INSPIRE INTELLIGENCE

Inspire Intelligence Journal
ISSN: 3110-942X, Vol. 1, No. 1, 2026
doi.org/10.65718/inspireAI.2026.1002

InspireQuill.org
Inspire Academic Press (Canada)

Research Paper

Interpretable Deep Learning for Breast Lesion Classification: A SHAP-Based Comparison of CESM and Digital Mammography

Samara Acosta-Jiménez [1], Miguel M. Mendoza-Mendoza [1], Carlos E. Galván-Tejada [1,*], José M. Celaya-Padilla [1], J. Rubén Delgado-Contreras [1] and Jorge I. Galván-Tejada [1]

[1] Software Engineering Department, University of Zacatecas, Mexica
[*] Corresponding Author: ericgalvan@uaz.edu.mx

Abstract

Accurate detection of malignant breast lesions is fundamental for timely intervention and improved patient outcomes. Digital Mammography (DM) remains the standard imaging modality in most clinical settings. In contrast, Contrast-Enhanced Spectral Mammography (CESM) provides functional imaging data, which has the potential to improve the performance of deep learning algorithms in lesion detection. The study compares a ResNet-18 CNN trained on DM versus CESM for breast-lesion classification, uses 80% stratified 3-fold cross-validation, evaluates the best model on a held-out 20% blind test, and assesses interpretability with SHapley Additive exPlanations (SHAP) attribution maps. On the blind test set, CESM achieved superior overall classification performance, with an AUC of 0.890 for malignant lesions compared to 0.790 for DM. The micro- and macro-averaged AUCs are 0.83 and 0.77 for CESM, respectively, compared to 0.68 and 0.68 for DM. SHAP analysis revealed that CESM yielded more focused and anatomically aligned model attributions, particularly in malignant cases. DM produced more diffuse, yet still lesion-centered, attribution patterns. CESM substantially improves CNN-based diagnostic accuracy and interpretability in breast lesion classification, with the most significant benefit observed in the detection of malignant lesions, which is the most clinically significant category. Although it is still challenging to classify benign lesions, DM can help identify malignancies without the need for contrast, making it useful for screening when CESM is not available. These results show the value of CESM-enhanced Artificial Intelligence (AI) systems and explainability tools for dependable use in clinics.

Keywords: Breast cancer; Convolutional neural network; SHAP; Digital mammography; Contrast-enhanced spectral mammography

1. Introduction

Breast cancer continues to be a major contributor to cancer-related morbidity and mortality worldwide, with an estimated 2.3 million new cases diagnosed each year and more than 685,000 deaths reported in 2020 alone [1]. Early detection is essential for improving patient outcomes, informing clinical decisions, and reducing mortality rates. To date, digital mammography (DM) is widely accessible and reasonably priced, it is used as the main

Academic Editor: Ghazanfar Latif

Received: 13/05/2025
Revised: 10/07/2025
Accepted: 06/11/2025
Published: 29/12/2025

Citation
Acosta-Jiménez, S., Mendoza-Mendoza, M. M., Galván-Tejada, C. E., Celaya-Padilla, J. R. D.-C., & Galván-Tejada, J. I. (2026). Interpretable deep learning for breast lesion classification: A SHAP-based comparison of CESM and digital mammography. *Inspire Intelligence Journal, 1*(1), 10-28.

technique in population-based screening programs [2]. However, there are drawbacks to traditional DM, particularly in women with dense breast tissue, where overlapping fibroglandular structures may mask subtle lesions, decreasing sensitivity and raising the possibility of false negative results [3, 4]. In response to these challenges, advanced imaging techniques such as Contrast-Enhanced Spectral Mammography (CESM) have been developed. Dual-energy acquisition using low and high energy x-ray spectra following intravenous iodinated contrast injection provides both anatomic and functional information in the same examination [5, 6]. Several studies have shown that CESM provides better diagnostic sensitivity and accuracy than DM, especially for preoperative staging and in dense breasts [7, 8].

Even with these advancements in technology, accurately interpreting mammograms is still a challenging task. Factors such as radiologists' workload and the subtle presentation of early-stage tumors can impact radiologists' diagnostic performance [9]. The growing clinical workloads highlight the necessity of computational tools to facilitate precise and effective diagnosis.

Convolutional Neural Networks (CNNs) and other Deep Learning (DL) techniques have demonstrated encouraging results in recent years on a variety of Artificial Intelligence (AI) tasks, including medical image analysis [10, 11]. CNN-based methods in breast imaging show great efficacy in identifying, classifying, and evaluating lesions, consistently yielding accurate and repeatable findings [12, 13]. Of these topologies, ResNet-18 strikes a balance between computational efficiency and representational capacity. The structure reduces overfit-ting risks in limited medical datasets, while maintaining sufficient capacity for com-plex pattern recognition. Furthermore, its low computational cost facilitates clinical integration, where both performance and transparency are critical. However, the opaque decision-making process of deep learning models is a major obstacle to clinical use [14]. Explainable artificial intelligence (XAI) techniques have been developed to improve model transparency in order to tackle this difficulty. SHapley Additive ExPlanations (SHAP) is one approach for analyzing model predictions [15]. SHAP quantifies each input feature's contribution to the model's output and offers both local and global explanations. This enables more rigorous clinical validation of AI-generated decisions.

This study investigates whether, despite the enhanced lesion visibility provided by CESM, a ResNet-18 CNN model trained on conventional DM can still detect subtle yet clinically meaningful patterns. Through SHAP analysis, the aim to demonstrate that attribution maps derived from DM, although potentially less pronounced, maintain a degree of concordance with the key diagnostic regions highlighted in CESM images. This comparison intends to uncover interpretable features learned by the AI from both imaging modalities, highlighting the potential of AI to extract valuable diagnostic information even from more accessible imaging techniques.

The primary objective of this study is to perform a comparative SHAP-based analysis of paired DM and CESM images from the same patients using a ResNet-18 CNN model. The aim to identify shared and distinct attribution patterns across the two modalities, assessing whether DM-based SHAP maps reflect similar clinically relevant features as those observed in CESM. This analysis offers a deeper understanding of AI decision, making processes and underscores the value of interpretable AI tools in resource, limited settings where advanced imaging is not widely available.

2. Materials and Methods

This study evaluates and compares the interpretability and classification performance of a CNN model for breast lesion classification using two mammographic modalities: DM and CESM. The same architecture is independently trained and tested on each modality to ensure a fair comparison of performance.

The methodological pipeline is illustrated in Figure 1. It comprises image acquisition, preprocessing, data augmentation, model training, cross-validation, blind test evaluation, and interpretability analysis using SHAP.

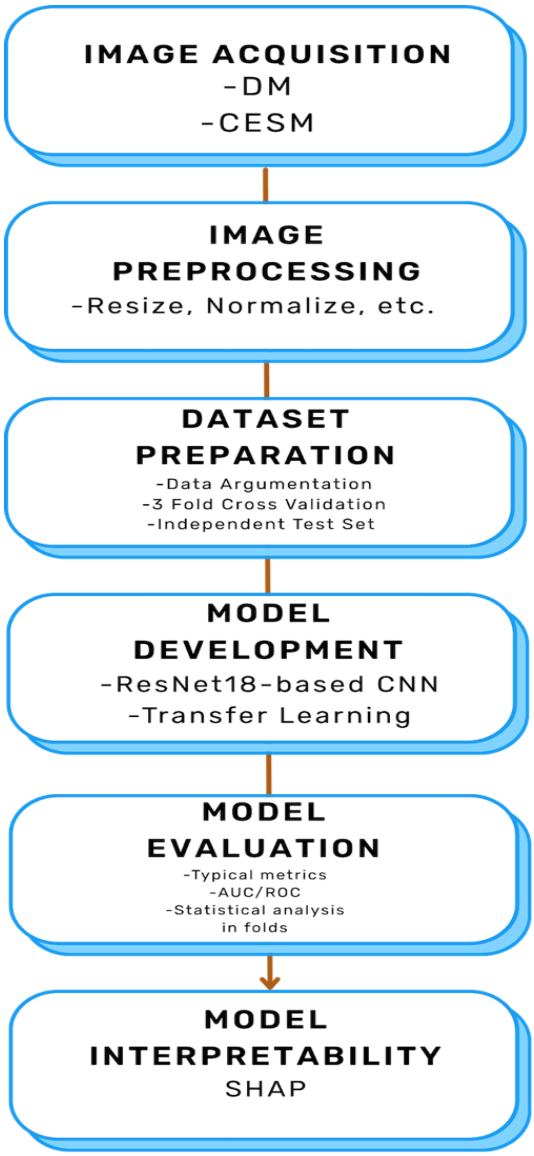

Figure 1. General workflow of the study: from DM and CESM image acquisition to model development, evaluation, and SHAP-based interpretability analysis.

2.1. Image Acquisition

The dataset used is publicly available from The Cancer Imaging Archive (TCIA), specifically the CDD-CESM collection [16]. It comprises images from 326 female patients aged 18 to 90, each undergoing both DM and CESM. Images were acquired using GE Senographe DS and Hologic Selenia Dimensions systems.

A total of 2,006 images are included: 1,003 low-energy (DM) and 1,003 CESM images. Standard craniocaudal (CC) and mediolateral oblique (MLO) views are provided, although not all views are present for each patient. All images are in JPEG format, 3-channel RGB, with an average resolution of 2355×1315 pixels. They are labeled as normal, benign , or malignant based on histopathology or expert annotation.

Table 1 summarizes the image distribution across the three diagnostic classes for each modality, allowing for a direct comparison of CNN performance on DM and CESM data.

Table 1. Number of images by class and modality.

Class	DM Images	CESM Images	Total
Normal	341	416	757
Benign	331	256	587
Malignant	331	331	662
Total	**1003**	**1003**	**2006**

2.2. Image Preprocessing

The preprocessing steps are summarized in Table 2. Data augmentation is only applied to training images, including random brightness/contrast changes, flips, and rotations, to improve robustness and prevent overfitting. The images are resized to 224×224 pixels and normalized with ImageNet statistics. Validation and test sets are only resized and normalized. Figure 2 shows an example of data augmentation applied to one mammography from each class.

Table 2. Image preprocessing transformations.

Transformation	Description	Parameters
Resize	Resize image to uniform input size	(224,224)
ColorJitter	Randomly alters brightness and contrast	Brightness: 0.05; Contrast: 0.15
RandomHorizontalFlip	Randomly flips image horizontall	-
RandomVerticalFlip	Randomly flips image vertically	-
RandomRotation	Rotates image within a fixed degree range	20
ToTensor	Converts image to PyTorch tensor	-
Normalize	Normalizes the image with ImageNet mean and std	Mean: [0.485, 0.456, 0.406]; Std: [0.229, 0.224, 0.225]

2.3. Model Architecture and Training

A ResNet18 model pretrained on ImageNet is used. Its final layer is replaced to output three classes: normal, benign, and malignant. The dataset is split at the patient level into 80% for training/validation and 20% for a blind test. Within the 80%, stratified 3-fold cross-validation is performed.

The best-performing model from each fold, as determined by the lowest validation loss, is saved during cross-validation. During the validation phases, the model with the highest macro-averaged F1-score is chosen for the blind test set final evaluation. Inverse frequency weighting is employed in the categorical cross-entropy loss function to address the class imbalance in the dataset and ensure minority classes are adequately represented during training. The model is optimized using the Adam algorithm with an initial learning rate of 1x10-3, and a step-based learning rate scheduler (StepLR) is employed to reduce the learning rate by a factor of 0.1 every 10 epochs. Each fold is trained for a total of 30 epochs using a mini-batch size of 8. A detailed summary of the training configuration and hyperparameters is provided in Table 3.

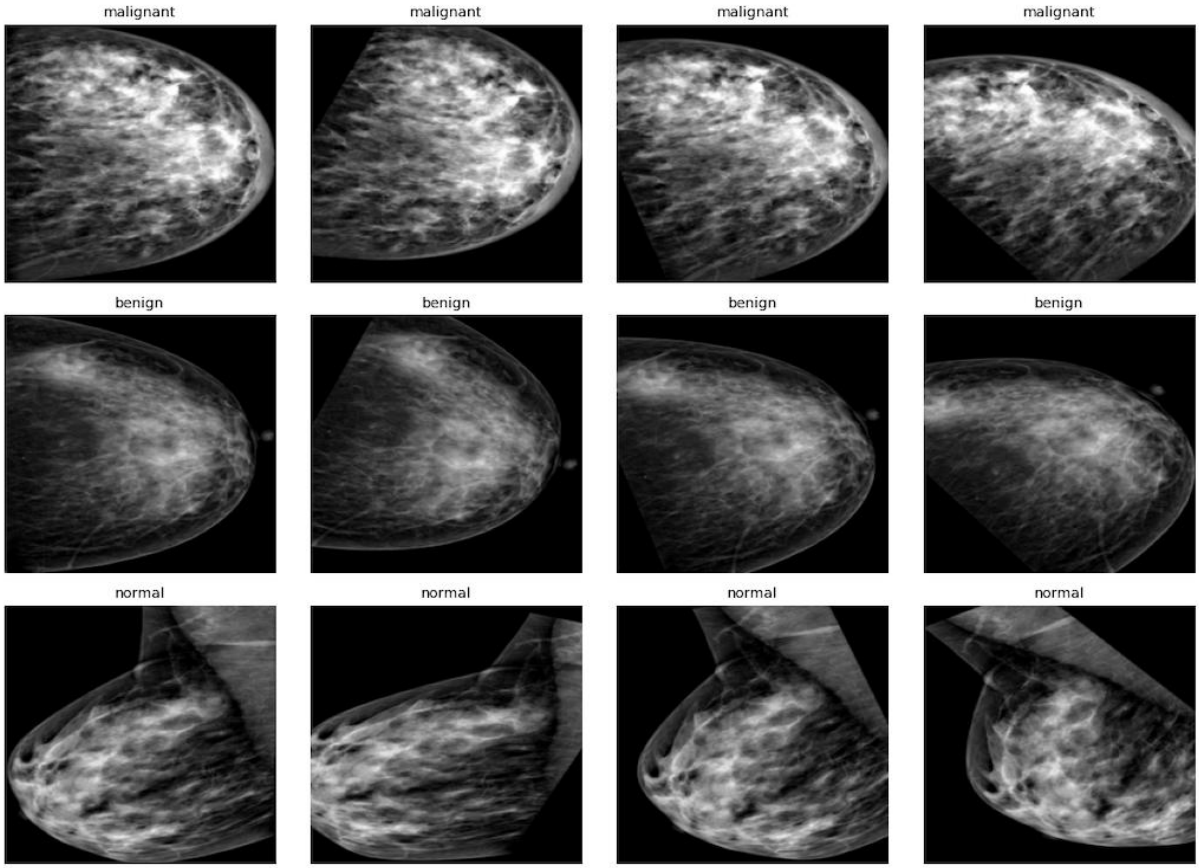

Figure 2. Example of data augmentation applied to breast mammography images. Each row corresponds to a representative case from the malignant, benign, and normal classes, respectively. From left to right: original image followed by three augmented versions generated through rotations.

Table 3. Training configuration.

Configuration	Value
Architecture	*ResNet18*
Cross-validation	*3-fold*
Loss function	*Weighted Cross-Entropy*
Optimizer	*Adam*
Initial learning rate	$1x10^{-3}$
Learning rate scheduler	*StepLR (step=10, gamma=0.1)*
Batch size	*8*
Epoch	*30*
Class balancing	*Inverse class frequency*

2.4. Evaluation Metrics

The performance of the proposed model is assessed using a comprehensive set of classification metrics, including accuracy, precision, recall (sensitivity), specificity, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). These metrics are computed for each fold during cross-validation as well as for the final evaluation on the blind test set.

$$\text{Accuracy} = TP + TN/TP + TN + FP + FN \quad (1)$$

$$\text{Precision} = TP/TP + FP \quad (2)$$

$$\text{Recall} = TP/TP + FN \quad (3)$$

$$\text{F1-score} = 2\ (\text{Precision} \times \text{Recall}) / \text{Precision} + \text{Recall} \quad (4)$$

Accuracy provides a general measure of the model's correctness by quantifying the proportion of true results among the total number of evaluated cases. Precision and recall offer insights into the model's performance regarding positive predictions, with recall emphasizing sensitivity to true positives and precision reflecting the reliability of positive predictions. The F1-score, as the harmonic mean of precision and recall, provides a balanced metric, especially useful in imbalanced datasets. Additionally, AUC-ROC is used to evaluate the model's discriminative ability across all classification thresholds by analyzing the trade-off between the true positive rate and the false positive rate [17].

2.5. Statistical Analysis

To robustly compare the classification performance between DM and CESM modalities, a comprehensive statistical analysis is conducted on the evaluation metrics.

For the cross-validation results, the mean and standard deviation of each metric (accuracy, precision, recall, F1-score, and AUC values) are computed across the three folds for both DM and CESM. A paired t-test (t-test for dependent samples) is used to assess the statistical significance of the differences in means between CESM and DM for each metric. Additionally, 95% confidence intervals (CIs) for the mean differences are computed to provide a range of plausible values for the true difference. A p-value of less than 0.05 ($p < 0.05$) is considered statistically significant. For the final evaluation on the independent test set, a direct comparison of the individual metrics for DM and CESM is performed. Given that only a single evaluation is conducted on this dataset, no statistical significance tests are applied to the test set metrics. The performance metrics and AUC values for both modalities are reported for observation.

All statistical analyses and data visualizations are performed using Python (version 3.9.6). Key libraries included NumPy (version 1.26.4), SciPy (version 1.13.1), Matplotlib (version 3.9.0), PyTorch (version 2.3.1), SHAP (version 0.45.1), and scikit-learn (version 1.5.0).

2.6. Model Interpretability

SHAP is used to address model interpretability in order to improve transparency and foster clinical trust in deep learning predictions. A game-theoretic framework called SHAP allows for a principled attribution of prediction responsibility by quantifying the contribution of each input feature to the model's output. In this study, the GradientExplainer from the SHAP library is employed, which is suitable for interpreting deep neural networks based on gradient information. Pixel-wise SHAP values are computed for selected test images to visualize local attributions. The entire training dataset is used as the background dataset for estimating the expected model output. Each SHAP value reflects the impact of a given input pixel on the model's confidence for a specific class: positive values indicate a contribution toward that prediction, while negative values suggest suppression of that class. The resulting SHAP values are visualized as heatmaps, in these visualizations, red regions represent features that strongly support a specific classification, while blue regions indicate features that contradict the classification. This method facilitates interpretation of individual predictions and assists in identifying clinically relevant regions, including lesions or abnormal tissue patterns in mammography images. For quantitative comparison across modalities, the raw pixel-wise SHAP values generated for paired DM and CESM images are directly analyzed.

This involves assessing the magnitude, distribution, and spatial localization of these SHAP values, particularly in relation to diagnostic findings. This allows for a direct numerical understanding of how the AI model attributes importance to specific image features in each modality and enables a structured comparison of these attribution patterns.

3. Results

3.1 Cross-validation Performance

The classification performance of the CNN model is evaluated using 3-fold cross-validation, applied independently to the DM and CESM datasets. As presented in Table 4, CESM consistently outperformed DM across all folds in terms of accuracy, precision, recall, and F1-score. On average, CESM achieved an accuracy of 0.63238 compared to 0.5233 for DM. The average F1-score increased from 0.5055 for DM to 0.5756 for CESM.

Table 4. Cross-validation performance metrics per fold for DM and CESM.

Fold	DM				CESM			
	Accuracy	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
0	0.5033	0.4950	0.5057	0.4910	0.6325	0.5914	0.5909	0.5866
1	0.5449	0.5408	0.5607	0.5274	0.6246	0.5658	0.5908	0.5655
2	0.5216	0.5002	0.5169	0.4981	0.6412	0.5914	0.5772	0.5748
Average	**0.5233**	**0.5120**	**0.5278**	**0.5055**	**0.6328**	**0.5829**	**0.5863**	**0.5756**

Table 5 reports the class-wise AUC values per fold. CESM consistently achieved higher AUCs across all classes and folds. The most notable improvements are observed in the malignant and normal categories. For example, the average AUC for malignant lesions increased from 0.77 with DM to 0.88 with CESM, and for normal cases, from 0.74 to 0.82. In contrast, the benign class showed a more modest improvement, from 0.57 to 0.59, suggesting that this class remains more difficult to distinguish, likely due to overlapping features with normal tissue.

Table 5. AUC values per class and fold for DM and CESM.

Fold	DM			CESM		
	Benign	**Malignant**	**Normal**	**Benign**	**Malignant**	**Normal**
0	0.57	0.76	0.75	0.63	0.85	0.84
1	0.57	0.76	0.75	0.59	0.89	0.82
2	0.58	0.79	0.72	0.56	0.91	0.82
Average	**0.57**	**0.77**	**0.74**	**0.59**	**0.88**	**0.82**

Figures 3 and 4 present the ROC curves for each fold and class using CESM and DM, respectively. These visualizations illustrate the model's discriminative performance across all thresholds. The CESM-based model demonstrates a clear advantage, particularly in the malignant class, where AUCs ranged from 0.85 to 0.91 across folds, compared to 0.76 to 0.79 for DM. Substantial improvements are also observed for the normal class, while the benign class yielded consistent but smaller gains.

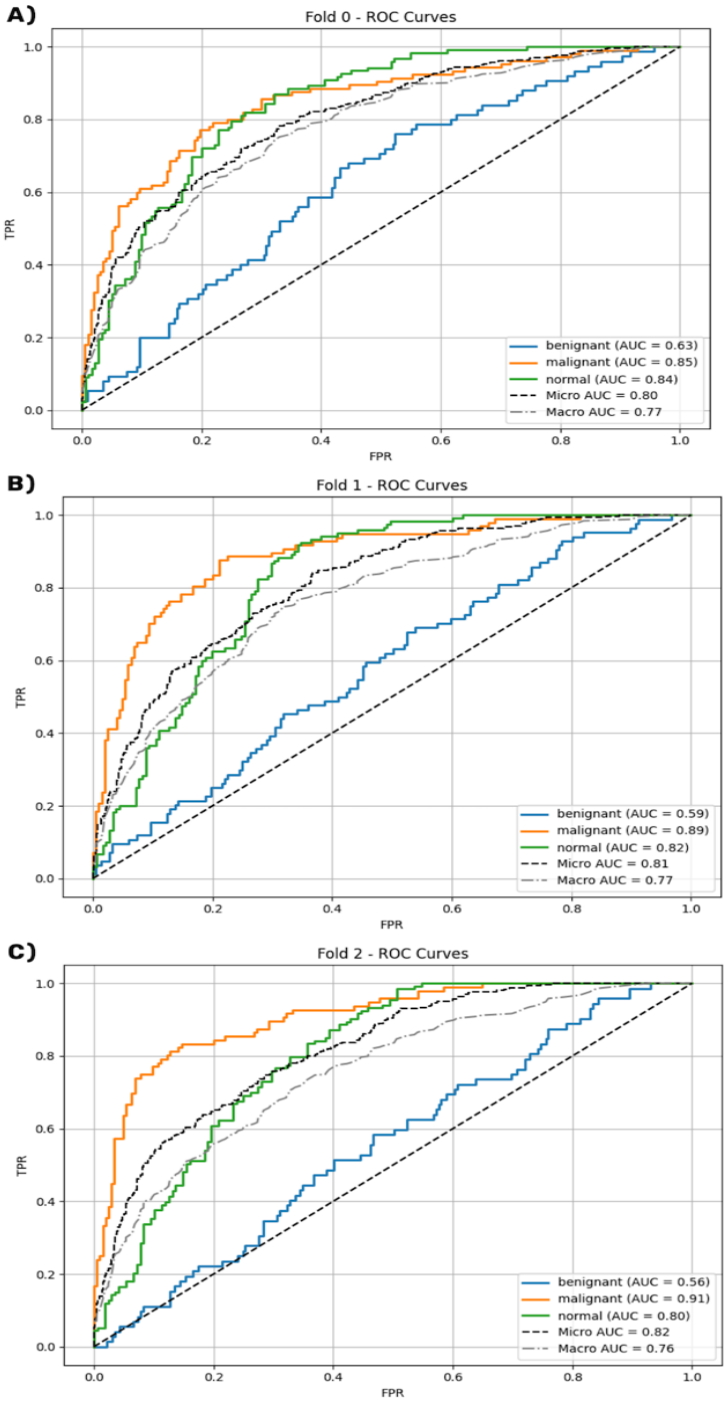

Figure 3. ROC curves for CESM across all three cross-validation folds. Each subfigure (A, B, C) corresponds to one fold and shows class-wise performance with associated AUC values. The curves highlight consistently superior discriminative performance, particularly for malignant and normal classes.

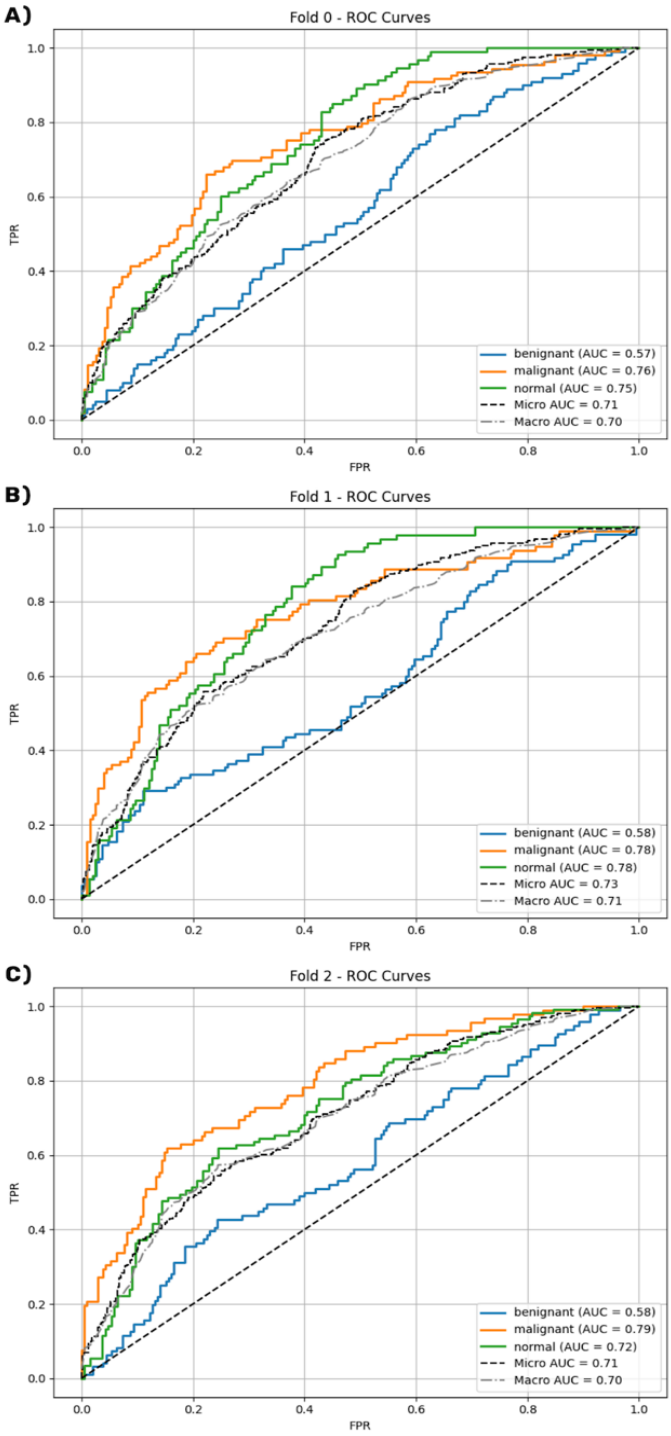

Figure 4. ROC curves for DM across all three cross-validation folds. Each subfigure (A, B, C) corresponds to one fold and displays class-wise performance with associated AUC values. While demonstrating reasonable classification performance, the DM-based model exhibits lower discriminative capacity, especially in malignant case detection.

3.2 Statistical Comparison Between Modalities

To assess the statistical significance of the observed performance differences between DM and CESM, a comparative analysis is conducted using paired t-tests across cross-validation folds. The results are summarized in Tables 6 and 7.

Table 6. CNN performance comparison between DM and CESM modalities.

Metric	DM	CESM	Difference	95% CI	p-value
Accuracy	0.523 ± 0.021	0.633 ± 0.008	+0.110 (+20.9%)	[0.044, 0.175]	**0.019**
Precision	0.512 ± 0.025	& 0.583 ± 0.015	+0.071 (+13.8%)	[-0.028, 0.170]	0.091
Recall	0.528 ± 0.029	0.586 ± 0.008	+0.059 (+11.1%)	[-0.010, 0.127]	0.067
F1-score	0.505 ± 0.019	0.576 ± 0.011	+0.070 (+13.9%)	[-0.003, 0.143]	0.054

Table 7. Comparison of class-wise AUC values between DM and CESM modalities.

Metric	DM	CESM	Difference	95% CI	p-value
Benign	0.573 ± 0.006	0.593 ± 0.035	+0.020 (+3.5%)	[-0.079, 0.119]	0.478
Malignant	0.770 ± 0.017	0.883 ± 0.031	+0.113 (+14.7%)	[0.062, 0.165]	**0.011**
Normal	0.740 ± 0.017	0.820 ± 0.020	+0.080 (+10.8%)	[0.055, 0.105]	**0.005**

Regarding global classification metrics, CESM yielded a statistically significant improvement in overall accuracy ($p = 0.019$). Improvements in F1-score ($p = 0.054$), recall ($p = 0.067$), and precision ($p = 0.091$) demonstrated positive trends but did not reach conventional significance thresholds. These results suggest a consistent advantage for CESM, though the limited number of folds (n=3) may affect statistical power.

In terms of class-wise AUC values, CESM significantly outperformed DM in the classification of malignant ($p = 0.011$) and normal cases ($p = 0.005$), highlighting the enhanced discriminative capacity of contrast-enhanced imaging in clinically relevant categories. No significant difference is observed for the benign class ($p = 0.478$), suggesting that this category remains challenging across modalities, likely due to its visual overlap with normal tissue features.

3.3 Independent Test Set Evaluation

The final evaluation on the independent blind test set confirmed the superior performance of CESM over conventional DM across all evaluation metrics, as summarized in Table 8. CESM achieved a markedly higher accuracy of 0.66, compared to 0.51 for DM, representing a 29.4% relative improvement. This substantial difference is consistent with the cross-validation trends, supporting the robustness and generalizability of the model's performance. CESM also outperformed DM in precision, recall, and F1-score, with relative improvements of 18.0%, 22.4%, and 16.0%, respectively. These results indicate that CESM not only improves overall classification but also yields a more balanced trade-off between sensitivity and specificity across all diagnostic categories.

Table 8. Detailed performance metrics of the CNN model on the test set for each modality

Fold	DM			CESM		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Benign	0.28	0.36	0.32	0.36	0.16	0.22
Malignant	0.57	0.53	0.53	0.76	0.85	0.80
Normal	0.67	0.59	0.62	0.65	0.80	0.72
Average	**0.50**	**0.49**	**0.50**	**0.59**	**0.60**	**0.58**
Accuracy			**0.51**			**0.66**

AUC analysis further reinforced these findings (Table 9). The malignant class exhibited the most pronounced improvement, with CESM achieving an AUC of 0.890 versus 0.790 for DM, corresponding to a clinically significant 12.7% relative gain in discriminative ability. The normal class also improved substantially (from 0.710 to 0.830; +16.9%), while the benign class showed a more modest increase (from 0.510 to 0.570; +11.8%), confirming its continued classification difficulty.

Table 9. AUC values by class for DM and CESM on the test set.

Modality	Benign	Malignant	Normal
DM	0.510	0.790	0.710
CESM	0.570	0.890	0.830

Figures 5 and 6 present the ROC curves and confusion matrices for the test set evaluation, illustrating the performance of CESM and DM, respectively. The superior discriminative capacity of CESM is clearly evident, with curves positioned closer to the upper-left corner of the ROC space across all classes. The malignant class demonstrates the most dramatic improvement, with CESM achieving a steep initial rise in the ROC curve compared to the more gradual ascent observed in DM. This pattern indicates enhanced sensitivity at lower false positive rates, which is particularly valuable for cancer screening applications. In other hand, the confusion matrices provide additional insights into the classification patterns, CESM, demonstrates superior performance in correctly identifying malignant cases, with 28 out of 33 malignant samples correctly classified (85% sensitivity), compared to only 17 out of 32 malignant cases (53% sensitivity) for DM. This represents a substantial improvement in the most clinically critical category. For normal tissue classification, CESM correctly identified 33 out of 41 normal cases (80% sensitivity), while DM achieved 24 out of 41 (59% sensitivity). However, the benign class presented consistent challenges for both modalities. CESM correctly classified only 4 out of 25 benign cases (16% sensitivity), while DM achieved 9 out of 25 (36% sensitivity). This counterintuitive result suggests that while CESM enhances contrast for malignant lesions, it may also accentuate features in benign lesions that make them appear more suspicious, leading to increased misclassification as malignant. This finding has important clinical implications, as it suggests that CESM-based automated systems may require additional refinement or complementary diagnostic approaches for optimal benign lesion characterization.

The micro-averaged and macro-averaged AUC values (0.83 and 0.77 for CESM vs. 0.68 and 0.68 for DM, respectively) further confirm the overall superior performance of CESM, particularly when considering the clinical importance of accurate malignant detection. These tests set results validate the cross-validation findings and confirm that CESM provides superior diagnostic input for CNN-based classification systems, especially for the critical task of malignant lesion detection.

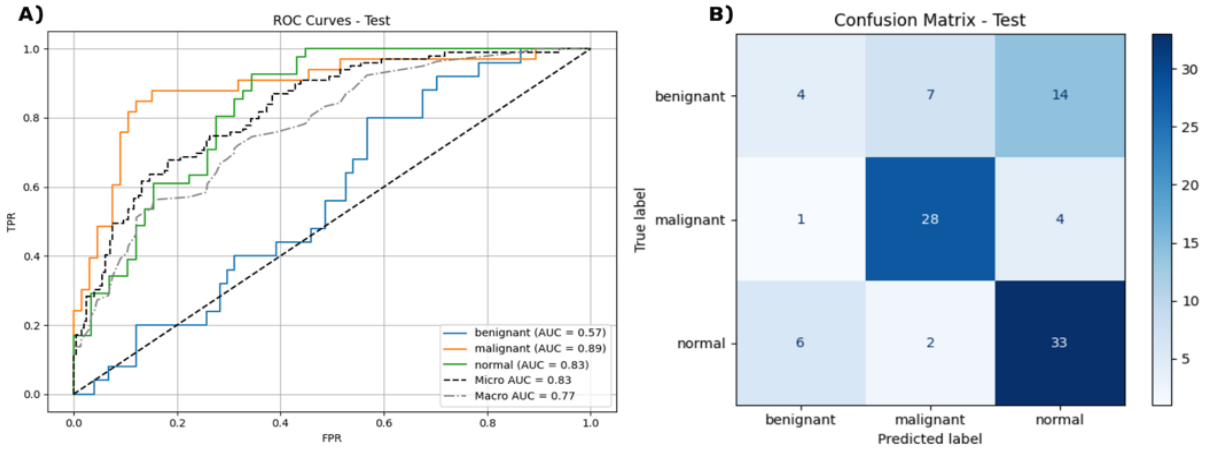

Figure 5. Test set evaluation results for CESM. (A) ROC curves show AUC values of 0.57 (benign), 0.89 (malignant), and 0.83 (normal). Micro-averaged AUC = 0.83; macro-averaged AUC = 0.77. (B) Confusion matrix shows high performance for malignant (28/33, 85%) and normal (33/41, 80%) classifications, with limited performance for benign (4/25, 16%).

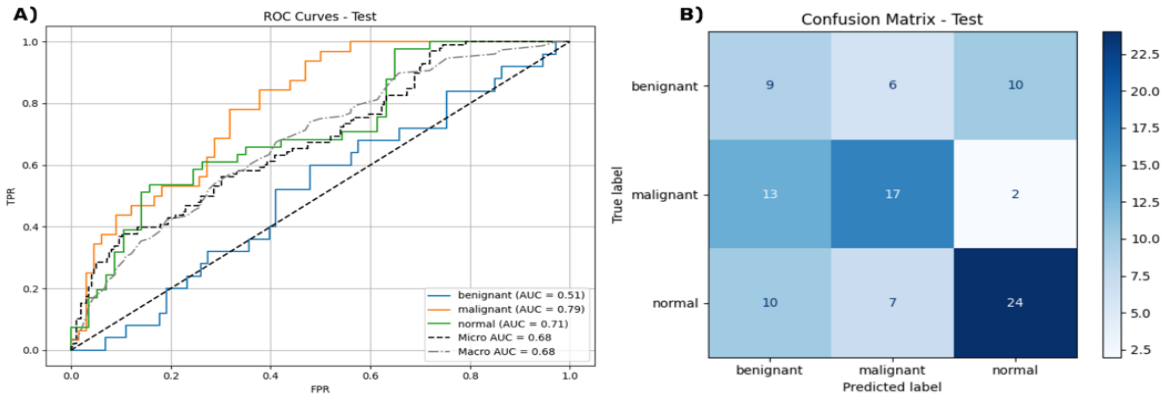

Figure 6. Test set evaluation results for DM. (A) ROC curves show AUC values of 0.51 (benign), 0.79 (malignant), and 0.71 (normal). Both micro- and macro-averaged AUC = 0.68. (B) Confusion matrix reflects moderate performance for malignant (17/32, 53%) and normal (24/41, 59%) classes, with relatively better performance for benign (9/25, 36%).

3.4 Model Interpretability

To enhance transparency and support clinical trust in deep learning predictions, model interpretability is investigated SHAP. SHAP is a game-theoretic approach that quantifies the contribution of each input feature to the model's output, allowing principled attribution of predictions. In this study, the GradientExplainer from the SHAP library is employed, which is suitable for interpreting deep neural networks through gradient-based analysis. Pixel-wise SHAP values are computed for selected test images to visualize local attributions. The entire training dataset served as the background for estimating expected model outputs. Positive SHAP values indicate that a given pixel contributes to a prediction, whereas negative values suppress that prediction. These values are visualized as heatmaps: red regions signify positive contributions, while blue regions denote inhibitory effects. This method provides intuitive explanations of individual predictions, highlighting relevant anatomical areas such as lesions or abnormal tissue patterns. For comparison between modalities, SHAP values from paired DM and

CESM images are directly analyzed, focusing on their magnitude, spatial distribution, and anatomical correspondence to diagnostic findings.

Figures 7, 8, and 9 present these case-specific SHAP visualizations by diagnostic category. Figure 7 shows attribution maps for two representative malignant cases. In subfigure A), CESM-based SHAP maps exhibit highly localized positive attributions (red areas) in the malignant column that align precisely with the lesion, reaching SHAP values of ±0.03. The normal column shows strong negative values in the same region, indicating suppression of non-pathological interpretation. In contrast, the DM-based map for the same patient shows more diffuse attributions with lower intensity (±0.04), yet it maintains similar spatial focus on the lesion area. Subfigure B) further supports this pattern: CESM produces sharp, anatomically aligned activations, while DM shows more scattered and less organized patterns. These results suggest that CESM enhances the model's ability to form more confident and radiologically interpretable features, consistent with its superior quantitative performance.

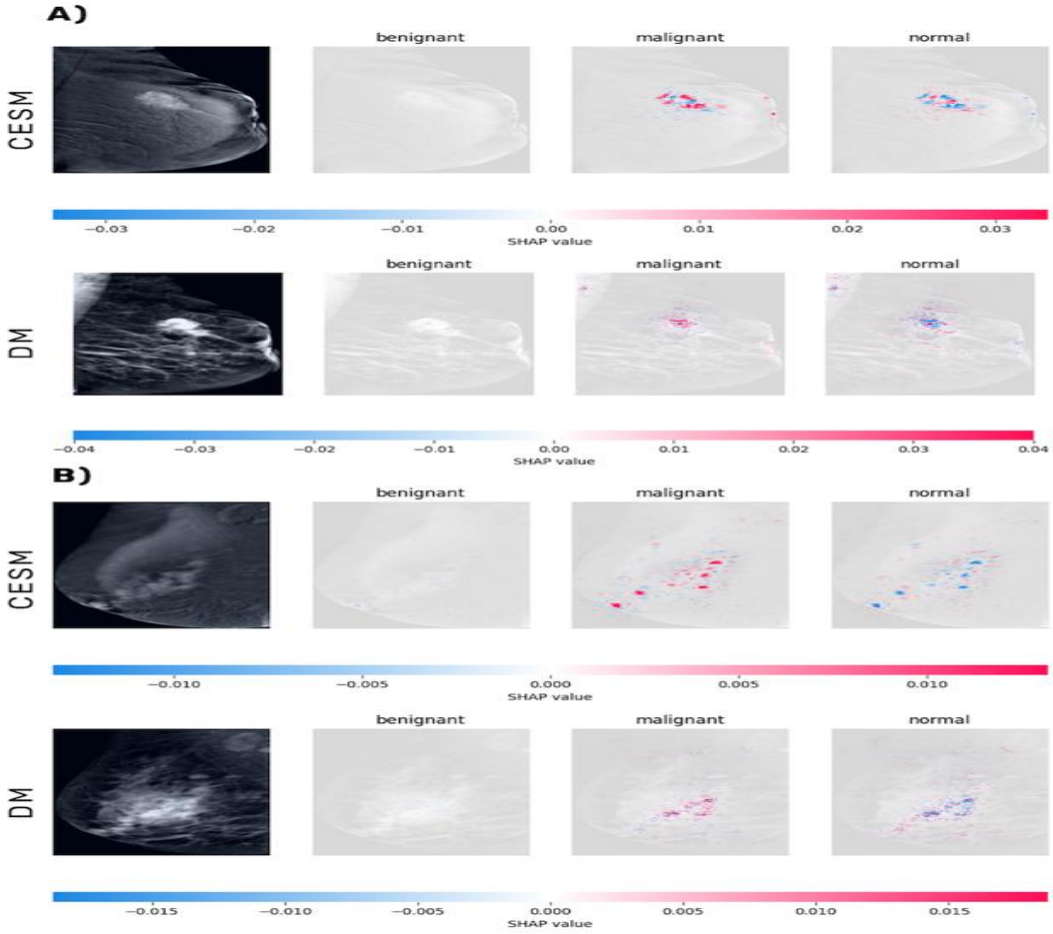

Figure 7. SHAP attribution maps for two representative malignant cases (A and B). Each row displays the original CESM and DM images, followed by SHAP maps for each predicted class (benign, malignant, and normal). In both cases, CESM-based SHAP maps show well-localized and intense positive attributions (red areas) within the lesion, particularly in the malignant class, suggesting the model correctly focuses on areas of enhancement. DM-based SHAP maps reveal more diffuse yet lesion-centered attributions, indicating the model is still able to detect structural tumor patterns in the absence of contrast.

Figure 8 displays attribution maps for two representative normal cases. In subfigure A), CESM exhibits low, balanced attributions across classes, with SHAP values up to ±0.010, reflecting appropriate diagnostic uncertainty. DM shows slightly higher, but more diffuse, SHAP values (±0.020), yet retains similar anatomical distribution. Subfigure B) mirrors this: CESM produces controlled and organized SHAP values (±0.006), while DM again shows scattered and stronger activations (±0.03). In both cases, the absence of focal attribution confirms appropriate model behavior in normal tissue. These findings illustrate that CESM enhances feature stability and interpretability even in the absence of pathology.

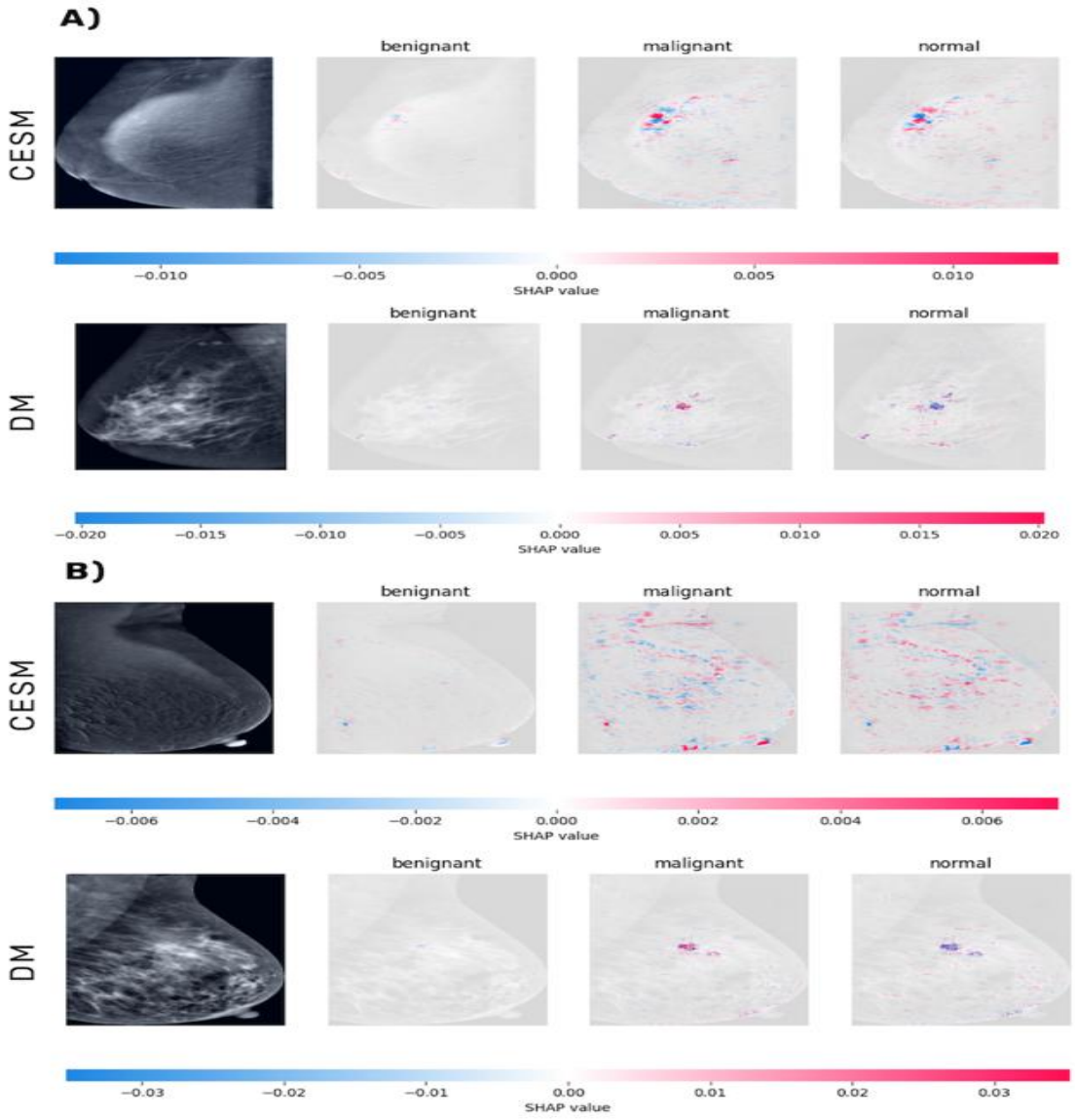

Figure 8. SHAP attribution maps for two representative normal cases (A and B). Each row presents CESM and DM images along with SHAP maps for the three diagnostic classes. Both modalities show low-magnitude and diffuse SHAP values, with no dominant focal regions, reflecting the absence of pathological findings. CESM maps are slightly more balanced, whereas DM shows more scattered attributions. These patterns confirm appropriate model behavior in normal cases, where low diagnostic confidence and non-focal attention are expected.

In contrast, Figure 9 highlights attribution maps for two benign cases. In subfigure A), CESM shows focal activations in both the malignant and normal columns (±0.015), suggesting confusion between these categories. DM shows more intense but diffuse activations ±0.020), also cantered on the lesion. Subfigure B) presents similar results: CESM shows localized SHAP values (±0.015), while DM produces scattered patterns of comparable magnitude. In both cases, the model fails to consistently attribute benign lesions to the correct class, instead showing conflicting signals. This likely reflects the morphological ambiguity of benign lesions, and in CESM, the enhanced contrast may inadvertently amplify features resembling malignancy, leading to misclassification. These insights explain the counterintuitive result that CESM performed worse than DM for benign lesions and highlight the need for further refinement in this diagnostic category.

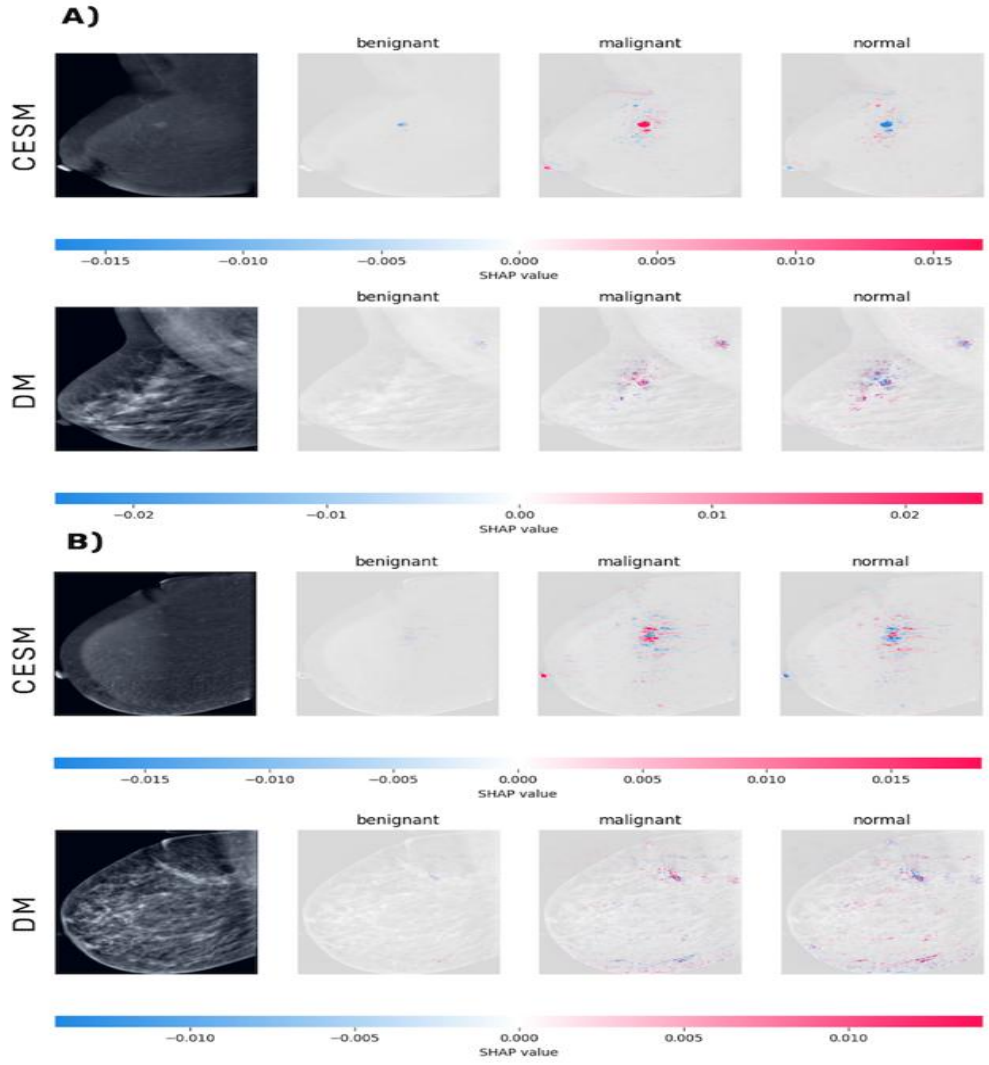

Figure 9. SHAP attribution maps for two representative benign cases (A and B). Each row shows CESM and DM images followed by SHAP maps for each class. Both modalities demonstrate overlapping and ambiguous SHAP attributions, particularly between the malignant and benign classes. CESM-based SHAP maps typically show focal activations, which the model may misinterpret due to mild contrast enhancement. DM maps are more dispersed but have similar intensity. These patterns emphasize the inherent difficulty in distinguishing benign lesions, which explains the model's poor classification performance in this class.

4. Discussion

This study presents a comparative evaluation of CNN-based breast lesion classification using Digital Mammography DM and CESM, emphasizing both quantitative performance and model interpretability through SHAP-based attribution analysis. The findings consistently demonstrate the superior diagnostic performance of CESM across multiple evaluation metrics, while also uncovering interpretability patterns that shed light on model behavior and limitations.

Quantitative results from stratified 3-fold cross-validation indicate that CESM out-performs DM in all major metrics, including accuracy, precision, recall, and F1-score. These improvements are clinically meaningful and statistically significant for overall accuracy ($p = 0.019$) and show a strong trend toward significance for F1-score and recall. Additionally, CESM demonstrates significantly higher AUCs for malignant and normal categories ($p = 0.011$ and $p = 0.005$, respectively), confirming its enhanced discriminative capability. These trends persist in the independent test set, where CESM achieves a 29.4% high-er accuracy than DM (0.660 vs. 0.510) and improves both recall and F1-score by over 20%.

These performance gains are especially important for detecting malignant lesions, the most clinically significant classification in breast cancer screening. CESM achieved an AUC of 0.890 for malignant cases in the test set, compared to 0.790 for DM, and showed significantly higher sensitivity (85% vs. 53%). This improvement highlights the clinical value of contrast enhancement in identifying aggressive lesions earlier and supports CESM's integration into AI-assisted screening programs. This improvement highlights the clinical value of contrast enhancement for early detection of biologically aggressive lesions and supports integrating CESM into AI-assisted screening programs [5, 6, 8, 18, 19].

However, benign lesion classification remained a challenge for both modalities. CESM, despite its advantages, had a lower recall than DM for benign cases (16% vs. 36%), which may lead to over-interpretation of moderate enhancement patterns that resemble malignancy [20, 21]. It is important to recognize this limitation when implementing AI tools in clinical decision-making, as accurately identifying benign lesions helps prevent overtreatment and unnecessary biopsies.

To further investigate these differences, SHAP-based interpretability analysis is conducted on representative test set cases from each diagnostic category. The primary goal of this analysis is to assess whether SHAP attributions from DM align with those from CESM in terms of spatial localization and intensity, particularly in clinically relevant regions. CESM served as a contrast-enhanced reference for lesion conspicuity, while DM provided a baseline for evaluating the model's capacity to detect structural abnormalities in standard grayscale imaging.

In malignant cases, CESM SHAP maps showed highly localized, high-confidence activations centred on the lesion, reflecting biologically meaningful features such as neovascularity. DM, although more diffuse, produced attribution maps with similar spatial focus, suggesting that even without contrast, the model could detect morphologic abnormalities associated with malignancy. This convergence of SHAP patterns between modalities reinforces the notion that, even in the absence of CESM technology, DM-based deep learning models may still offer valuable insights for malignant lesion detection, especially in resource-limited settings.

In both modalities, normal cases showed diffuse, low-magnitude SHAP values, indicating appropriate model behavior when pathology is absent. Both modalities generated strong class activations, consistent with the high specificity observed in test metrics. CESM maps appeared marginally more ordered, likely due to improved signal clarity.

Benign lesion cases revealed the greatest ambiguity. SHAP attributions are scattered across multiple prediction columns, indicating the model's uncertainty in distinguishing benign features from malignant or normal ones. CESM sometimes accentuated this confusion by enhancing benign features that resemble malignant patterns,

which may explain its poorer performance in this category. Although DM produced less intense and more conservative activations, there is still insufficient class-specific separation. This finding highlights a critical gap in clinical applicability, the urgent need to increase model robustness in the classification of benign lesions.

These findings yield several important implications. First, the consistent superiority of CESM, particularly for malignant lesions, supports its integration into deep learning-assisted workflows for diagnostic support. CNNs' use of contrast information enables more precise and anatomically relevant feature extraction, consistent with radiological criteria and clinical expectations [10, 11].

Second, the use of SHAP as an interpretability tool not only enhanced transparency but also revealed model vulnerabilities, especially in the benign class. These findings point to possible directions for model improvement, like adding regions of interest annotated by radiologists, investigating domain-specific loss functions that penalize incorrectly classifying benign lesions, or using class-specific regularization to enhance decision boundaries [14, 18, 22].

Importantly, the alignment between SHAP attention and lesion location, especially in CESM, reinforces clinical trust in AI outputs. Models that consistently attend to pathologically relevant regions while suppressing irrelevant areas provide a critical foundation for real-world adoption of explainable AI in breast imaging [15, 23, 24].

Limitations of this study must be acknowledged. The limited dataset size reduces generalizability and increases the risk of overfitting or model instability, particularly within benign categories. Only the ResNet18 architecture can be used, which restricts the evaluation of model variability. The model must be externally validated using larger, multi-institutional cohorts in order to show its robustness.

Future research should address these limitations by utilizing the following approaches:

- Evaluating different CNN architectures (e.g., EfficientNet, DenseNet, Vision Transformers) for generalizability and performance enhancements.
- Using pairwise classification strategies (e.g., malignant vs. normal, benign vs. malignant) to reduce intra-class ambiguity.
- Multimodal analysis can be facilitated and a more thorough decision-making con-text can be provided by integrating radiomic features or clinical metadata.
- Classification accuracy may be increased by using contrastive or ensemble learning techniques that are adapted to fine diagnostic boundaries.

In summary, this study shows that deep learning models trained on standard DM images still have the ability to discriminate, especially for malignant lesions, where SHAP attributions show similar attention patterns to CESM, even though CESM performs noticeably better overall than DM. This understanding supports the usefulness of DM-based models in screening workflows aimed at early cancer detection and is essential in situations where CESM is not available. However, benign classification is still a major problem, and future research should focus on methods to increase clinical decision safety and model reliability in this area.

5. Conclusion

This study shows that CESM greatly enhances the interpretability and performance of CNN-based breast lesion classification when compared to traditional DM. Superior quantitative metrics and SHAP attribution maps demonstrate that CESM offers more accurate and anatomically precise predictions, even though DM still has clinical value, especially for identifying malignant lesions in settings with limited resources. There is a need for focused model improvements because both imaging modalities show limitations in the classification of benign lesions. Future model development and clinical validation are informed by the incorporation of explainability tools

like SHAP, which also improves model transparency and makes it easier to identify algorithmic limitations. These results underline the need for ongoing development to guarantee safe and dependable application across all diagnostic categories and support the use of CESM-enhanced AI systems in breast cancer screening.

Data Availability Statement
Not applicable.

Funding
This work was supported without any funding.

Conflicts of Interest
The author declares no conflicts of interest.

Ethical Approval and Consent to Participate
Not applicable.

References

[1]. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA. Cancer J. Clin. 74, 229–263 (2024).
[2]. Smith, R.A., Andrews, K.S., Brooks, D., Fedewa, S.A., Manassaram-Baptiste, D., Saslow, D., Wender, R.C.: Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. CA. Cancer J. Clin. 69, 184–210 (2019).
[3]. Brown, A.L., Vijapura, C., Patel, M., De La Cruz, A., Wahab, R.: Breast cancer in dense breasts: detection challenges and supplemental screening opportunities. Radiographics. 43, e230024 (2023).
[4]. Bodewes, F., Van Asselt, A., Dorrius, M., Greuter, M., De Bock, G.: Mammographic breast density and the risk of breast cancer: a systematic review and meta-analysis. The Breast. 66, 62–68 (2022).
[5]. Jochelson, M.S., Lobbes, M.B.: Contrast-enhanced mammography: state of the art. Radiology. 299, 36–48 (2021).
[6]. Liu, J., Xiao, R., Yin, H., Hu, Y., Zhen, S., Zhou, S., Han, D.: Meta-analysis and systematic review of the diagnostic value of contrast-enhanced spectral mammography for the detection of breast cancer. BMJ Open. 14, e069788 (2024).
[7]. Fallenberg, E., Dromain, C., Diekmann, F., Engelken, F., Krohn, M., Singh, J., Ingold-Heppner, B., Winzer, K., Bick, U., Renz, DM: Contrast-enhanced spectral mammography versus MRI: initial results in the detection of breast cancer and assessment of tumour size. Eur. Radiol. 24, 256–264 (2014).
[8]. Gluskin, J., Rossi Saccarelli, C., Avendano, D., Marino, M.A., Bitencourt, A.G., Pilewskie, M., Sevilimedu, V., Sung, J.S., Pinker, K., Jochelson, M.S.: Contrast-enhanced mammography for screening women after breast conserving surgery. Cancers. 12, 3495 (2020).
[9]. Alabousi, M., Zha, N., Salameh, J.-P., Samoilov, L., Sharifabadi, A.D., Pozdnyakov, A., Sadeghirad, B., Freitas, V., McInnes, M.D., Alabousi, A.: Digital breast tomosynthesis for breast cancer detection: a diagnostic test accuracy systematic review and meta-analysis. Eur. Radiol. 30, 2058–2071 (2020).
[10]. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., others: International evaluation of an AI system for breast cancer screening. Nature. 577, 89–94 (2020).
[11]. Tsuneki, M.: Deep learning models in medical image analysis. J. Oral Biosci. 64, 312–320 (2022).

[12]. Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R.: A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 292, 60–66 (2019).

[13]. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging. 35, 1299–1312 (2016).

[14]. Sadeghi, Z., Alizadehsani, R., Cifci, M.A., Kausar, S., Rehman, R., Mahanta, P., Bora, P.K., Almasri, A., Alkhawaldeh, R.S., Hussain, S., others: A review of Explainable Artificial Intelligence in healthcare. Comput. Electr. Eng. 118, 109370 (2024).

[15]. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30, (2017).

[16]. Khaled, R., Helal, M., Alfarghaly, O., Mokhtar, O., Elkorany, A., El Kassas, H., Fahmy, A.: Categorized Digital Database for Low energy and Subtracted Contrast Enhanced Spectral Mammography images. Cancer Imaging Arch. (2021). https://doi.org/10.7937/29kw-ae92.

[17]. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. Int. J. Data Min. Knowl. Manag. Process. 5, 1 (2015).

[18]. Wang, Q., Dong, J., Lin, X., Wang, M., Zhou, Y., Lin, S.: Diagnostic Accuracy of Contrast-Enhanced Spectral Mammography versus Digital Mammography for Breast Cancer: A Systematic Review and Meta-analysis. J. Breast Cancer. 25, 280–292 (2022).

[19]. Xiao, H., Chen, Y., Zhang, L., Chen, W., Mao, W.: Diagnostic performance of contrast-enhanced spectral mammography for breast cancer: A systematic review and meta-analysis. Breast Cancer Res. Treat. 199, 31–45 (2023).

[20]. Luczynska, E., Rzepka, S., Mrozek, J., Jasinski, P., Malecka, A., Adamczewski, G.: The challenge of benign enhancing lesions in contrast-enhanced mammography: A systematic review. Insights Imaging. 12, 1–11 (2021).

[21]. Mann, R.M., Kuhl, C.K., Kaiser, W.A.: Contrast-enhanced mammography (CEM): a new breast imaging tool. Review and recommendations. Eur. J. Radiol. 106, 28–36 (2018).

[22]. Shen, W., Ma, R., Zhang, P., Yan, S., Xie, Z., Tan, M., Xiao, L., Li, J.: Deep learning for breast cancer detection and diagnosis: A review. Med. Image Anal. 73, 102192 (2021).

[23]. Holzinger, A., Langs, G., Kittenberger, R., Langs, S., Denk, H.: Towards an Explainable AI in Medical Imaging. Sci. Rep. 9, 1–13 (2019).

[24]. Ghassemi, M., Naumann, T., Doshi-Velez, F.: Foundations of explainable artificial intelligence for medical imaging. Nat. Biomed. Eng. 5, 1116–1129 (2021).