanations (SHAP) for a convolutional neural network model

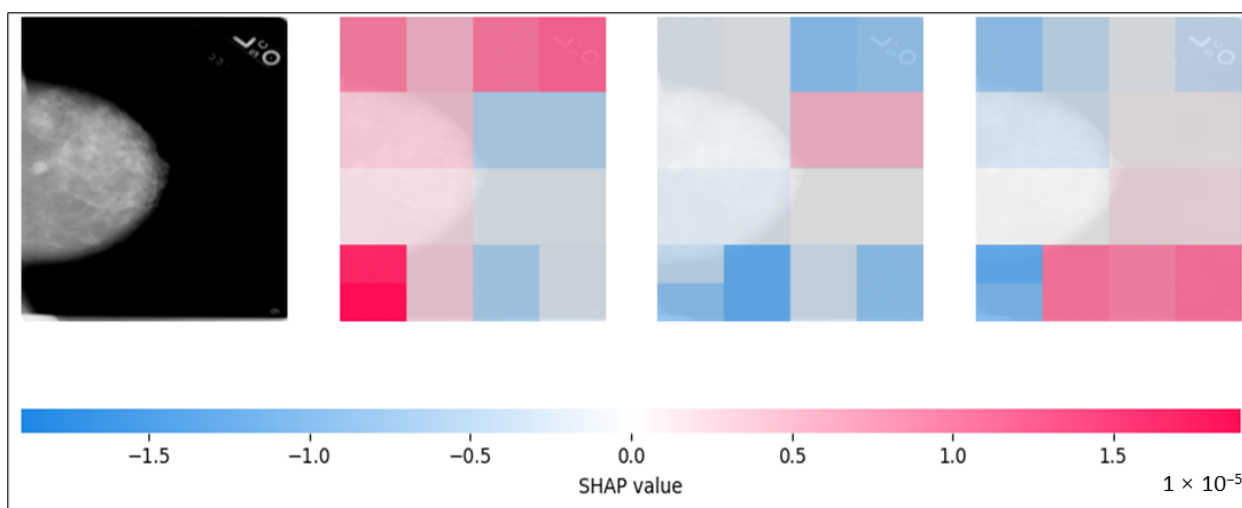


Figure 24. Generating SHapley Additive exPlanations (SHAP) for a vision transformer model

similarity, and negative values imply anti-correlation. Low SSIM values may arise from multiple factors, including genuine architectural differences in feature representation, as well as noisy or unstable Grad-CAM heatmaps resulting from limited training data, moderate model performance, or visualization artifacts. Accordingly, we interpreted the reported SSIM values as evidence of low similarity between the generated explanation maps, rather than definitive proof of distinct attention mechanisms. Here, SSIM was computed between the Grad-CAM heatmaps produced by the CNN and ViT models for the same input image and target class. These heatmaps visually represent

the regions each model deems important for a specific prediction. Table 4 shows the SSIM calculated for two samples from the original image, considered as TRUE, and the class as malignant. The table shows the calculation of SSIM between a CNN's heatmap and a ViT's heatmap as a numerical measure of how structurally similar their "attention" or "focus areas" are. A high SSIM score suggests that both models attend to similar parts of the image when making decisions for a particular class. A low SSIM score indicates they attend to different regions, reflecting their architectural differences (convolutional feature learning vs. patch-based attention).

Table 2. SHapley Additive exPlanations (SHAP)-based model interpretation

Class	SHAP observation	Interpretation	Clinical relevance
Benign	Positive SHAP values (pink/red regions) are concentrated in the upper-central and central breast regions.	These image patches increase the probability of the lesion being classified as benign. The model identifies smooth tissue patterns and homogeneous density characteristics associated with non-cancerous lesions.	Suggests normal tissue architecture or benign masses without suspicious margins.
Benign-without-callback	Moderate positive SHAP contributions appear in central and upper-right regions with relatively low magnitude.	The model recognizes features indicative of a normal screening mammogram that does not require further diagnostic evaluation.	Consistent with routine screening cases exhibiting no significant abnormalities.
Malignant	Strong negative SHAP values (blue regions) dominate the upper-right patch, while weaker contributions appear near the lesion area.	Features in these patches decrease confidence in the malignant class. The absence of strong positive malignant evidence suggests the model does not strongly associate the current lesion with cancer-related characteristics.	Indicates a lack of prominent malignant indicators such as spiculated margins, architectural distortion, or suspicious calcifications.
Background regions	Large gray areas correspond to SHAP values close to zero.	These regions contribute minimally to the model decision and are effectively ignored during classification.	Demonstrates that the model focuses primarily on breast tissue rather than irrelevant background pixels.

Table 3. Overall model decision analysis

Aspect	Observation
Primary attention region	Central breast tissue containing the visible lesion.
Most influential class	Benign and benign-without-callback receive stronger positive contributions than malignant.
Malignant evidence	Weak; no dominant positive SHAP regions support malignancy.
Model confidence pattern	The model appears to rely on localized tissue texture and density characteristics rather than suspicious cancer-specific features.
Explainability insight	SHAP confirms that the classifier focuses on clinically relevant breast tissue regions rather than image artifacts or background areas.

Abbreviation: SHAP: SHapley Additive exPlanations.

Table 4. SSIM for two samples

Samples	Class	Original image: TRUE	CNN prediction	ViT prediction	SSIM
Sample 1	Benign	Malignant	Benign-without-callback	Benign	0.0012
	Benign-without-callback	Malignant	Benign-without-callback	Benign	0.0104
	Malignant	Malignant	Benign-without-callback	Benign	0.0154
Sample 2	Benign	Malignant	Benign-without-callback	Benign-without-callback	0.0059
	Benign-without-callback	Malignant	Benign-without-callback	Benign-without-callback	0.0635
	Malignant	Malignant	Benign-without-callback	Benign-without-callback	-0.0207

Abbreviations: CNN: Convolutional neural network; SSIM: Structural similarity index; ViT: Vision transformer.

4. Conclusion

We propose an explainable deep learning framework for classifying breast cancer using mammographic images from the CBIS-DDSM dataset. The framework integrates state-of-the-art XAI methods, such as MCD, Grad-CAM,

and SHAP, with CNN and ViT architectures to improve interpretability, prediction, and uncertainty performance, and uses SSIM to quantitatively compare Grad-CAM heatmaps. Convolutional architectures are well suited to small medical imaging datasets, as the experiments show

that the CNN model's classification accuracy (60%) is better than the ViT model's (40%). XAI analyses showed that ViTs utilize distributed patch-based attention mechanisms, while CNN models attend to localized lesion regions. MCD captured the predictive uncertainty accurately, and SHAP provided pixel-level feature attributions. The proposed framework provided accurate classification and decision explanations that are understandable to humans, offering a transparent and reliable approach to breast cancer diagnosis.

The proposed work mainly focused on uncertainty estimation to study dropout behavior and overfitting in the available dataset, providing researchers with insights to improve diagnosis. However, the proposed work has a few limitations, such as its reliance on a small dataset with some class imbalance and the need for cross-validation to obtain more reliable performance estimates. The current experimental conditions evaluate the relative behavior rather than integrating into a diagnostic model with a larger, independent, and clinically representative dataset to classify the presence of benign and malignant breast cancer. Future work should address these limitations by using a larger dataset, repeated *k*-fold cross-validation, confidence interval estimation, and appropriate statistical significance tests (e.g., McNemar's test or bootstrap analysis) to support more robust, statistically grounded conclusions, as well as advanced augmentation strategies to enable a more balanced comparison. Additional methods such as saliency maps, LIME, integrated gradients, occlusion sensitivity, and transformer-specific techniques such as attention rollout, as well as non-parametric statistics, could provide further insights into the consistency and robustness of the learned representations. Future work should also address complementary explainability validation techniques, such as overlap metrics, faithfulness analyses, expert visual assessment, and explanation stability evaluations, to strengthen interpretation and provide a more balanced discussion of the SSIM results and their limitations.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Arthi Rengaraj

Data curation: Manoj Kumar Devendran

Formal analysis: Manoj Kumar Devendran

Methodology: Arthi Rengaraj

Software: Manoj Kumar Devendran

Validation: Praveenkumar Babu, Vijayan Thiruvengadam

Writing-original draft: Mahesh Kumar Nadesan

Writing-review & editing: Vijayan Thiruvengadam

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

All data were obtained from kagglehub/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset/versions/1

References

1. Saharan S, Wani NA, Chatterji S, Kumar N, Almuhaideb AM. A deep learning and explainable artificial intelligence based scheme for breast cancer detection. *Sci Rep.* 2025;15(1):32125. doi: 10.1038/s41598-024-80535-7
2. Karthiga R, Narasimhan K, Thanikaiselvan V, Hemalatha M, Amirtharajan R. Review of AI and XAI-based breast cancer diagnosis methods using various imaging modalities. *Multimed Tools Appl.* 2025;84(5):2209-2260. doi: 10.1007/s11042-024-20271-2
3. Fatima M, Zia R, Usmani IA. A hybrid framework integrating deep learning and XAI for breast cancer detection with stratified cross-validation. *Syst Soft Comput.* 2025;7:200412. doi: 10.1016/j.sasc.2025.200412
4. Arravalli T, Chadaga K, Muralikrishna H, et al. Detection of breast cancer using machine learning and explainable artificial intelligence. *Sci Rep.* 2025;15(1):26931. doi: 10.1038/s41598-025-12644-w
5. Talukder MA. An improved XAI-based DenseNet model for breast cancer detection using reconstruction and fine-tuning. *Results Eng.* 2025;26:104802. doi: 10.1016/j.rineng.2025.104802
6. Ahmmed J, Ahmed F, Kabir MA, et al. LBNet: an optimized lightweight CNN for mammographic breast cancer classification with XAI-based interpretability. *Sci Rep.* 2025;16(1):2013. doi: 10.1038/s41598-025-31642-6
7. Manojee KS, Kannan AR. Patho-Net: enhancing breast cancer classification using deep learning and explainable artificial intelligence. *Am J Cancer Res.* 2025;15(2):754-768.

- doi: 10.62347/XKFN1793
8. Singh SK, Patnaik KS. MammXAI: an XAI integrated adaptive multi-model deep learning approach for breast cancer detection using multi-modality images. *Biomed Signal Process Control*. 2026;113:109173.
doi: 10.1016/j.bspc.2025.109173
9. Tchokponhoue GAD, Idri A. On the value of uncertainty quantification in deep learning-based breast cancer molecular subtype classification. *Appl Soft Comput*. 2026;186:114249.
doi: 10.1016/j.asoc.2025.114249
10. Sabry M, Balaha HM, Ali KM, *et al*. AI-driven breast cancer diagnosis: a systematic review of imaging modalities, deep learning, and explainability. *Cancers*. 2026;18(8):1305.
doi: 10.3390/cancers18081305
11. Barseghyan S, Babajanyan A, Balyan Z, Asatryan A, Kumar S. Explainable AI for breast cancer risk prediction: evaluating the accuracy-explainability trade-off. *Mach Learn Comput Sci Eng*. 2025;1(2):25.
doi: 10.1007/s44379-025-00028-w
12. Bouallegue B, Abd El-Latif YM, El-Sofany H, Taj-Eddin IA, Vocaturo E. An enhanced approach for predicting breast cancer using different deep learning algorithms and explainable AI techniques in an IoT environment. *Int J Intell Syst*. 2025;2025(1):8884481.
doi: 10.1155/int/8884481
13. Ergün U, Çoban T, Kayadibi İ. BCECNN: an explainable deep ensemble architecture for accurate diagnosis of breast cancer. *BMC Med Inform Decis Mak*. 2025;25(1):374.
doi: 10.1186/s12911-025-03186-2
14. Ghasemi A, Hashtarkhani S, Schwartz DL, Shaban-Nejad A. Explainable artificial intelligence in breast cancer detection and risk prediction: a systematic scoping review. *Cancer Innov*. 2024;3(5):e136.
doi: 10.1002/cai2.136
15. Chai H, Lin S, Lin J, *et al*. An uncertainty-based interpretable deep learning framework for predicting breast cancer outcome. *BMC Bioinformatics*. 2024;25(1):88.
doi: 10.1186/s12859-024-05716-7
16. Hamedani-KarAzmoddehFar F, Tavakkoli-Moghaddam R, Tajally AR, Aria SS. Breast cancer classification by a new approach to assessing deep neural network-based uncertainty quantification methods. *Biomed Signal Process Control*. 2023;79:104057.
doi: 10.1016/j.bspc.2022.104057
17. Chegini M, Mahloojifar A. Uncertainty-aware deep learning-based CAD system for breast cancer classification using ultrasound and mammography images. *Comput Methods Biomech Biomed Eng Imaging Vis*. 2024;12(1):2297983.
doi: 10.1080/21681163.2023.2297983
18. Di Giammarco M, Vitulli C, Cirnelli S, *et al*. Explainable deep learning for breast cancer classification and localization. *ACM Trans Comput Healthc*. 2025;6(1):1-18.
doi: 10.1145/3702237
19. Alom MR, Farid FA, Rahaman MA, *et al*. An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images. *Sci Rep*. 2025;15(1):17531.
doi: 10.1038/s41598-025-97718-5
20. Murugan TK, Karthikeyan P, Sekar P. Efficient breast cancer detection using neural networks and explainable artificial intelligence. *Neural Comput Appl*. 2025;37(5):3759-3776.
doi: 10.1007/s00521-024-10790-2
21. Thiagarajan P, Khairnar P, Ghosh S. Explanation and use of uncertainty quantified by Bayesian neural network classifiers for breast histopathology images. *IEEE Trans Med Imaging*. 2021;41(4):815-825.
doi: 10.1109/TMI.2021.3123300
22. Fuentes-Fino R, Calderón-Ramírez S, Domínguez E, López-Rubio E, Elizondo D, Molina-Cabello MA. An uncertainty estimator method based on the application of feature density to classify mammograms for breast cancer detection. *Neural Comput Appl*. 2023;35(30):22151-22161.
doi: 10.1007/s00521-023-08904-3
23. López-Pérez M, Morales-Álvarez P, Cooper LA, Molina R, Katsaggelos AK. Deep Gaussian processes for classification with multiple noisy annotators: application to breast cancer tissue classification. *IEEE Access*. 2023;11:6922-6934.
doi: 10.1109/ACCESS.2023.3237990
24. Li J, Arefan D, Wu S. Uncertainty quantification for deep learning-based medical imaging classification model evaluation and individualized risk estimation. In: 2025 IEEE 7th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA). 2025:717-723.
doi: 10.1109/TPS-ISA67132.2025.00088
25. Kurz A, Hauser K, Mehrtens HA, *et al*. Uncertainty estimation in medical image classification: systematic review. *JMIR Med Inform*. 2022;10(8):e36427.
doi: 10.2196/36427
26. Tajally A, Zarean J, Bozorgi-Amiri A, Tavakkoli-Moghaddam R. Deep uncertainty quantification algorithms for confidence-aware hope classification of breast cancer patients based on their cognitive features. *Appl Soft Comput*. 2025;172:112860.
doi: 10.1016/j.asoc.2025.112860

27. Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE Jr, Burnside ES. Breast cancer risk estimation with artificial neural networks revisited: discrimination and calibration. *Cancer*. 2010;116(14):3310-3321.
doi: 10.1002/cncr.25081
28. Musah T, Kalaiwo C, Akram M, *et al.* Towards trustworthy breast tumor segmentation in ultrasound using Monte Carlo dropout and deep ensembles for epistemic uncertainty estimation. In: *Lecture Notes in Computer Science*. Springer Nature Switzerland; 2026:41-51.
doi: 10.1007/978-3-032-13654-1_5
29. Ali ZA, Sallow AB, Hassan MM. Bayesian deep learning with selective inference for uncertainty-aware breast ultrasound classification. *J Radiat Res Appl Sci*. 2026;19(2):102308.
doi: 10.1016/j.jrras.2026.102308
30. Zaheer Sajid M, Fareed Hamid M, Qureshi I. Explainable and uncertainty-aware ensemble framework with causal analysis for breast cancer detection. *Front Oncol*. 2025;15:1751090.
doi: 10.3389/fonc.2025.1751090
31. Joshi P, Dhar R. EpICC: a Bayesian neural network model with uncertainty correction for a more accurate classification of cancer. *Sci Rep*. 2022;12(1):14628.
doi: 10.1038/s41598-022-18874-6
32. Ünal S, Gürfidan R. Uncertainty-aware machine learning for predicting axillary lymph node metastasis using breast MRI. *BCTT*. 2026;18:1-19.
doi: 10.2147/bctt.s599211
33. Fooladgar F, Jamzad A, Connolly L, *et al.* Uncertainty estimation for margin detection in cancer surgery using mass spectrometry. *Int J Comput Assist Radiol Surg*. 2022;17(12):2305-2313.
doi: 10.1007/s11548-022-02764-3
34. Pflieger LT, Mason CC, Facelli JC. Uncertainty quantification in breast cancer risk prediction models using self-reported family health history. *J Clin Transl Sci*. 2017;1(1):53-59.
doi: 10.1017/cts.2016.9
35. Yao N, Hu H, Chen K, *et al.* A robust deep learning method with uncertainty estimation for the pathological classification of renal cell carcinoma based on CT images. *J Imaging Inform Med*. 2025;38(3):1323-1333.
doi: 10.1007/s10278-024-01276-7
36. Fordellone M, Chiodini P. Unsupervised hierarchical classification approach for imprecise data in breast cancer detection. *Entropy*. 2022;24(7):926.
doi: 10.3390/e24070926
37. Borquez S, Pezoa R, Salinas L, Torres CE. Uncertainty estimation in the classification of histopathological images with HER2 overexpression using Monte Carlo dropout. *Biomed Signal Process Control*. 2023;85:104864.
doi: 10.1016/j.bspc.2023.104864
38. Prabakaran I, Wu Z, Lee C, *et al.* Gaussian mixture models for probabilistic classification of breast cancer. *Cancer Res*. 2019;79(13):3492-3502.
doi: 10.1158/0008-5472.CAN-19-0573
39. Verboom SD, Kroes J, Broeders MJ, Sechopoulos I. Estimating deep learning model uncertainty of breast lesion classification to guide reading strategy in breast cancer screening. In: *Medical Imaging 2024: Computer-Aided Diagnosis*. Proceedings of SPIE; 2024;12927:467-472.
doi: 10.1117/12.3006713
40. Pocevičiūtė M, Eilertsen G, Jarkman S, Lundström C. Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology. *Sci Rep*. 2022;12(1):8329.
doi: 10.1038/s41598-022-11826-0