IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 861~877
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp861-877

Hybrid texture-deep feature fusion for mammogram classification: a patient-level, calibrated evaluation

Muhammad Subali[1], Lulu Mawaddah Wisudawati[2], Teresa[3]
[1]Department of Informatics Engineering, Faculty of Engineering, Cendekia Abditama University, Tangerang, Indonesia
[2]Department of Informatics, Faculty of Industrial Technology, Gunadarma University, Depok, Indonesia
[3]Department of Nursing, Faculty of Nursing, Cendekia Abditama University, Tangerang, Indonesia

Article Info

Article history:

Received Aug 18, 2025
Revised Dec 29, 2025
Accepted Jan 22, 2026

Keywords:

Calibration
Computer-aided diagnosis
DWT-GLCM
Feature fusion
Mammography
ResNet-50

ABSTRACT

We propose a lightweight computer-aided diagnosis (CAD) framework that fuses four sub-band discrete wavelet transform gray-level co-occurrence matrix (DWT–GLCM) texture features with fine-tuned ResNet-50 embeddings under a strict, patient-level, leak-free evaluation protocol. Experiments were conducted on two public datasets: mammographic image analysis society (MIAS) (normal vs. abnormal) and curated breast imaging subset of the digital database for screening mammography (CBIS-DDSM) (benign vs. malignant). Five-fold cross-validation (CV) was confined to the training portion, operating thresholds were fixed on the validation split to target high recall, and the held-out test set was evaluated once. Performance was assessed using accuracy, F1-score, receiver operating characteristic (ROC)-area under the curve (AUC) with bootstrap 95% confidence intervals (CI), precision-recall (PR)-AUC, and calibration metrics (Brier score, expected calibration error). The proposed fusion model achieved ROC-AUC on MIAS (0.992) and strong performance on CBIS-DDSM (0.896), with consistent PR characteristics. Calibration analysis indicated reliable probability estimates and clinically interpretable decisions at a 95% sensitivity operating point. Ablation experiments revealed substantial gains over texture-only baselines and parity with convolutional neural network (CNN)-only models, highlighting fusion as a simple yet well-calibrated alternative for screening-oriented workflows. This study underscores the necessity of patient-level evaluation, explicit operating-point selection, and calibration reporting to ensure clinically meaningful CAD performance in mammography.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Muhammad Subali
Department of Informatics Engineering, Faculty of Engineering, Cendekia Abditama University
Islamic Raya St., Kelapa Dua, Tangerang, Banten, Indonesia
Email: subali@uca.ac.id

1. INTRODUCTION

Breast cancer remains the most frequently diagnosed cancer among women and a leading cause of cancer-related mortality worldwide [1]. This burden underscores the urgency of effective screening and early detection strategies. Mammography remains the gold standard for population-based screening and clinical work-up, with reporting and management commonly standardized using the breast imaging reporting and data system (BI-RADS) lexicon [2]. However, mammograms are often low-contrast and affected by tissue superposition. This makes interpretation difficult and increases inter-reader variability. Therefore, computer-aided diagnosis (CAD) systems have been developed to support radiologists in lesion detection and risk stratification. This is particularly important under limited data or complex imaging scenarios.

Early CAD approaches primarily relied on handcrafted descriptors to capture textural and multiresolution characteristics. Gray-level co-occurrence matrix (GLCM) statistics remain a foundational tool for quantifying textural relationships [3], while wavelet decompositions provide spatial–frequency representations well suited for breast tissue analysis [4]. Building on these ideas, Berbar [5] introduced hybrid sub-image descriptors (wavelet-contourlet (CT)/CT2 and statistical-texture (ST)-GLCM) coupled with support vector machines (SVMs) and reported strong results on digital database for screening mammography (DDSM) and mammographic image analysis society (MIAS). Abdullah *et al.* [6] proposed a texture-analysis pipeline using a multi-class SVM and reported 98% accuracy for early lesion detection, although the evaluation protocol was not clearly specified. Wisudawati *et al.* [7] demonstrated robust classification of normal/abnormal and benign/malignant [8] tumors using 2D-discrete wavelet transform (DWT)–GLCM combined with artificial neural network (ANN) (backpropagation neural network (BPNN)). These studies underscore the value of engineered textural features, though generalization is often limited by dataset size and evaluation inconsistencies.

With the advent of deep learning, mammography CAD has undergone a paradigm shift. Convolutional neural networks (CNNs) learn hierarchical semantics directly from images, enabling stronger generalization compared to handcrafted descriptors. Muduli *et al.* [9] employed CNNs on multi-modal inputs (mammography+ultrasound) with accuracies above 90% across datasets. Mahmood *et al.* [10] leveraged transfer learning, augmentation, and preprocessing to achieve area under the curve (AUC)≈0.99 for benign-malignant discrimination. Petrini *et al.* [11] introduced a two-view EfficientNet architecture that aggregates bilateral craniocaudal (CC) and mediolateral oblique (MLO) views, achieving AUC 0.934 under 5-fold cross-validation (CV) and 0.848 on the official split on curated breast imaging subset (CBIS) of the DDSM. Comparative evaluations show the superiority of residual networks such as ResNet-50 over visual geometry group version 16 (VGG16) on MIAS [12], while Saber *et al.* [13] benchmarked multiple transfer-learning backbones, achieving near-ceiling image-level metrics (AUC≈0.995). Recent works also highlight architectural innovations such as CNN–transformer hybrids [14] and fast leaky residual network with class imbalance reduction (FastLeakyResNet-CIR) [15] and systematic reviews emphasize both their promise and limitations [16], [17].

Alongside CNN-only pipelines, hybrid frameworks have been explored to combine complementary strengths of handcrafted and deep features [18], [19]. For example, Sajid *et al.* [18] integrated handcrafted and deep features in a unified framework for breast cancer classification. Shaukat *et al.* [19] combined deep CNN features with handcrafted texture features (e.g., Gabor and wavelet) for breast cancer detection in mammogram and ultrasound images. More recently, Das *et al.* [20] applied ResNet-50 to breast cancer magnetic resonance imaging (MRI) images and reported 92.01% accuracy. These results support the robustness of residual architectures in clinical imaging. Deshpande *et al.* [21] employed transfer learning with ResNet-50 on mammograms and reported 93.4% accuracy. Transfer learning also reduced the training burden compared with training from scratch. To address class imbalance, Alshamrani and Alshomrani [22] introduced a dual ResNet-50+synthetic minority over-sampling technique (SMOTE) framework. The study reported 99% accuracy on balanced sets and 90% on imbalanced ones. This result highlights the importance of balancing strategies in CAD.

Beyond deep CNNs, feature-level fusion strategies that combine handcrafted and deep representations have emerged as a promising direction [23], [24]. Razali *et al.* [23] fused CNN embeddings with wavelet-scattering features and reported 98-99% accuracy on INbreast. Vijayalakshmi *et al.* [24] combined shearlet transforms with GLCM/gray-level run-length matrix (GLRLM) and a bidirectional long short-term memory (BiLSTM)-CNN model, achieving 97.14% on MIAS. At patient level, Wimmer *et al.* [25] demonstrated multi-task fusion pipelines that aggregated predictions across tasks and views, yielding AUC 0.962 for lesion presence and 0.791 for malignancy on DDSM/CBIS-DDSM. Sun *et al.* [26] extended this with an attention-guided dual-branch CNN for craniocaudal (CC) and mediolateral oblique (MLO) fusion. Collectively, these studies suggest that hybrid pipelines can outperform CNN-only models, especially when evaluated under diverse or imbalanced datasets, by leveraging the complementary nature of texture-based and deep-learned representations.

In parallel, recent AI trends relevant to mammography include supervised contrastive pre-training frameworks that boost screening performance [27]. Domain adaptation and domain generalization are also explored to mitigate distribution shifts and improve external validity [28], [29]. In addition, vision transformer (ViT) models and multi-view architectures have shown promising results in recent comparative works and surveys [30], [31]. These developments motivate evaluation frameworks that balance architectural advances with clinical practicality, emphasizing interpretability, calibration, and reproducibility.

However, despite these advances, three limitations remain unaddressed. First, many studies rely on regions of interest (ROI)/patch-level splits that risk patient leakage and may inflate performance, while patient-level, leak-free evaluation is rarely enforced. Second, probability calibration is often neglected even though calibrated outputs are essential for threshold selection and risk communication [32], [33]. Third, in imbalanced clinical datasets, performance is frequently over-reported using receiver operating characteristic (ROC) curves alone, while precision-recall (PR) analysis provides a more realistic view of positive-class retrieval [34]–[36]. These limitations hinder clinical adoption and motivate the need for robust, interpretable, and deployment-oriented CAD pipelines.

This study proposes a hybrid fusion pipeline that concatenates DWT–GLCM texture features with fine-tuned ResNet-50 embeddings, classified via a shallow ANN under a patient-level, leak-free protocol (stratified 60/20/20 split; 5-fold CV on training/validation; thresholds fixed on validation and applied once to the held-out test). Beyond ROC-AUC, we provide calibration analysis (reliability diagrams, Brier score, and expected calibration error (ECE)) and precision-recall (PR)-based evaluation to reflect class imbalance. Our contributions are threefold:

i) A simple yet effective feature-level fusion of handcrafted multiresolution textures and deep embeddings for mammogram classification.
ii) Patient-level, leak-free evaluation ensuring reproducibility, and clinical relevance (with external validation (EV) across datasets, where applicable).
iii) Comprehensive assessment including calibration, high-recall operation, and PR-metrics, addressing critical gaps in the mammography CAD literature.

We explicitly position the novelty of this work in its calibration-aware, patient-level evaluation and clinical interpretability, and we outline attention-based gating or feature selection as pragmatic extensions to the present fusion design. Taken together, this work situates hybrid feature fusion within a clinically grounded evaluation framework, aligning with the broader push toward safe, reliable, and patient-centered AI for breast cancer screening. Section 2 describes the materials and methods (datasets, preprocessing/ROI extraction, feature extraction, fusion/classification, and evaluation protocol). Section 3 reports the results and discussion, including overall performance, ablations, calibration/PR analysis, comparative studies, EV, and qualitative error analysis. Section 4 concludes with implications and future research directions.

2. METHOD

We adopt a fusion-based CAD framework in Figure 1, which integrates statistical texture features extracted via DWT–GLCM with deep embeddings obtained from a fine-tuned ResNet-50, to address two binary classification tasks: normal vs. abnormal on the MIAS dataset and benign vs. malignant on the CBIS-DDSM dataset [37], [38]. ROIs are prepared following our earlier pipelines—automatic cropping, intensity-guided localization, and mask refinement as in Wisudawati *et al.* (normal–abnormal) [7] and Wisudawati *et al.* (benign–malignant) [8] then converted to 8-bit grayscale and resized to 224×224 (converted gray→RGB for ResNet-50). For textures, each ROI is decomposed by a one-level db4 DWT into low-low (LL), low-high (LH), high-low (HL), and high-high (HH). On each sub-band, a GLCM with 256 gray levels are computed at pixel distance d=1 in four orientations (0°, 45°, 90°, 135°). From the normalized GLCM $p(i,j)$, we extract contrast, correlation, energy, and homogeneity [3] and average across orientations, yielding 16 features per ROI (4 sub-bands×4 statistics). For deep features, a ResNet-50 (ImageNet) is fine-tuned on train; 2048-D avg_pool embedding is extracted per ROI. The 16-D texture vector and 2048-D deep vector are concatenated (2064-D) and z-scored using train statistics only, then classified by a shallow ANN (input 2064→hidden 128 (ReLU)→SoftMax-2) with inverse-frequency class weights.

For a leak-free evaluation protocol, we employ a patient-level approach, stratified 60/20/20 split (train/val/test). Five-fold CV is confined to train for model selection and ablation summaries. The operating threshold τ is chosen on validation to target 95% sensitivity, then applied unchanged to test, which is evaluated once.

Accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC, PR-AUC with 95% bootstrap confidence intervals (CI), and assess calibration is reported via Brier score and ECE (10 bins). For ablations (texture-only/CNN-only/fusion), we use paired fold-wise AUC tests (paired t-test or Wilcoxon, as appropriate). In addition, the decision threshold τ* fixed on val (≈95% sensitivity) is applied unchanged to both test and EV. EV strictly prevents patient overlap with the source cohort, uses the same normalization and τ* without refitting, and reports AUC-EV with 95% bootstrap CIs (B=2,000). Beyond the main ablation, three lightweight variants are evaluated under the identical protocol A is deeper head; B is Monte Carlo (MC) feature-dropout with mean aggregation; C is α-gated fusion with minimum redundancy maximum relevance (mRMR) (α, K chosen on val) and a qualitative error analysis (gradient-weighted class activation mapping (Grad-CAM)) is provided to illustrate true positive/true negative/false positive/false negative failure modes. Normality is screened via a Lilliefors test to choose between paired t-test and Wilcoxon.

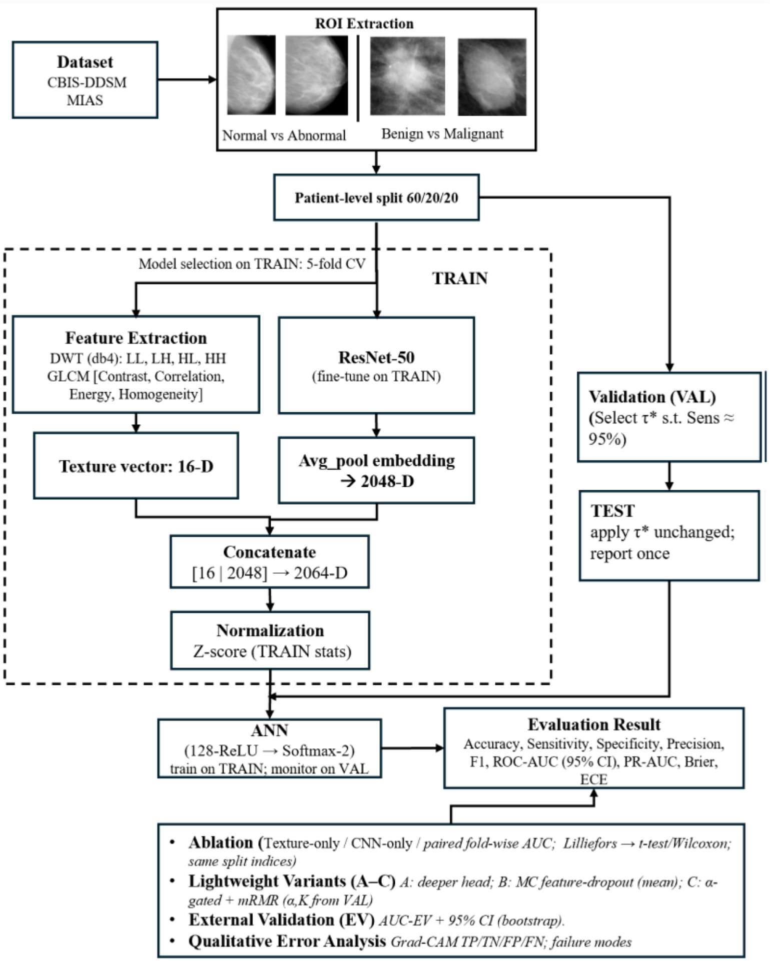

Dataset
CBIS-DDSM
MIAS
ROI Extraction
Normal vs Abnormal
Benign vs Malignant
Patient-level split 60/20/20
Model selection on TRAIN: 5-fold CV
TRAIN
Feature Extraction
DWT (db4): LL, LH, HL, HH
GLCM [Contrast, Correlation, Energy, Homogeneity]
ResNet-50
(fine-tune on TRAIN)
Texture vector: 16-D
Avg_pool embedding → 2048-D
Concatenate
[16 | 2048] → 2064-D
Normalization
Z-score (TRAIN stats)
Validation (VAL)
(Select τ* s.t. Sens ≈ 95%)
TEST
apply τ* unchanged; report once
ANN
(128-ReLU → Softmax-2)
train on TRAIN; monitor on VAL
Evaluation Result
Accuracy, Sensitivity, Specificity, Precision, F1, ROC-AUC (95% CI), PR-AUC, Brier, ECE
Ablation (Texture-only / CNN-only / paired fold-wise AUC; Lilliefors → t-test/Wilcoxon; same split indices)
Lightweight Variants (A–C) A: deeper head; B: MC feature-dropout (mean); C: α-gated + mRMR (α,K from VAL)
External Validation (EV) AUC-EV + 95% CI (bootstrap).
Qualitative Error Analysis Grad-CAM TP/TN/FP/FN; failure modes

Figure 1. Overview of the proposed fusion CAD framework integrating DWT-GLCM texture features and ResNet-50 deep embeddings

2.1. Dataset

This study employs two publicly available mammogram datasets: the MIAS [37] and the CBIS-DDSM [38]. MIAS, though lower in resolution, is widely used for normal vs. abnormal classification (107 normal and 89 abnormal). CBIS-DDSM provides high-resolution images with pathology-confirmed ROIs, supporting benign vs. malignant classification (481 benign and 527 malignant). All available images from MIAS and CBIS-DDSM were included. No samples were excluded except for corrupted or unreadable files, ensuring that the datasets were fully representative.

2.2. Preprocessing and ROI extraction

Preprocessing and ROI extraction followed validated procedures. Images were converted to grayscale, contrast-enhanced with clip-limited adaptive histogram equalization (CLAHE) and denoised using a 3×3 median filter. For normal vs. abnormal classification, the full breast region was segmented via adaptive thresholding, automatic cropping, and morphology [7], while for benign vs. malignant, suspicious masses were isolated [8]. ROIs were then used in two pipelines: DWT-GLCM on the ROI, and CNN-based models (ResNet-50) with resized (224×224) and normalized inputs. Figure 2 illustrates representative preprocessing and ROI extraction steps for the two classification tasks. In Figure 2(a), the breast area is segmented from the background and artifacts (e.g., labels and scanning noise) using adaptive thresholding, morphological operations, and automatic cropping, providing a clean ROI for normal vs. abnormal classification on MIAS. In Figure 2(b), a suspicious region is isolated from the mammogram using adaptive thresholding and morphological segmentation, producing an ROI that corresponds to the benign vs. malignant classification task on CBIS-DDSM. These steps ensure consistent ROI quality across datasets and provide standardized inputs for subsequent texture- and deep-feature pipelines.

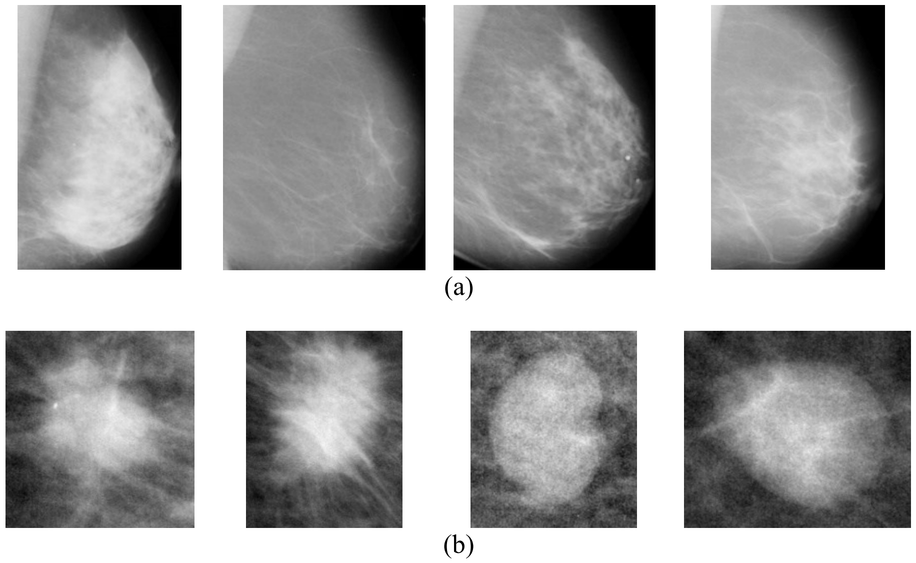


Figure 2. Examples of ROI extraction of (a) segmented breast area for normal vs. abnormal classification, and (b) isolated mass for benign vs. malignant classification

2.3. Feature extraction

2.3.1. DWT-GLCM texture features

Each 8-bit ROI was decomposed using a one-level 2D DWT with a Daubechies-4 (db4) basis, producing four sub-bands: LL, LH, HL, and HH. For each sub-band, the wavelet coefficients were linearly rescaled to the range [0, 255], and a GLCM with 256 gray levels was computed at a pixel distance of d=1 in four orientations (0°, 45°, 90°, and 135°). From the normalized GLCM $p(i,j)$, four standard Haralick statistics [3] were extracted: contrast (local intensity variation), correlation (linear dependency between pixel pairs), energy (textural uniformity), and homogeneity (closeness of the distribution to the diagonal). These statistics were averaged across the four orientations, resulting in a total of 16 texture features per ROI (4 sub-bands×4 statistics). Here, $p(i,j)$ denotes the normalized GLCM for a given sub-band, while $\mu_x,\ \mu_y$ are means, and $\sigma_x,\ \sigma_y$ are standard deviations of marginal distributions.

2.3.2. Fine-tuned ResNet-50 deep features

To capture high-level morphological patterns, a ResNet-50 pretrained on ImageNet is fine-tuned on the training split only, following the patient-level separation protocol. The final fully connected layer is replaced with a two-class classifier, and the network is trained for 5 epochs using Adam (learning rate 1×10^{-4}), batch size of 8, and cross-entropy loss. After fine-tuning, each ROI is forwarded through the network, and the resulting 2,048-D embedding is taken from the global average pooling (avg_pool) layer.

2.4. Feature fusion and classification

The final feature representation is obtained by concatenating the 16-D texture vector and the 2,048-D deep embedding, resulting in a 2,064-D vector. Z-score normalization is applied using the mean and standard deviation from the training set, and the same parameters are applied unchanged to validation and test sets to prevent leakage. Classification is performed using a shallow ANN with an input layer of size 2,064, one hidden layer with 128 ReLU units, and a softmax-2 output layer. Class imbalance is addressed by inverse-frequency weighting in the loss function. Training uses Adam optimization (learning rate 1×10^{-4}), batch size 32, and a maximum of 30 epochs.

This lightweight ANN head was deliberately chosen to ensure computational efficiency and practical deployability in CAD workflows, where rapid training and inference can be achieved on modest hardware, including CPU-only environments. While the current fusion mechanism is implemented through straightforward feature concatenation for simplicity and reproducibility, we intentionally adopt simple concatenation given our deployment-oriented scope and page constraints; lightweight attention-based fusion is deferred to future work. More advanced strategies such as attention-based gating, multi-branch fusion, or mRMR feature selection may further improve interpretability and discriminative power. These alternatives, together with deeper classifier designs (e.g., multi-layer perceptrons with dropout or transformer-based heads), are acknowledged as potential future extensions to strengthen robustness and AI novelty without compromising the clinical efficiency of the pipeline. As exploratory ablations, we also tested two minimal extensions: i) a scalar gate that re-weights the texture block before concatenation and ii) optional mRMR top-k feature selection on the fused vector.

2.5. Evaluation strategy and metrics

We report accuracy, sensitivity (recall), specificity, precision, F1-score, ROC-AUC, and PR-AUC as primary performance metrics. CV outcomes are summarized as mean±standard deviation across five folds on the training portion (train). On the held-out test set, 95% bootstrap CI for ROC-AUC are computed using B=2,000 patient-level resamples [36]. For comparative analysis, fusion is evaluated against texture-only and CNN-only baselines using paired statistical tests on fold-wise AUC. A Lilliefors normality test (α=0.05) [39] is used to determine whether a paired t-test or a Wilcoxon signed-rank test [40] is applied.

Calibration is assessed via Brier score and ECE (ECE; 10 bins), with reliability diagrams included where relevant [32], [33]. For clinical interpretability, a decision threshold is selected on the validation split to target 95% sensitivity and is applied unchanged to the test set, with specificity, precision, and F1-score reported at this fixed operating point. For EV, we used the validation-fixed threshold (τ) without any refitting and ensured no patient overlap between source and external cohorts; all normalization parameters were computed on train only and applied unchanged to val/test/EV.

As an exploratory step, we also applied post-hoc temperature scaling (TS) to calibrate the SoftMax outputs. The temperature parameter T was optimized on the validation split by minimizing the negative log-likelihood, and calibration was subsequently evaluated using the Brier score and (ECE; 10 bins). Since the primary results are reported in the uncalibrated setting, TS is presented only to illustrate feasibility and to highlight future directions in calibration-aware CAD design.

All experiments were implemented in MATLAB R2023a using the deep learning toolbox and image processing toolbox. Standard functions were employed for ROC/PR computation and bootstrap CI. Preprocessing, normalization, and threshold fitting were derived exclusively from the training/validation data and applied unchanged to the held-out test (and EV) set to ensure a fully leak-free evaluation and reproducibility of all steps.

3. RESULTS AND DISCUSSION

3.1. Overall performance of the proposed fusion model

This section reports the primary performance of the proposed fusion-based CAD framework on both datasets. We summarize CV discrimination (CV-AUC, mean±SD), and held-out test metrics accuracy, sensitivity, specificity, precision, F1-score, and ROC-AUC-each with 95% CI. The test set size (N) is shown in the dataset label for clarity. Unless otherwise noted, all results are pre-calibration. Table 1 presents the overall results, showing that the proposed model achieves high discrimination across both MIAS and CBIS-DDSM datasets, with minimal degradation between CV and held-out evaluation. CV-AUC is mean±SD over 5 folds. 95% CI for accuracy/sensitivity/specificity/precision: Wilson; F1-score: bootstrap; ROC-AUC: bootstrap (B=2000). Test set size is shown in the dataset label.

Table 1. Overall performance of the proposed fusion model on MIAS and CBIS-DDSM datasets

Dataset/Task	CV-AUC (mean±SD)	Test accuracy % (95% CI)	Test sensitivity % (95% CI)	Test specificity % (95% CI)	Test precision % (95% CI)	Test F1-score % (95% CI)	ROC-AUC (95% CI)
MIAS (N=39) Normal vs. Abnormal	0.997±0.007	97.44 (86.5–99.9)	100.00 (80.5–100)	95.45 (77.2–99.9)	94.44 (72.7–99.9)	97.14 (90.0–100.0)	0.992 [0.968–1.000]
CBIS-DDSM (N=201) Benign vs. Malignant	0.992±0.004	84.58 (78.8–89.3)	86.67 (78.6–92.5)	82.29 (73.2–89.3)	84.26 (76.0–90.6)	85.45 (79.8–90.2)	0.896 [0.845–0.938]

3.1.1. MIAS—normal vs. abnormal

The fusion model achieved CV-AUC=0.997±0.007 across five stratified folds, indicating high stability during training. On the held-out test set, AUC=0.992 (95% CI: 0.968-1.000), with accuracy=97.44% (86.5-99.9), sensitivity=100.00% (80.5-100), specificity=95.45% (77.2-99.9), precision=94.44% (72.7-99.9), and F1-score=97.14% (90.0-100.0). Figure 3 presents the performance of the proposed fusion model on MIAS. The ROC curve in Figure 3(a) demonstrates near-perfect class separation, with the curve closely approaching the top-left corner. The confusion matrix in Figure 3(b) confirms that all abnormal cases were correctly identified, with no false negatives—a critical requirement in screening scenarios. The learning curve in Figure 4 indicates rapid convergence and minimal overfitting throughout training.

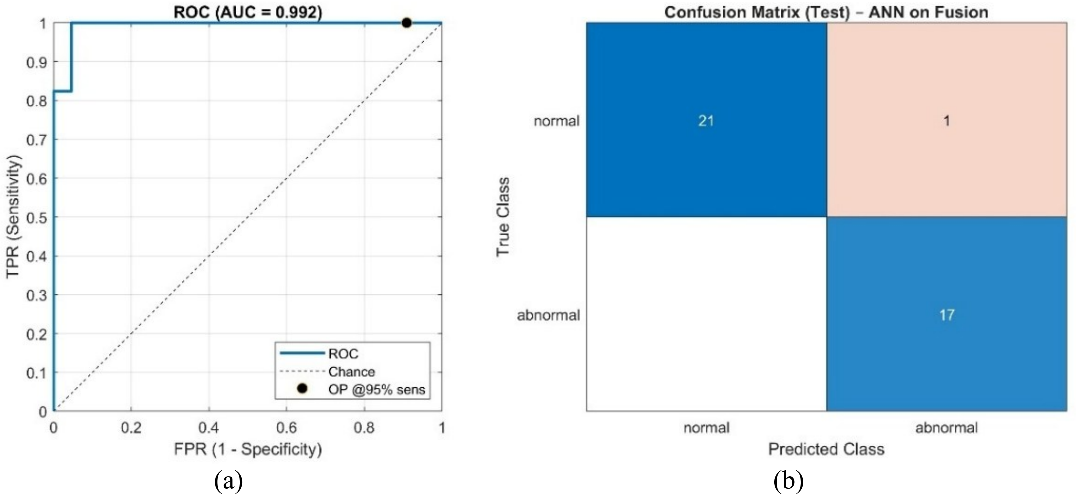


Figure 3. Proposed fusion model on MIAS of (a) ROC curve and (b) confusion matrix

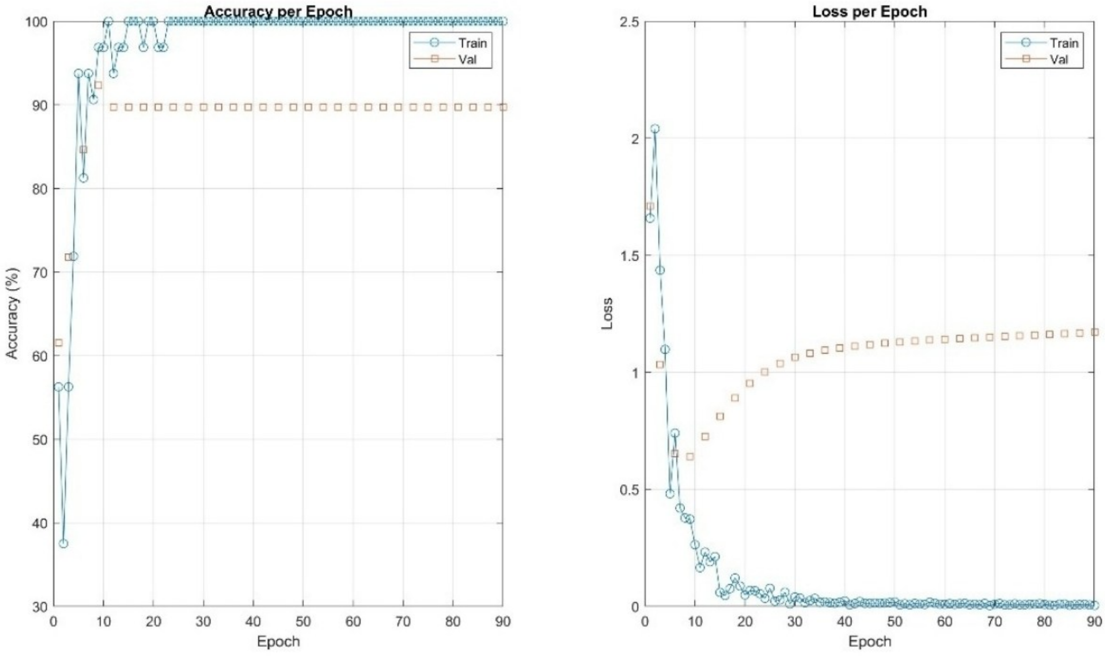
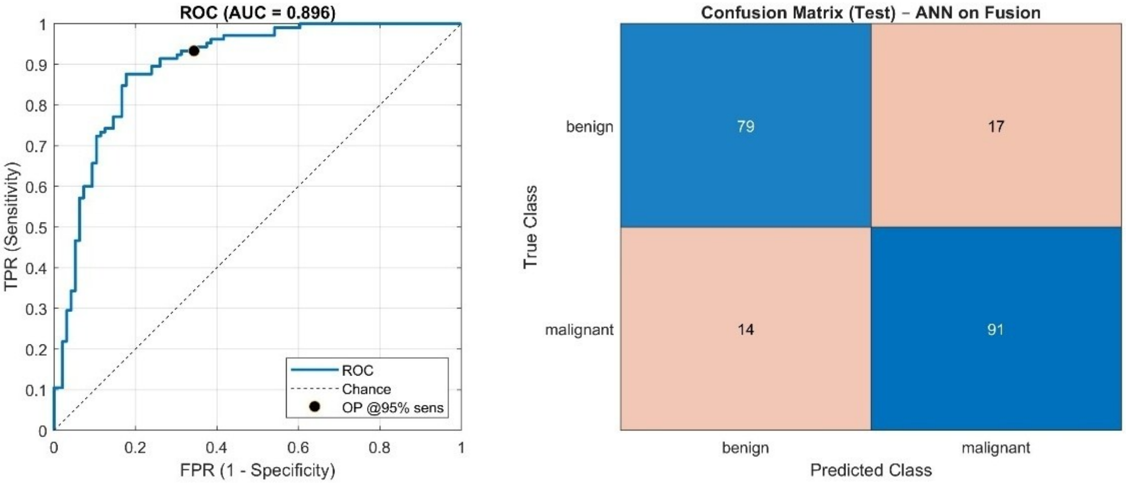


Figure 4. Learning curves for the ANN head on fusion features (MIAS)

3.1.2. CBIS-DDSM—benign vs. malignant

For CBIS-DDSM, CV-AUC=0.992±0.004 and test AUC=0.896 (95% CI: 0.845–0.938). Test performance was accuracy=84.58% (78.8-89.3), sensitivity=86.67% (78.6-92.5), specificity=82.29% (73.2-89.3), precision=84.26% (76.0–90.6), and F1-score=85.45% (79.8–90.2). Figure 5 presents the performance of the proposed fusion model on CBIS-DDSM. The ROC curve in Figure 5(a) demonstrates high discrimination but with a less steep rise compared to MIAS, reflecting the greater difficulty of benign–malignant separation. The confusion matrix in Figure 5(b) shows balanced detection performance but with some false positives and false negatives. Learning curves in Figure 6 indicate stable training with validation accuracy around 85-90%, suggesting the model generalizes reasonably well but could benefit from additional regularization. Overall, the proposed fusion approach demonstrates excellent generalization in normal vs. abnormal classification (MIAS) and competitive performance in the more challenging benign vs. malignant task (CBIS-DDSM). The integration of handcrafted DWT-GLCM features with deep semantic ResNet-50 features contributes to robust detection across different lesion types.

Notably, the MIAS results show the system's ability to achieve perfect sensitivity, a critical factor in early breast cancer detection. In CBIS-DDSM, while sensitivity remained high, a trade-off with specificity was observed, consistent with typical screening system behavior favoring recall. These findings support the clinical potential of hybrid fusion CAD systems, particularly in scenarios where both texture and deep features carry complementary diagnostic information.



(a) (b)

Figure 5. Proposed fusion model on CBIS-DDSM of (a) ROC curve and (b) confusion matrix

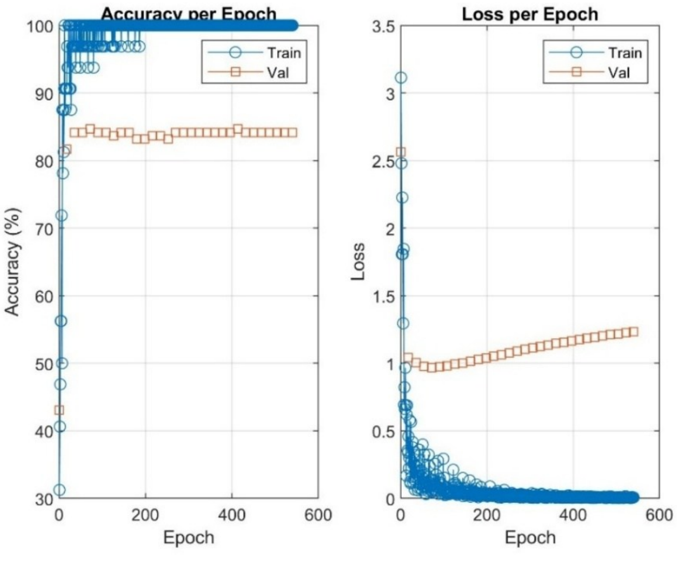


Figure 6. Learning curves for the ANN head on fusion features (CBIS-DDSM)

3.2. Ablation study—texture-only vs. fine-tuned ResNet-50 vs. fusion

We evaluated three architectural variants under an identical, leak-free protocol (patient-level 60/20/20 split; stratified 5-fold CV on the training set; identical ANN head and schedule): i) texture-only: 4-subband DWT–GLCM features→ANN, ii) CNN-only: fine-tuned ResNet-50 embeddings→ANN, and iii) fusion: concatenation of texture and CNN features→ANN. This setup isolates the contributions of handcrafted texture descriptors and learned deep representations. Table 2 presents ablation summary (CV-AUC and test AUC).

Table 2. Ablation summary (CV-AUC and test AUC) (Values are averaged over 5 folds)

Dataset	Method	CV-AUC (mean±SD)	Test AUC
MIAS	Texture-only (DWT–GLCM→ANN)	0.785±0.085	0.853
MIAS	CNN-only (ResNet-50 ft→ANN)	0.992±0.017	0.984
MIAS	Fusion (Texture+CNN→ANN)	0.997±0.004	0.987
CBIS-DDSM	Texture-only (DWT–GLCM→ANN)	0.835±0.033	0.808
CBIS-DDSM	CNN-only (ResNet-50 ft→ANN)	0.996±0.002	0.895
CBIS-DDSM	Fusion (Texture+CNN→ANN)	0.992±0.006	0.905

On the MIAS dataset (normal vs. abnormal), fusion achieved the highest cross-validated AUC (0.997±0.004), marginally outperforming CNN-only (0.992±0.017) and substantially exceeding texture-only (0.785±0.085). Pairwise comparisons showed significant gains of both CNN-only and fusion over texture-only (ΔAUC≈+0.21, p ≤0.043), while CNN-only versus fusion was not statistically significant (p=0.853). On the held-out test, AUCs were 0.853 (texture-only), 0.984 (CNN-only), and 0.987 (fusion). For the CBIS-DDSM dataset (benign vs. malignant), CNN-only and Fusion performed similarly across folds (0.996±0.002 vs. 0.992±0.006; ΔAUC=-0.004, p=0.205), with both significantly outperforming texture-only (ΔAUC≥+0.157, p ≤0.0005). Test AUCs were 0.808 (texture-only), 0.895 (CNN-only), and 0.905 (fusion).

Paired statistics are computed on fold-wise AUC (K=5). Normality was screened with Lilliefors; when violated we used Wilcoxon signed-rank, otherwise paired t-tests. Differences between CNN-only and fusion were not significant on either dataset, indicating that deep features dominate performance while texture cues provide complementary but not consistently significant gains under this protocol.

3.2.1. Combined results for lightweight variants (A-C)

We assessed three lightweight variants under an identical patient-level, leak-free protocol, with operating points fixed on the internal validation (val) split and then evaluated once on the held-out test set:

i) Variant A: deeper head: replacing the baseline ANN head (128–ReLU→softmax) with a deeper head (256–ReLU→Dropout (0.3)→128–ReLU→SoftMax).

ii) Variant B: uncertainty (MC feature-dropout): aggregating T=30 stochastic forward passes via the mean score (MC-mean). For MIAS, decision threshold τ is selected on the val split of the MC-mean model; for CBIS-DDSM, τ follows the val operating point of the deeper variant (as indicated in Table 3).

iii) Variant C: lightweight fusion (α-gated+mRMR): late fusion with a gating factor α and mRMR feature selection of size K, where (α,K) are chosen on val; the test split remained unseen until final evaluation.

Table 3 shows the combined results for experiments A-C on MIAS and CBIS-DDSM (patient-level, leak-free, and thresholds fixed on validation). All metrics are computed on the held-out test set; no refitting on test. For MIAS-B, τ is selected on the MC-mean VAL model; for CBIS-B, τ follows the VAL operating point of the deeper variant, as indicated in the "key" column.

Table 3. Combined results for experiments A–C on MIAS and CBIS-DDSM (patient-level, leak-free; thresholds fixed on validation)

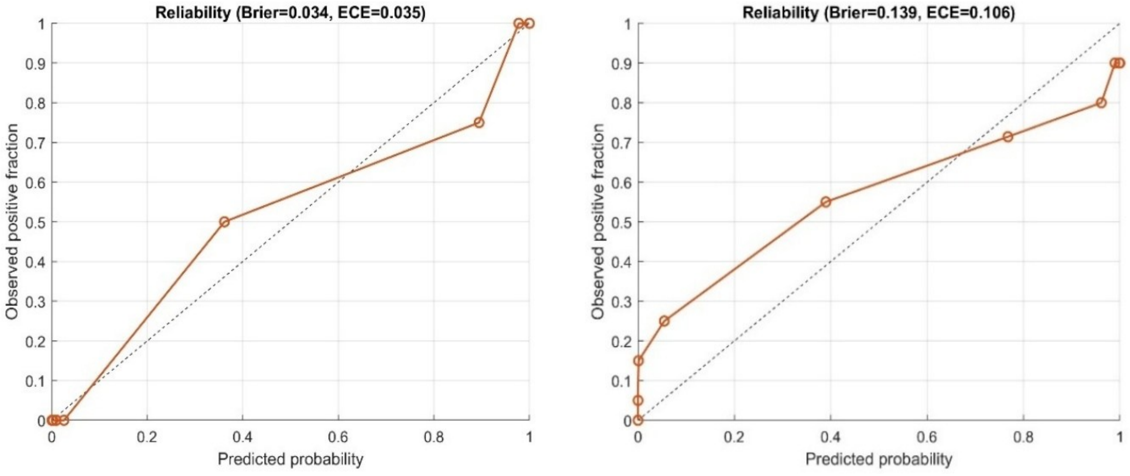
Section	Variant	Key	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	AUC
Normal vs. abnormal (MIAS)								
A: deeper-vs.-baseline	A: baseline head	128-ReLU→softmax, τ@VAL-95%Sens	90.77	89.51	92.05	92.36	90.91	0.929
A: deeper-vs.-baseline	A: deeper head	256→dropout(0.3)→128	90.38	90.24	90.52	90.91	90.58	0.927
B: uncertainty	B: MC-feature-dropout (mean)	T=30, rate=0.30, τ@VAL (MC-mean)	87.00	90.00	84.00	84.91	87.39	0.890
C: gated+mRMR	C: α-gated+mRMR (best VAL)	α=0.05, K=128, τ@VAL	90.00	89.02	90.99	90.91	89.96	0.926
Benign vs. malignant (CBIS-DDSM)								
A: deeper-vs.-baseline	A: baseline head	128-ReLU→softmax, τ@VAL-95%Sens	83.65	92.71	74.59	79.59	85.62	0.856
A: deeper-vs.-baseline	A: deeper head	256→dropout(0.3)→128	82.02	92.71	71.32	77.17	84.50	0.842
B: uncertainty	B: MC-feature-dropout (mean)	T=30, rate=0.30, τ from VAL (Deeper)	79.60	95.24	62.50	73.53	82.99	0.895
C: gated+mRMR	C: α-gated+mRMR (best VAL)	α=0.20, K=64, τ@VAL	81.43	88.57	63.54	72.66	79.83	0.842

Table 3 consolidates the results on MIAS (normal vs. abnormal) and CBIS-DDSM (benign vs. malignant). Variant A (deeper head) performs essentially on par with the baseline head (MIAS: accuracy 90.77% vs. 90.38%, AUC 0.929 vs. 0.927; CBIS-DDSM: 83.65% vs. 82.02%, 0.856 vs. 0.842), indicating that performance is largely driven by the fine-tuned ResNet-50 representations rather than head depth; differences between the deeper and baseline heads were not statistically significant, consistent with the small absolute gaps. Variant B (uncertainty, MC-feature-dropout, MC-mean aggregation) provides an uncertainty-aware operating point with clinically interpretable trade-offs (MIAS: accuracy 87.00%, sensitivity 90.00%, specificity 84.00%, precision 84.91%, F1-score 87.39%, AUC 0.890; CBIS-DDSM: 79.60%, 95.24%, 62.50%, 73.53%, 82.99%, 0.895), and the quartile-wise error trends support ordered risk stratification suitable for triage or second-reader use. Variant C (α-gated+mRMR) yields a pragmatic, compute-light enhancement while preserving the same decision protocol (best VAL settings: MIAS α=0.05, K=128→test accuracy 90.00%, sensitivity 89.02%, specificity 90.99%, precision 90.91%, F1-score 89.96%, AUC 0.926; CBIS-DDSM α=0.20, K=64→81.43%, 88.57%, 63.54%, 72.66%, 79.83%, 0.842). Overall, i) deeper heads offer negligible gains over a well-tuned baseline, ii) MC-mean uncertainty adds actionable interpretability with controllable recall–specificity trade-offs, and iii) α-gated+mRMR provides modest but consistent improvements without altering the leak-free evaluation.

3.3. Calibration and operating-point analysis

Calibration performance was assessed using reliability diagrams with 10 equal-frequency bins, along with Brier score and ECE. For clinical interpretability, the decision threshold was fixed on validation set to achieve 95% sensitivity and then applied unchanged to the held-out test set as shown on Figure 7. On the MIAS dataset (normal vs. abnormal), the fusion model showed good calibration (Brier score=0.034; ECE=3.5%). At the validation-selected threshold, the model achieved a high-recall operating point with sensitivity=100.0%, specificity=9.09%, precision=45.95%, and F1-score=62.96%. This trade-off emphasizes sensitivity, which is desirable for screening, while also reflecting the small test size (N=39).

On the CBIS-DDSM dataset (benign vs. malignant), calibration was moderate (Brier score=0.139; ECE=10.6%), with mild overconfidence at higher predicted probabilities. At the validation-selected threshold (τ=0.063), the test set performance was sensitivity=93.33%, specificity=65.62%, precision=74.81%, and F1-score=83.05%. The reliability diagram in Figure 7(a) shows well-calibrated estimates for MIAS, whereas Figure 7(b) confirms this overconfidence, particularly at higher probability bins.



(a) (b)

Figure 7. Reliability (calibration) curves with Brier score and ECE of (a) MIAS and (b) CBIS-DDSM

As an exploratory step, we also applied post-hoc TS for calibration. TS was fitted on the validation split by minimizing the negative log-likelihood, rescaling logits as $pT = softmax(z/T)$. Because TS is a monotonic transformation, discrimination metrics such as ROC-AUC/PR-AUC remain unchanged. On CBIS-DDSM, TS with $T = 3.183$ improved calibration (Brier: 0.1389→0.1285; ECE: 10.6%→5.1%) while preserving the same val-fixed operating-point metrics on test (sensitivity=93.33%, specificity=65.62%, precision=74.81%, and F1-score=83.05%). On MIAS, TS with $T = 2.863$ degraded calibration

(Brier: 0.0335→0.0503; ECE: 3.5%→9.8%) despite leaving ROC-AUC constant (0.992), suggesting over-smoothing when baseline calibration is already good. Since these adjustments did not improve overall discrimination, our primary results remain uncalibrated, with TS reported only to illustrate feasibility for future calibration-aware CAD design.

3.4. Precision-recall analysis

Given the inherent class imbalance in both datasets, we complement ROC-based evaluation with PR analysis, which is often more informative in such settings as shown in Table 4. For CBIS-DDSM, the proposed fusion model achieved a PR AUC of 0.868; in the high-recall regime—critical for screening applications—at recall≈0.95 (0.952 exactly), the model maintained a precision of 73.53%. For MIAS, the PR AUC was 0.930, and at recall=1.000 the model achieved a precision of 94.44%.

These results demonstrate that the proposed framework sustains high precision even under sensitivity-prioritized operation. This underscores its suitability for clinical screening where minimizing false negatives is critical. Compared with ROC-based evaluation, PR metrics provide a more realistic view of positive-class retrieval in imbalanced datasets, thereby reinforcing the robustness and clinical applicability of the proposed CAD system.

Table 4. PR metrics on the held-out test set for the fusion model

Dataset	PR-AUC	Precision @ recall≈0.95
MIAS (N=39)	0.930	94.44% (recall=1.000)
CBIS-DDSM (N=201)	0.868	73.53% (recall=0.952)

3.5. Comparison with previous studies

Direct comparison across studies is constrained by differences in dataset partitioning, evaluation protocols (patient-level vs. ROI/patch-level), and class definitions. Nevertheless, Table 5 summarizes representative results on MIAS, CBIS-DDSM, and related datasets. Handcrafted pipelines have demonstrated strong performance: Berbar [5] reported 97.89% accuracy using wavelet-CT/GLCM with SVM [5]; Abdullah *et al.* [6] achieved 98.00% with a texture-based multi-class SVM; Wisudawati *et al.* [7], [8] obtained 93.8% for three-class MIAS classification using 2D-DWT+GLCM+ANN, and subsequently 88.7% on DDSM for benign-malignant discrimination using BPNN.

Table 5. Comparative performance of mammographic breast cancer classification

Study (Year)	Dataset	Task	Method	AUC/accuracy (%)
Berbar [5]	MIAS	Normal vs. malignant	Hybrid Wavelet–CT1/CT2, ST-GLCM+SVM	97.89 (accuracy)
Abdullah *et al.* [6]	MIAS	Multi-class classification	Texture features+multi-class SVM	98.00 (accuracy)
Wisudawati *et al.* [7]	MIAS	Normal vs. benign vs. malignant	2D-DWT+GLCM+ANN	93.80 (accuracy)
Wisudawati *et al.* [8]	DDSM	Benign vs. malignant	2D-DWT+GLCM+BPNN	88.70 (accuracy)
Muduli *et al.* [9]	DDSM	Normal, benign, malignant	Deep CNN	90.68 (accuracy)
Mahmood *et al.* [10]	MIAS+Private	Benign vs. malignant	ConvNet (transfer learning)	0.990 (AUC)
Petrini *et al.* [11]	CBIS-DDSM	Diagnosis (two-view) malignant vs. non malignant	End-to-end EfficientNet (two-view)	0.9344 (5-fold CV), 0.8483 (official)
Vijayalakshmi *et al.* [24]	MIAS	Normal vs. abnormal	Shearlet+GLCM/GLRLM+BiLSTM-CNN	97.14 (accuracy)
Ashwini *et al.* [12]	MIAS	Normal vs. benign vs. malignant	VGG16 vs. ResNet-50 (TL)	Acc 91.23% (VGG16); 99.01% (ResNet-50)
Saber *et al.* [13]	MIAS	Normal vs. benign vs. malignant	Novel deep learning (ResNet-50)	0.970 (AUC)
Razali *et al.* [23]	INbreast	Mass classification	CNN+wavelet scattering+texture fusion	98.00 (accuracy)
Wimmer *et al.* [25]	CBIS-DDSM	Lesion presence/malignancy	Multi-task fusion deep pipeline	0.962 (AUC, presence); 0.791(AUC, malignancy)
This work	MIAS	Normal vs. abnormal	DWT–GLCM+fine-tuned ResNet-50 (fusion)	0.992 (AUC)
This work	CBIS-DDSM	Benign vs. malignant	DWT–GLCM+fine-tuned ResNet-50 (fusion)	0.896 (AUC)

Note: prior results may use non-comparable protocols (e.g., ROI/patch-level or non-patient-level splits); values are provided for context and are not direct head-to-head claims.

With introduction of deep learning, transfer learning substantially improved baselines. Muduli *et al.* [9] achieved 90.68% on DDSM (three-class) using CNNs; Mahmood *et al.* [10] reported AUC 0.990 on MIAS and a private dataset with a transfer-learning ConvNet and targeted preprocessing/augmentation; and Petrini *et al.* [11] employed an end-to-end two-view EfficientNet on CBIS-DDSM, yielding AUC 0.9344 (5-fold CV) and 0.8483 (official split). Further gains were demonstrated by Vijayalakshmi *et al.* [24] (97.14% with Shearlet+GLCM/GLRLM+BiLSTM-CNN), Ashwini *et al.* [12] (99.01% with ResNet-50 vs. 91.23% with VGG16), and Saber *et al.* [13] (AUC 0.970 with a ResNet-50-based model).

Hybrid and patient-level fusion approaches continue to show promise. Razali *et al.* [23] achieved 98% on INbreast by combining CNN embeddings with wavelet-scattering features, demonstrating that multi-resolution handcrafted descriptors and deep spatial features can capture both fine texture variations and global structural patterns. Wimmer *et al.* [25] reported AUC 0.962 for lesion presence and 0.791 for malignancy using a multi-task fusion pipeline on DDSM/CBIS-DDSM, highlighting the value of integrating multi-view, multi-task predictions to align model outputs with radiological workflows.

Under a patient-level, leak-free protocol, our proposed DWT–GLCM+fine-tuned ResNet-50 fusion model achieves AUC 0.992 on MIAS (normal vs. abnormal) and AUC 0.896 on CBIS-DDSM (benign vs. malignant). This performance is competitive with, and in some cases exceeds, the performance of recent deep and hybrid approaches. These results underscore that integrating multi-scale texture cues from wavelet-GLCM with deep semantic embeddings from a fine-tuned CNN provides complementary information, resulting in more robust and clinically relevant mammogram classification.

3.6. External validation

To assess cross-dataset generalizability, we trained the fusion model on one dataset and evaluated it on an independent cohort. Specifically, the MIAS trained model (normal vs. abnormal) was tested on 50 DDSM cases (25 normal and 25 abnormal), and the CBIS-DDSM trained model (benign vs. malignant) was tested on 50 MIAS cases (25 benign and 25 malignant). Table 6 shows the EV performance (cross-dataset) of the proposed fusion model. Values are percentages with 95% Wilson CI for accuracy, sensitivity, specificity, and precision; F1-score uses bootstrap CI. AUC EV+95% CI is reported. Performance is summarized in Table 6.

Table 6. EV performance (cross-dataset) of the proposed fusion model

Task (Training→external test)	N (pos/neg)	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC EV+95% CI (bootstrap)
Normal vs. abnormal (MIAS→DDSM)	50 (25/25)	86.0% (73.8–93.0)	100.0% (86.7–100.0)	72.0% (52.4–85.7)	78.1% (61.2–89.0)	87.7% (80.6–94.3)	0.890 [0.796-0.984]
Benign vs. malignant (CBIS-DDSM→MIAS)	50 (25/25)	90.0% (78.6–96.7)	100.0% (86.7–100.0)	80.0% (60.9–91.1)	83.3% (66.4–92.7)	90.9% (84.7–98.0)	0.934 [0.861-1.000]

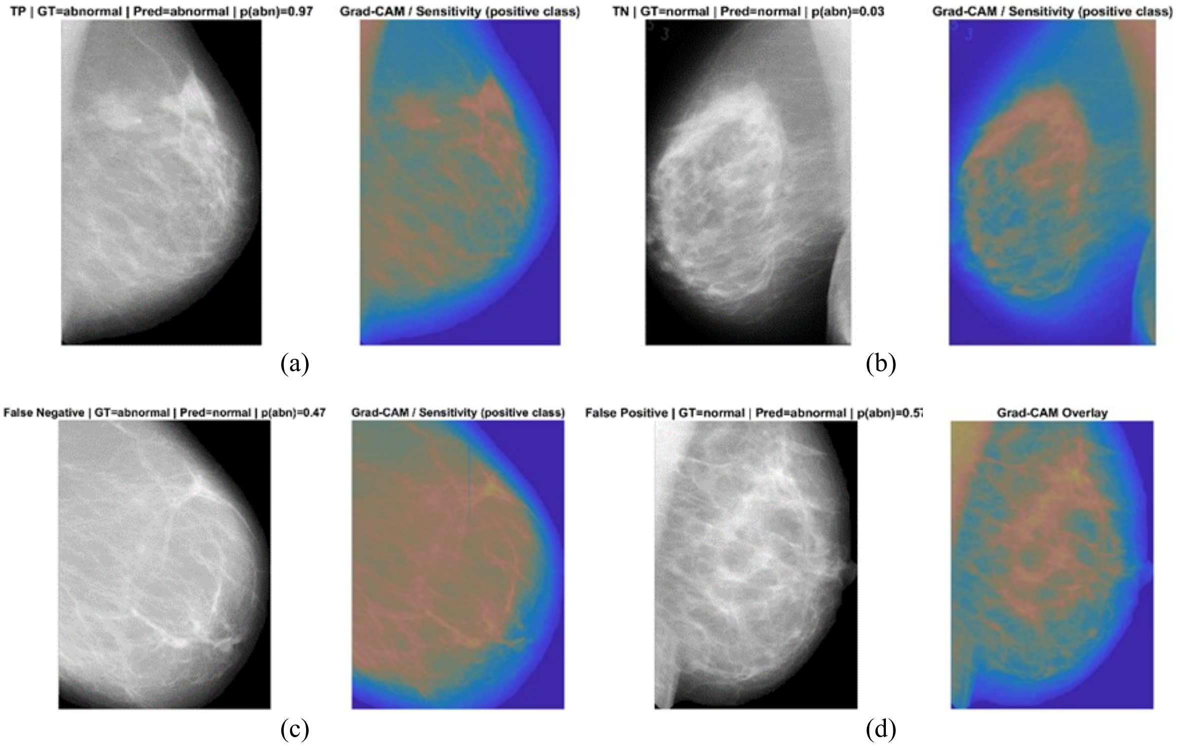
For the normal vs. abnormal task (MIAS→DDSM), external evaluation achieved 86.0% accuracy (95% CI: 73.8-93.0) with 100.0% sensitivity (86.7-100.0) and 72.0% specificity (52.4-85.7). Precision was 78.1% (61.2-89.0) and F1-score 87.7% (80.6-94.3). The lower specificity reflects more false positives, consistent with domain shift between MIAS and digitized film based DDSM images. Consistently, the threshold free ROC performance is moderate, AUC 0.890 [0.796-0.984] under the validation-fixed, high-recall operating point, further reflecting inter-dataset domain shift. For the benign vs. malignant task (CBIS-DDSM→MIAS), performance reached 90.0% accuracy (78.6-96.7) with 100.0% sensitivity (86.7-100.0) and 80.0% specificity (60.9-91.1). Precision was 83.3% (66.4-92.7) and F1-score 90.9% (84.7-98.0). Sensitivity remained high, indicating stable detection of malignant cases under cross-dataset testing, while a modest specificity reduction suggests increased false positives.

Overall, these EV results indicate that the proposed fusion framework maintains clinically meaningful sensitivity under cross-dataset evaluation. Some degradation is expected due to dataset heterogeneity. Nevertheless, the model preserves reliable detection, underscoring the value of domain adaptation and larger multi-center validation for deployment readiness.

3.7. Qualitative error analysis

Beyond quantitative metrics, we also provide qualitative visualizations of representative cases covering both correct and incorrect predictions (Figure 8). Figure 8(a) shows a true positive case (GT=abnormal, Pred=abnormal; p(abn)≈0.97), where the Grad-CAM heatmap exhibits concentrated attention over lesion-bearing regions. Figure 8(b) presents a true negative case (GT=normal, Pred=normal;

p(abn)≈0.03), characterized by a low abnormality score; however, the abnormal class heatmap may still highlight benign parenchymal structures with minimal influence on the final decision. In contrast, Figure 8(c) illustrates a false negative case (GT=abnormal, Pred=normal; p(abn)≈0.47), typically associated with subtle or low-contrast findings that evade detection. Finally, Figure 8(d) depicts a false positive case (GT=normal, Pred=abnormal; p(abn)≈0.57), which often occurs in dense parenchyma or vessel overlaps misinterpreted as abnormal tissue. These qualitative observations illustrate common failure modes in mammography CAD and underscore the importance of interpretability in supporting clinical adoption.


TP | GT=abnormal | Pred=abnormal | p(abn)=0.97
Grad-CAM / Sensitivity (positive class)
TN | GT=normal | Pred=normal | p(abn)=0.03
Grad-CAM / Sensitivity (positive class)
False Negative | GT=abnormal | Pred=normal | p(abn)=0.47
Grad-CAM / Sensitivity (positive class)
False Positive | GT=normal | Pred=abnormal | p(abn)=0.5
Grad-CAM Overlay

(a) (b)

(c) (d)

Figure 8. Heatmap visualizations (Grad-CAM, abnormal class) of (a) true positive, (b) true negative, (c) false negative, and (d) false positive

3.8. Discussion

Here we present a patient-level, leak-free fusion pipeline that integrates wavelet-GLCM texture features with fine-tuned ResNet-50 embeddings through a lightweight ANN. The approach delivers strong discrimination on MIAS and competitive performance on CBIS-DDSM, with the operating point fixed on the validation split for high recall and then applied unchanged to the test set. In addition to standard accuracy and AUC, we report calibration and PR analyses to align evaluation with screening needs and clinical decision-making.

Relative to prior work, often based on ROI/patch-level splits that risk leakage—our protocol enforces patient-level separation and therefore offers a stricter, more clinically relevant assessment while remaining competitive with recent deep and hybrid pipelines. Ablation findings clarify what drives performance: CNN embeddings are the principal source of discrimination, whereas handcrafted texture features add complementary signal that stabilizes results but do not consistently surpass CNN-only models. This suggests that future gains are more likely to come from richer context modeling (e.g., multi-view fusion, attention-based integration, and lightweight transformer heads) than from further variants of classic descriptors alone. Beyond the tested α-gated and mRMR variants, we defer full attention-based cross-modal fusion (e.g., cross-attention between texture and deep features) to future work in order to preserve deployability and the computational footprint of the system.

Under the same leak-free protocol, we evaluated three lightweight variants: a deeper classification head, MC-mean uncertainty scoring, and α-gated fusion with optional mRMR on the fused vector. The deeper head yielded negligible gains, indicating that performance is dominated by representation quality rather than head depth; α-gated+mRMR delivered small but consistent test-set improvements under the same operating-point protocol; and MC-mean uncertainty improved operational flexibility (e.g., high-recall triage) rather than raw AUC. All operating points including the MC-mean threshold were fixed on the internal validation split to preserve a leak-free evaluation. All experiments were carried out on a consumer-grade laptop (Intel Core Ultra 7-155H, 32 GB RAM, integrated Intel Arc GPU) operated in CPU-only mode. Under the patient-level, leak-free protocol (stratified 60/20/20 split with 5-fold CV confined to the training portion), the proposed fusion pipeline required ≈22 minutes (1,314 s) to complete training and fine-tuning. Per-ROI inference remained under 1 s, indicating that the approach is executable on modest, non-GPU clinical hardware and can be deployed on standard workstations. Running the same pipeline on a GPU workstation would be expected to further reduce training time. We also plan to benchmark lightweight ViT/ConvNeXt heads and self-supervised contrastive pretraining under the same patient-level, leak-free protocol, aiming to improve out-of-distribution robustness without sacrificing calibration and efficiency.

Calibration analyses show that the model is well-calibrated on MIAS and mildly overconfident on CBIS-DDSM at higher scores. Post-hoc TS, fitted on the validation set by minimizing negative log-likelihood, behaved as expected for monotonic rescaling of logits: discrimination (ROC-AUC/PR-AUC) did not change. On CBIS-DDSM, TS improved calibration (ECE 10.6%→5.1%; Brier 0.1389→0.1285), whereas on MIAS already well-calibrated, it slightly degraded calibration (ECE 3.5%→9.8%; Brier 0.0335→0.0503). Because these adjustments did not improve discrimination and the effects were dataset-dependent, we report primary results without TS and treat it as a feasibility check for calibration-aware deployment.

Given class imbalance, PR analysis provides a more faithful view of positive-class retrieval and confirms that the fusion model sustains high precision in the high-recall regime appropriate for screening. Cross-dataset tests preserve clinically meaningful sensitivity but reduce specificity, consistent with domain shift and increased false positives, underscoring the need for domain adaptation and multi-center validation to improve robustness across acquisition settings. Qualitatively, false negatives tend to involve subtle or low-contrast findings, while false positives often arise in dense parenchyma or vessel overlaps patterns that motivate refinements in preprocessing, thresholding, and explainability to support radiologist trust.

The work has limitations: a single split per dataset, reliance on pre-segmented ROIs rather than end-to-end detection→classification, lack of explicit domain-shift modeling, and potential run-to-run variability due to stochastic training. The fusion mechanism is a straightforward concatenation, and the ANN head is shallow design choices made for computational efficiency and deployability. Future directions include attention-based gating, systematic feature selection (e.g., mRMR), lightweight deeper heads (multilayer perceptron (MLP)/transformer), and self-supervised pretraining (simple framework for contrastive learning of visual representations (SimCLR)/bootstrap your own latent (BYOL)) to improve generalization without additional labels, alongside broader uncertainty quantification, expanded explainability, comparative calibration methods, and multi-center/domain-adaptation studies aimed at reducing false positives while maintaining high sensitivity. Overall, the main contribution is a calibrated, reproducible, and clinically oriented evaluation of a simple yet robust fusion pipeline under a patient-level, leak-free protocol that aligns with the requirements of sensitivity-focused breast-cancer screening.

4. CONCLUSION

We presented a data-efficient fusion framework that integrates wavelet-GLCM texture features with fine-tuned ResNet-50 embeddings, evaluated under a strict, leak-free, patient-level protocol. Across MIAS and CBIS-DDSM, the approach achieved strong discrimination with consistent behavior between CV and held-out testing (e.g., MIAS AUC=0.992; CBIS-DDSM AUC=0.896). Calibration analyses showed well-behaved probability estimates and supported clinically interpretable, high-recall operation (thresholds fixed on validation and applied once to test). External, cross-dataset validation (MIAS→DDSM; CBIS-DDSM→MIAS) further demonstrated robustness (accuracy 86.0% and 90.0%, respectively) while preserving 100% sensitivity in both directions. Ablation experiments indicated that deep CNN embeddings are the dominant source of discriminative power, with handcrafted textures providing complementary cues—especially in challenging cases. We also explored lightweight variants (scalar gating and mRMR feature selection); although they improved validation metrics in some settings, they did not consistently enhance test performance, supporting the choice of simple concatenation with a shallow ANN as the most stable and deployment-ready design. Overall, the principal contribution of this work lies in establishing a reproducible, calibration-aware, and clinically interpretable evaluation framework under realistic patient-level constraints, rather than proposing an entirely new network architecture. Limitations include the reliance on pre-segmented ROIs, the use of a shallow classifier head, and residual sensitivity to domain shift. Future work will therefore pursue multi-center validation, domain adaptation, uncertainty quantification, and more advanced fusion strategies (e.g., attention-guided, multi-branch, and transformer-based) to further improve robustness and generalizability. For completeness, we note that the primary discrimination results are reported without post-hoc TS; TS improved calibration on CBIS-DDSM but not on MIAS, highlighting dataset-dependent effects.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the curators and maintainers of the CBIS-DDSM and MIAS mammographic datasets for making these resources publicly available to the research community.

FUNDING INFORMATION

This work was supported by the Directorate of Research and Community Service or Direktorat Penelitian dan Pengabdian kepada Masyarakat (DPPM), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, under contract No. 8018/LL4/PG/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, minimize authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Subali	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Lulu Mawaddah Wisudawati	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Teresa				✓		✓	✓	✓		✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

The mammographic datasets utilized in this study are publicly accessible. The curated breast imaging subset of the digital database for screening mammography (CBIS-DDSM) can be obtained from The Cancer Imaging Archive at https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM. The mammographic image analysis society (MIAS) database is available at http://peipa.essex.ac.uk/info/mias.html.

REFERENCES

[1] F. Bray *et al.*, "Global cancer statistics 2022: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, May 2024, doi: 10.3322/caac.21834.

[2] C. D'Orsi, E. Sickles, E. Mendelson, and E. Morris, *ACR BI-RADS atlas: breast imaging reporting and data system*, Reston, United States: American College of Radiology, 2013.

[3] R. M. Haralick, I. Dinstein, and K. Shanmugam, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[4] S. Mallat, *A wavelet tour of signal processing*, Florida, United States: Academic Press, 2008.

[5] M. A. Berbar, "Hybrid methods for feature extraction for breast masses classification," *Egyptian Informatics Journal*, vol. 19, no. 1, pp. 63–73, Mar. 2018, doi: 10.1016/j.eij.2017.08.001.

[6] A. K. Abdullah, R. M. Azawi, I. T. Ibrahim, and A. A. Ajwad, "Mammography images classification system based texture analysis and multi class support vector machine," in *AIP Conference Proceedings*, 2023, doi: 10.1063/5.0110733.

[7] L. M. Wisudawati, S. Madenda, E. P. Wibowo, and A. A. Abdullah, "Feature extraction optimization with combination 2D-discrete wavelet transform and gray level co-occurrence matrix for classifying normal and abnormal breast tumors," *Modern Applied Science*, vol. 14, no. 5, Apr. 2020, doi: 10.5539/mas.v14n5p51.

[8] L. M. Wisudawati, S. Madenda, E. P. Wibowo, and A. A. Abdullah, “Benign and malignant breast tumors classification based on texture analysis and backpropagation neural network,” *Computer Optics*, vol. 45, no. 2, pp. 227–234, Apr. 2021, doi: 10.18287/2412-6179-CO-769.

[9] D. Muduli, R. Dash, and B. Majhi, “Automated diagnosis of breast cancer using multi-modal datasets: a deep convolution neural network based approach,” *Biomedical Signal Processing and Control*, vol. 71, Jan. 2022, doi: 10.1016/j.bspc.2021.102825.

[10] T. Mahmood, J. Li, Y. Pei, F. Akhtar, M. Ur Rehman, and S. H. Wasti, “Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach,” *PLoS ONE*, vol. 17, no. 1, Jan. 2022, doi: 10.1371/journal.pone.0263126.

[11] D. G. P. Petrini, C. Shimizu, R. A. Roela, G. V. Valente, M. A. A. K. Folgueira, and H. Y. Kim, “Breast cancer diagnosis in two-view mammography using end-to-end trained efficientnet-based convolutional network,” *IEEE Access*, vol. 10, pp. 77723–77731, 2022, doi: 10.1109/ACCESS.2022.3193250.

[12] P. Ashwini, N. Suguna, and N. Vadivelan, “Detection and classification of breast cancer types using VGG16 and ResNet50 deep learning techniques,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 5, pp. 5481–5488, Oct. 2024, doi: 10.11591/ijece.v14i5.pp5481-5488.

[13] A. Saber, M. Sakr, O. M. A.-Seida, A. Keshk, and H. Chen, “A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique,” *IEEE Access*, vol. 9, pp. 71194–71209, 2021, doi: 10.1109/ACCESS.2021.3079204.

[14] A. Zeynali, M. A. Tinati, and B. M. Tazehkand, “Hybrid CNN-transformer architecture with xception-based feature enhancement for accurate breast cancer classification,” *IEEE Access,* vol. 12, pp. 189477-189493, 2024, doi: 10.1109/ACCESS.2024.3516535.

[15] R. Zeng *et al*., “FastLeakyResNet-CIR: a novel deep learning framework for breast cancer detection and classification,” *IEEE Access,* vol. 12, pp. 70825-70832, 2024, doi: 10.1109/ACCESS.2024.3401729.

[16] A. Kumar, R. Saini, and R. Kumar, “A systematic review of breast cancer detection using machine learning and deep learning,” *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON),* 2023, pp. 1128-1133, doi: 10.1109/UPCON59197.2023.10434530.

[17] X. Wen, X. Guo, S. Wang, Z. Lu, and Y. Zhang “Breast cancer diagnosis: a systematic review,” *Biocybernetics and Biomedical Engineering*, vol. 44, no. 1, pp. 119-148, 2024, doi: 10.1016/j.bbe.2024.01.002.

[18] U. Sajid, R. A. Khan, S. M. Shah, and S. Arif, “Breast cancer classification using deep learned features boosted with handcrafted features,” *Biomedical Signal Processing and Control*, vol. 84, 2023, doi: 10.1016/j.bspc.2023.105353.

[19] A. Shaukat *et al.*, “HyFusion-X: hybrid deep and traditional feature fusion with ensemble classifiers for breast cancer detection using mammogram and ultrasound images,” *Scientific Reports,* vol. 15, Nov. 2025, doi: 10.1038/s41598-025-22262-1.

[20] T. Das, D. S. K. Nayak, A. Kar, L. Jena, and T. Swarnkar, “ResNet-50: the deep networks for automated breast cancer classification using mr images,” *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, 2024, doi: 10.1109/ASSIC60049.2024.10507980.

[21] S. Deshpande, R. K. Patel, S. S. Chouhan, and H. Vishwakarma, “Transfer learning with ResNet50 for enhanced mammographic breast cancer identification,” in *2024 5th International Conference on Circuits, Control, Communication and Computing (I4C)*, Oct. 2024, pp. 58–63. doi: 10.1109/I4C62240.2024.10748454.

[22] A. F. A. Alshamrani and F. S. Z. Alshomrani, “Optimizing breast cancer mammogram classification through a dual approach: a deep learning framework combining ResNet50, smote, and fully connected layers for balanced and imbalanced data,” *IEEE Access*, vol. 13, pp. 4815–4826, 2025, doi: 10.1109/ACCESS.2024.3524633.

[23] N. F. Razali, I. S. Isa, S. N. Sulaiman, N. K. A. Karim, and M. K. Osman, “CNN-wavelet scattering textural feature fusion for classifying breast tissue in mammograms,” *Biomedical Signal Processing and Control*, vol. 83, May 2023, doi: 10.1016/j.bspc.2023.104683.

[24] S. Vijayalakshmi, B. K. Pandey, D. Pandey, and M. E. Lelisho, “Innovative deep learning classifiers for breast cancer detection through hybrid feature extraction techniques,” *Scientific Reports*, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-06669-4.

[25] M. Wimmer *et al.*, “Multi-task fusion for improving mammography screening data classification,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 937–950, Apr. 2022, doi: 10.1109/TMI.2021.3129068.

[26] L. Sun *et al.*, “Two-view attention-guided convolutional neural network for mammographic image classification,” *CAAI Transactions on Intelligence Technology*, vol. 8, no. 2, pp. 453–467, Jun. 2023, doi: 10.1049/cit2.12096.

[27] Z. Cao, Z. Deng, Z. Yang, J. Ma, and L. Ma, “Supervised contrastive pre-training models for mammography screening,” *Journal of Big Data*, vol. 12, no. 1, Feb. 2025, doi: 10.1186/s40537-025-01075-z.

[28] H. Guan and M. Liu, “Domain adaptation for medical image analysis: a survey,” *IEEE Transactions on Biomedical Engineering,* vol. 69, no. 3, pp. 1173-1185, Mar. 2022, doi: 10.1109/TBME.2021.3117407.

[29] L. Garrucho, K. Kushibar, S. Jouide, O. Diaz, L. Igual, and K. Lekadir, “Domain generalization in deep learning based mass detection in mammography: a large-scale multi-center study,” *Artificial Intelligence in Medicine*, vol. 132, Jul. 2022, doi: 10.1016/j.artmed.2022.102386.

[30] B. Abdikenov, T. Zhaksylyk, A. Imasheva, Y. Orazayev, and T. Karibekov, “Innovative multi-view strategies for ai-assisted breast cancer detection in mammography,” *Journal of Imaging*, vol. 11, no. 8, Jul. 2025, doi: 10.3390/jimaging11080247.

[31] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, and O. Al-Kadi, “Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases,” *Journal of Imaging Informatics in Medicine*, vol. 15, no. 2, Jan. 2025, doi: 10.1007/s10278-025-01481-y.

[32] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, Jan. 1950, doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[33] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *34th International Conference on Machine Learning, ICML 2017*, 2017, pp. 2130–2143.

[34] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets,” *PLOS ONE*, vol. 10, no. 3, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[35] E. R. DeLong, D. M. DeLong, and D. L. C.-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach,” *Biometrics*, vol. 44, no. 3, Sep. 1988, doi: 10.2307/2531595.

[36] S. T. Buckland, A. C. Davison, and D. V. Hinkley, “Bootstrap methods and their application,” *Biometrics*, vol. 54, no. 2, Jun. 1998, doi: 10.2307/3109789.

[37] J. Suckling, J. Parker, D. Dance, S. Astley, and I. Hutt, “*Mammographic image analysis society (MIAS) database v1.21*,” University of Cambridge Repository, 2015, doi: 10.17863/CAM.105113.

[38] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, “A curated mammography data set for use in computer-aided detection and diagnosis research,” *Scientific Data*, vol. 4, 2017, doi: 10.1038/sdata.2017.177.

[39] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, Jun. 1967, doi: 10.1080/01621459.1967.10482916.

[40] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, Dec. 1945, doi: 10.2307/3001968.

BIOGRAPHIES OF AUTHORS

Muhammad Subali received his bachelor's degree in Physics from Universitas Indonesia in 1990, his master's degree in Electrical Engineering from Universitas Trisakti in 1997, and his doctorate from Universitas Gunadarma in 2007. He is a head lecturer in the Faculty of Informatics Engineering at Cendekia Abditama University, Tangerang, Banten, Indonesia. His primary research interests include signal and image processing, machine learning, artificial intelligence, internet of things (IoT), and deep learning. He can be contacted at email: subali@uca.ac.id.

Lulu Mawaddah Wisudawati received her bachelor's degree in Informatics Engineering from Gunadarma University, Indonesia, in 2008. She obtained a double master's degree in Information Systems Management from Gunadarma University, Indonesia, and in Computer Vision from Université de Bourgogne, France, in 2013. She earned her doctoral degree in Industrial Technology from Gunadarma University in 2020. She is currently a lecturer and researcher in the Department of Informatics at Gunadarma University. Her research interests include medical imaging, machine learning, artificial intelligence, internet of things (IoT), image processing, and deep learning. She can be contacted at email: lulu_mawadah@staff.gunadarma.ac.id.

Teresa received her bachelor's degree in Nursing Science (PSIK) from the Faculty of Medicine, Universitas Indonesia in 1992 and her master's degree in Leadership and Nursing Management from the Faculty of Nursing, Universitas Indonesia in 2020. She is currently a lecturer at the Faculty of Nursing, Universitas Cendekia Abditama, Tangerang, Banten, Indonesia, where she also serves as the head of the Research and Community Service Unit (UPPM). Her main research interests focus on nursing management, particularly in the relationship between information technology and the improvement of healthcare service quality in hospitals. She can be contacted at email: teresa@uca.ac.id.