Muhammad Hamza, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Explainable Modality-Specific DenseNet Models and Population-Level Decision Fusion for Breast Cancer Classification Across Mammography, Ultrasound, and Histopathology

Muhammad Hamza[1,2], Sufiyan Hamza[3]

[1]Dr. A.Q Khan Institute of Computer Science and Information Technology, University of Engineering and Technology Taxila, Pakistan

[2]Capital University of Science and Technology, Islamabad, Pakistan

[3]School of Information Engineering, Southwest University of Science and Technology, Mianyang, P.R. China

Abstract- Breast cancer diagnosis draws on complementary information from mammography, ultrasound, and histopathology, yet multimodal artificial-intelligence studies can become clinically ambiguous when unrelated public cohorts are fused as though they represented the same patients. This study developed an explainable framework based on independently trained DenseNet-121 specialists for the three imaging modalities and evaluated feature- and decision-level fusion as population-level ensemble experiments. The harmonized analytical collection contained 34,207 publicly available images: 24,576 mammograms, 1,722 ultrasound images, and 7,909 histopathology images. Modality-specific preprocessing, class weighting, augmentation, and ImageNet transfer learning were applied, while the histopathology branch additionally incorporated a domain-adversarial objective. Grad-CAM++ was used for qualitative auditing of the radiological specialists. From the detailed test confusion matrices, ultrasound achieved 90.8% accuracy, 88.2% sensitivity, and 92.9% specificity; mammography achieved 96.0% accuracy, 96.3% sensitivity, and 95.6% specificity; and histopathology achieved 94.8% accuracy, 96.5% sensitivity, and 91.5% specificity. Wilson 95% confidence intervals, balanced accuracy, and Matthews correlation coefficients were derived from these counts. The domain-adversarial histopathology ablation showed an approximately two-percentage-point increase in rounded accuracy, although the archived experimental record did not preserve the precise source-target domain assignment. Attention-based feature fusion achieved 94.2% accuracy and assigned the largest learned weights to histopathology and mammography. Across eight exploratory fixed decision-weight configurations, the highest observed population-level accuracy was 98.8% for a mammography-emphasized rule. Because modality predictions originated from unmatched cohorts and the fixed-weight sweep was not documented as using a separate selection set, fusion findings are interpreted as exploratory sensitivity analyses rather than same-patient clinical performance. The results support transparent modality-specific evaluation and motivate future validation on deduplicated, patient-linked, externally collected cohorts with calibrated probabilities and reproducible split manifests.

Keywords— Breast cancer; Mammography; Ultrasound; Histopathology; DenseNet-121; Domain-adversarial neural network; Grad-CAM++; Multimodal fusion; Explainable artificial intelligence

I. INTRODUCTION

Breast cancer remains one of the most frequently diagnosed malignancies in women and a major contributor to cancer-related mortality worldwide. Global estimates for 2022 reported approximately 2.3 million new cases and about 670,000 deaths, and the absolute burden is expected to rise as populations age and screening coverage changes [1-3]. Earlier detection and accurate characterization increase the likelihood of breast-conserving treatment, reduce the probability of advanced-stage presentation, and support more appropriate decisions regarding biopsy, surgery, systemic therapy, and surveillance. Nevertheless, breast

cancer is not represented by a single imaging phenotype. Lesions vary in size, density, margin, echogenicity, calcification pattern, cellular organization, and biological behavior, so contemporary diagnostic pathways rely on multiple sources of evidence rather than on a single examination.

Mammography is the principal population-screening modality and is particularly effective for identifying microcalcifications, architectural distortion, asymmetry, and spiculated masses. Its limitations are well recognized in dense breasts, where overlapping fibroglandular tissue can obscure lesions and reduce sensitivity [4-6]. Ultrasound is therefore frequently used as a complementary examination. It differentiates cystic from solid lesions, characterizes margins and internal echo patterns, evaluates posterior acoustic behavior, and guides needle biopsy without ionizing radiation. Ultrasound, however, is operator dependent and is strongly influenced by speckle, gain settings, probe pressure, lesion depth, and equipment variation [7]. Histopathology provides the microscopic evidence required for definitive diagnosis. It permits direct evaluation of nuclear atypia, gland formation, mitotic activity, stromal response, and invasive growth, but digitized tissue images vary with fixation, staining, scanner optics, magnification, and laboratory practice [8-12].

The complementary roles of these modalities have encouraged the development of computer-aided diagnosis systems based on convolutional neural networks, transfer learning, transformers, and multimodal fusion. Deep models can learn discriminative representations directly from images and have achieved high reported performance in mammography, ultrasound, and computational pathology [9,13-17]. DenseNet architectures are particularly attractive because dense connectivity promotes feature reuse, strengthens gradient propagation, and preserves information from earlier layers. These characteristics are useful when diagnostically relevant patterns range from fine microcalcifications to large masses and from nuclear morphology to tissue-level architecture. Transfer learning from ImageNet can further stabilize optimization when labelled medical data are limited. Recent multimodal and hybrid breast-imaging systems have reported high performance using feature, decision, and transformer-based integration strategies [18-31].

Despite rapid progress, published breast-imaging studies frequently contain methodological weaknesses that complicate interpretation. Medical repositories may include multiple views, magnifications, patches, or repeated examinations from the same patient. Random image-level splitting can place highly related samples in both development and test partitions and produce optimistic performance. Patient-level grouping is especially important for BreakHis, where each patient contributes images at four magnifications, and for mammography, where multiple projections may belong to one examination. Class and modality stratification are also necessary to maintain representative validation and test sets [32].

Label harmonization creates an additional challenge. Mammography and ultrasound repositories may include normal, benign, and malignant categories, whereas histopathology collections are commonly binary. Combining these data without a transparent rule can yield inconsistent targets or allow normal cases to dominate one modality. In a benign-versus-malignant study, normal examinations should not automatically be treated as benign tumors because the two categories have different clinical meanings. The final label mapping, exclusions, and original data provenance therefore require explicit documentation.

Multimodal fusion also demands careful interpretation. Fusion may occur at the input, feature, or decision level [33]. Early fusion can learn cross-modal interactions, but it is clinically meaningful only when inputs correspond to the same patient and preferably the same lesion. Concatenating a mammogram from one individual, an ultrasound image from another, and histology from a third creates a synthetic observation without a patient-level interpretation. Similar concerns apply to intermediate feature fusion when embeddings are aligned by class or index rather than by identity.

Decision-level fusion preserves the independent predictions and more closely resembles a multidisciplinary workflow, but even late fusion requires patient-linked probabilities when the reported result is intended to represent an individual diagnosis. With unmatched public cohorts, fusion should instead be described as a population-level ensemble experiment.

Domain shift is a further concern, particularly in histopathology. Stain concentration, scanner characteristics, tissue preparation, magnification, and laboratory protocols can change image appearance without changing the underlying diagnosis. Conventional stain normalization may reduce color variability but can also alter subtle morphology or introduce artifacts. Domain-adversarial neural networks offer an alternative by encouraging the feature extractor to retain diagnostic information while making acquisition-domain membership difficult to predict [10,11,34,35]. The approach is conceptually attractive, although its benefit must be interpreted in relation to clearly defined domains and external validation.

Interpretability is equally important. High classification accuracy does not guarantee that a network relies on clinically plausible evidence; models may exploit borders, labels, scanner artifacts, or dataset-specific backgrounds. Gradient-based class activation mapping provides a qualitative view of the spatial regions that influence a prediction and can be used to examine whether networks attend to lesion margins, calcifications, glandular structures, or nuclei [14,36-39]. Such maps support visual auditing but do not prove causal reasoning. Their value is greatest when evaluated against annotations and reviewed by domain specialists.

The present study was designed around these concerns. Separate DenseNet-121 specialists were developed for ultrasound, mammography, and histopathology, each with a modality-specific preprocessing and augmentation strategy. The histopathology branch incorporated an exploratory domain-adversarial objective, and Grad-CAM++ was used to visualize class-relevant regions in the radiological specialists. Attention-based feature fusion and two reproducibly specified decision-level rules - fixed weighting and majority voting - were examined after independent specialist evaluation. The analysis therefore addresses four questions: how accurately a common DenseNet backbone can classify each modality after modality-specific adaptation; whether the adversarial histopathology branch changes classification performance; whether explanation maps highlight plausible diagnostic structures; and how feature and decision fusion behave when specialist outputs are combined at the population level. Detailed test confusion matrices are used as the primary numerical evidence whenever source summaries contain inconsistent headline values.

II. MATERIALS AND METHODS

This retrospective computational study used publicly available, de-identified image repositories. No participants were recruited and no new clinical intervention was performed. The analytical unit for optimization was the image, while an available patient identifier was used for grouping whenever the source metadata supported it. The design consisted of modality-specific data preparation, stratified development of three specialist classifiers, qualitative explainability analysis, and exploratory fusion. Because the mammography, ultrasound, and histopathology repositories did not contain the same patients, specialist performance was evaluated independently and the fusion results were not treated as same-patient clinical accuracy.

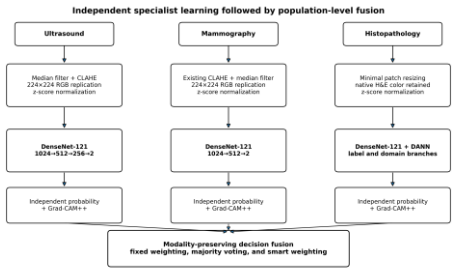

Figure 1: Overall study design. Three independently trained modality specialists generated predictions and Grad-CAM++ explanations before exploratory feature-level and population-level decision fusion.

The initial collection combined mammography, ultrasound, and histopathology data from established public repositories. Mammography images were assembled from INbreast, MIAS, DDSM, and CBIS-DDSM [31,40-42]. The consolidated source contained 26,602 preprocessed mammograms distributed as 13,710 malignant, 10,866 benign, and 2,026 normal images; normal images were excluded from the binary experiment, leaving 24,576 mammograms. Ultrasound images were obtained from the Breast Ultrasound Images dataset, BUS-UCLM, and BUS-UC [43-45]; after removal of normal images and label harmonization, the ultrasound cohort contained 969 benign and 753 malignant images. Histopathology used BreakHis, which contains benign and malignant breast-tumor images from 82 patients at 40x, 100x, 200x, and 400x magnification [46]. The binary histopathology cohort comprised 2,480 benign and 5,429 malignant images. Because CBIS-DDSM is a curated subset derived from DDSM, case-level overlap between the two mammography sources is a potential provenance concern. The available manifests did not permit a complete cross-source duplicate audit, so absence of overlap is not assumed.

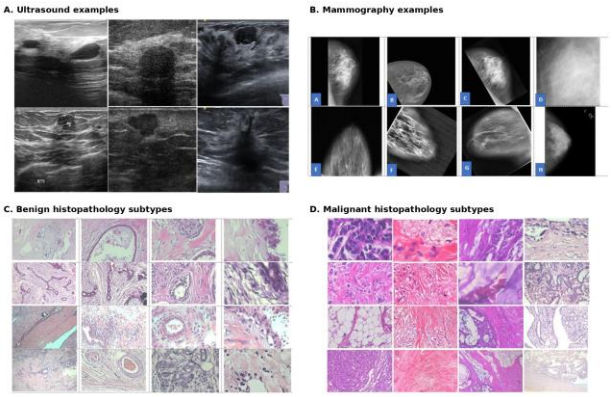

Figure 2: Representative images from the analytical collection: ultrasound, mammography, benign histopathology subtypes, and malignant histopathology subtypes. The panels illustrate the substantial differences in scale, contrast, texture, and biological content across modalities.

Original labels were mapped to two diagnostic classes, benign and malignant. Normal mammography and ultrasound images were excluded rather than merged with benign lesions because a normal examination and a pathologically benign tumor are clinically distinct. Each retained record stored the image path, modality, binary label, and patient identifier when available. A consolidated manifest was then generated to support consistent splitting and auditability. The final analytical collection contained 34,207 images, with mammography accounting for most observations and ultrasound representing the smallest modality.

Table 1. Distribution of the harmonized benign-versus-malignant image collection.

Modality	Benign	Malignant	Total	Primary sources
Mammography	10,866	13,710	24,576	INbreast, MIAS, DDSM/CBIS-DDSM
Ultrasound	969	753	1,722	BUSI, BUS-UCLM, BUS-UC
Histopathology	2,480	5,429	7,909	BreakHis, 40×-400×
Combined	14,315	19,892	34,207	Binary harmonized collection

The 2,026 normal mammography images and normal ultrasound images were excluded from the binary experiment.

Three train-validation-test allocations were examined: 80/10/10, 70/15/15, and 60/20/20. Images were first stratified by modality and diagnostic label. When an explicit patient identifier was available, all images belonging to that patient were retained in the same partition; this was essential for BreakHis because each patient contributes multiple tissue images and magnifications. For mammography and ultrasound sources in which repeated-patient mappings could not be recovered from the archived manifests, splitting was necessarily performed at the image-record level. Consequently, the corresponding evaluations cannot be assumed to be fully patient independent. Split quality was

assessed descriptively by comparing modality-label distributions with the complete collection through a chi-square statistic [32]. The 80/10/10 allocation had the smallest reported statistic and was used for subsequent experiments.

The selected split yielded 27,526 training images, 3,228 validation images, and 3,453 test images. Figure 3 shows the overall allocation and preservation of modality-label proportions. The 70/15/15 split approached the distribution of the full collection but had a larger chi-square statistic, whereas the 60/20/20 split showed the greatest deviation. Degrees of freedom and corresponding P values were not preserved in the study records, so the chi-square values are used only as descriptive split-comparison statistics rather than formal hypothesis tests.

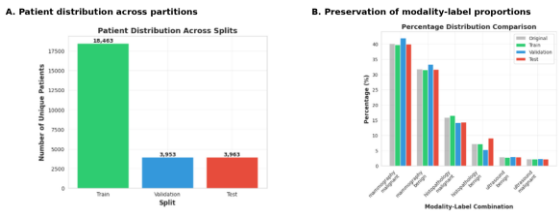

Figure 3: Evaluation of the data partition. (A) Numbers assigned to the training, validation, and test sets. (B) Comparison of modality-label proportions between the complete collection and the selected partitions.

Table 2. Candidate train-validation-test allocations and distributional comparison.

Split	Training	Validation	Test	Total	Reported χ^2
80/10/10	27,526	3,228	3,453	34,207	3.179
70/15/15	24,153	4,900	5,154	34,207	6.007
60/20/20	20,859	6,539	6,809	34,207	18.12

Preprocessing was deliberately modality specific because x-ray attenuation, ultrasound backscatter, and hematoxylin-and-eosin staining produce fundamentally different image statistics. Extensive manual segmentation and region-of-interest cropping were avoided to reduce dependence on observer-defined masks and to preserve anatomical context. All images were ultimately represented as 224×224×3 tensors compatible with ImageNet-initialized DenseNet-121. Global modality-specific means and standard deviations were estimated from the training data using Welford's online algorithm, and the same statistics were applied to validation and test images.

Ultrasound images were denoised with a 3×3 median filter to reduce high-frequency speckle while retaining lesion boundaries. Contrast-limited adaptive histogram equalization with a reported clipping value of 2 was then used to improve local contrast and emphasize subtle echogenicity differences. The grayscale images were resized to 224×224 pixels, replicated across three channels, and normalized by modality-specific z-scoring. Training augmentation included rotations of 90°, 180°, and 270° together with intensity or speckle-noise variation. Class weights of approximately 0.880 for benign and 1.144 for malignant images were used to reduce the influence of class imbalance. These operations are consistent with commonly used medical-image preprocessing strategies [47].

The mammography collection had already undergone CLAHE enhancement before model development, so subsequent processing was restricted to median filtering, resizing, three-channel replication, and z-score normalization. Pectoral-muscle removal, breast masking, and lesion cropping were not performed. This choice preserved global architectural context but also left acquisition borders and labels available to the network, an issue that was later examined qualitatively through Grad-CAM++. Augmentation consisted of horizontal and vertical reflection and rotations of 90°, 180°, and 270°. The reported class weights were approximately 1.131 for benign and 0.896 for malignant images. Median filtering and restrained normalization were selected to avoid repeated aggressive contrast manipulation [48].

Histopathology images were resized into 224x224 RGB patches and normalized without aggressive stain standardization. Macenko or Reinhard-type transformations were avoided to preserve native hematoxylin-and-eosin color and texture relationships. Horizontal and vertical flipping, rotation, and color jitter expanded the training distribution, and class weights of approximately 1.613 for benign and 0.725 for malignant images were applied. Inclusion of all four BreakHis magnifications increased morphological diversity but also increased within-patient correlation, reinforcing the importance of patient grouping.

$$Z = \frac{x - \mu_m}{\sigma_m} \qquad (1)$$

In Equation (1), x denotes a pixel intensity and μ_m and σ_m are the mean and standard deviation estimated from the training images of modality m.

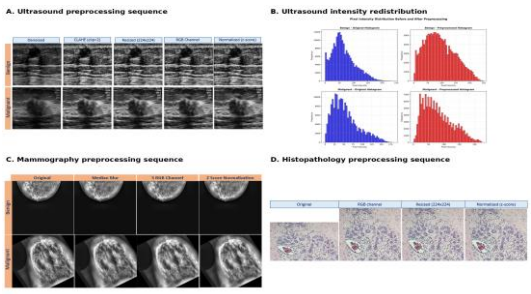

Figure 4: Modality-specific preprocessing. (A) Ultrasound denoising, local contrast enhancement, resizing, channel conversion, and normalization. (B) Ultrasound intensity distributions before and after processing. (C) Mammography preprocessing. (D) Minimal histopathology processing that retains native stain relationships.

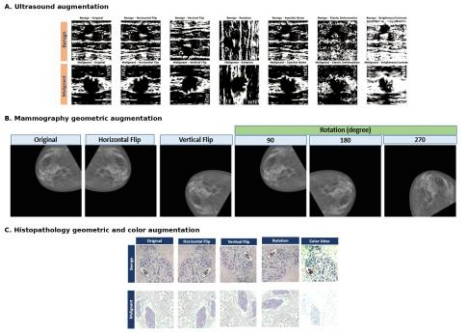

Figure 5: Representative training augmentation for ultrasound, mammography, and histopathology. The transformations were selected to increase acquisition and geometric diversity without changing the diagnostic label.

Each specialist used an ImageNet-1K initialized DenseNet-121 backbone. Input images passed through the initial convolution and pooling layers, four densely connected blocks, and the associated transition layers. The final dense block generated a 7×7×1,024 feature tensor, and global average pooling converted this tensor to a 1,024-dimensional embedding. Dense connectivity was chosen because it reuses earlier features, facilitates gradient propagation, and supports simultaneous representation of fine and coarse patterns. The final spatial feature map was retained for Grad-CAM++ visualization. DenseNet-121 was selected for its dense feature reuse and gradient propagation properties [49].

The ultrasound specialist used a multilayer classifier with dimensions 1,024→512→256→2. Linear transformations were followed by rectified-linear activations and dropout. The network was trained for 20 epochs with a batch size of 32; the learning rate was reduced to 5×10^{-5} after epoch 12. The initial learning rate and exact dropout probabilities were not stated in the study record and are therefore not inferred. The mammography specialist used a 1,024→512→2 head with reported dropout probabilities of 0.5 and 0.4. It was optimized using Adam at an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, cosine-annealing scheduling, a batch size of 16, and early-stopping patience of 10 epochs.

The histopathology specialist combined DenseNet-121 with a diagnostic label branch and a binary domain-discrimination branch. The label head used a 1,024→512→2 architecture. The domain branch applied a gradient-reversal layer to the pooled embedding and then used a 1,024→8→2 classifier. During the forward pass the gradient-reversal layer acts as the identity, whereas during backpropagation it reverses the domain gradient so that the feature extractor is optimized to retain diagnostic information while reducing domain-discriminative information [34,35]. The adaptation coefficient

increased gradually from zero to a cap of 0.3, and optimization used Adam at 1x10-4 with a batch size of 8 for 15 epochs. The study documentation identifies the discriminator as source versus target patient domains but does not specify the rule used to assign patients to those domains. The DANN experiment is therefore treated as an exploratory adversarial-regularization ablation rather than conclusive evidence of stain-, scanner-, or institution-invariant generalization.

$$\Delta\theta_f = \frac{\partial L_y}{\partial \theta_f} - \lambda \frac{\partial L_d}{\partial \theta_f} \qquad (2)$$

$$J = \frac{1}{N}\sum_{i=1}^{N} L_y\left(G_y\left(G_f(x_i)\right), y_i\right) - \lambda \frac{1}{N}\sum_{i=1}^{N} L_d\left(G_d\left(G_f(x_i)\right), d_i\right) \qquad (3)$$

Equations (2) and (3) express the adversarial update. L_y is the diagnostic loss, L_d is the domain loss, θ_f denotes the feature-extractor parameters, and λ controls the contribution of the domain objective.

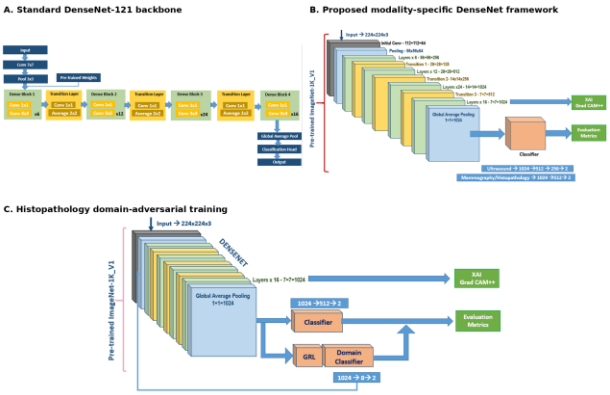

Figure 6: Network design. (A) Standard DenseNet-121 backbone. (B) Modality-specific DenseNet framework with global average pooling and classifier heads. (C) Histopathology domain-adversarial training with diagnostic and domain branches.

Table 3. Specialist-model configuration.

Element	Ultrasound	Mammography	Histopathology
Input	224×224×3	224×224×3	224×224×3
Backbone	DenseNet-121	DenseNet-121	DenseNet-121 + DANN
Head	1024→512→256→2	1024→512→2	Label 1024→512→2; domain 1024→GRL→8→2
Class weights	0.880 / 1.144	1.131 / 0.896	1.613 / 0.725
Batch size	32	16	8
Optimization	20 epochs; LR reduced after epoch 12	Adam; LR 10^{-4}; weight decay 10^{-4}; cosine annealing	Adam; LR 10^{-4}; 15 epochs; λ ≤ 0.3

Grad-CAM++ was applied to the final 7x7x1,024 convolutional feature map to obtain class-discriminative heat maps [50]. Class-specific weights were derived from higher-order gradient information, the weighted feature maps were combined and rectified, and the resulting heat map was resized to the input-image dimensions. Qualitative interpretation was retained only for ultrasound and mammography because those visual outputs could be verified against the corresponding source modality. The available histopathology heat-map panel duplicated ultrasound-like source images and was therefore excluded from microscopic interpretation. No lesion-mask overlap, pointing-game score, insertion-deletion test, or blinded specialist review was available, so Grad-CAM++ is treated as qualitative model auditing rather than localization validation.

For feature-level fusion, the 1,024-dimensional embeddings from ultrasound, mammography, and histopathology were concatenated to form a 3,072-dimensional vector. Because patient identities did not correspond across modalities, each modality was balanced and trimmed to the same class counts and then aligned by class and index rather than by person. An attention mechanism learned modality weights, and a feed-forward classifier reduced the fused representation through 3,072→1,024→512→2. This experiment was designed as a representation-level sensitivity analysis and not as patient-level multimodal validation.

Decision fusion combined two-class probabilities from the independent specialists. The available prediction records were balanced to 96 benign and 75 malignant observations, producing 171 population-level fusion instances. Fixed weighted voting computed a convex combination of modality probabilities, and eight pre-specified weight vectors were compared. Majority voting assigned the class selected by at least two specialists. Only these fully specified rules are retained in the primary analysis. The experimental record does not document a separate validation-only stage for selecting among the eight fixed-weight vectors; consequently, the fixed-weight table is interpreted as an exploratory sensitivity analysis rather than an unbiased optimized test estimate. Because the modality probabilities originated from different patient cohorts, all decision-fusion results remain population-level ensemble simulations.

$$P_f(c) = \sum_m w_m p_m(c)\,, \sum_m w_m = 1 \quad (4)$$

In Equation (4), pm(c) is the probability assigned to class c by modality m and wm is its non-negative fusion weight, with the weights constrained to sum to one.

Malignant disease was treated as the positive class. Accuracy, sensitivity, specificity, precision, F1 score, negative predictive value (NPV), and area under the receiver-operating-characteristic curve (AUROC) were used for evaluation. Threshold-dependent metrics were recalculated directly from the detailed confusion matrices whenever those counts were available; AUROC values were retained from the recorded evaluation plots or tables because they cannot be reconstructed from a single confusion matrix. Classification performance was evaluated using accuracy, sensitivity, specificity, precision, and F1 score, which were calculated as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, (5)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN},\ \text{Specificity} = \frac{TN}{TN+FP}, (6)$$

$$\text{Precision} = \frac{TP}{TP+FP},\ \text{F1} = \frac{2TP}{2TP+FP+FN}, (7)$$

$$\text{NPV} = \frac{TN}{TN+FN}, (8)$$

$$\text{Balance Accuracy} = \frac{\text{Sensitivity}+\text{Specificity}}{2}, (9)$$

$$\text{MCC} = \frac{TP\ TN - FP\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \quad (10)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Wilson 95% confidence intervals were calculated post hoc from the confusion-matrix counts for accuracy, sensitivity, and specificity. Balanced accuracy and MCC were derived from the same binary counts to provide class-imbalance-aware summaries. These uncertainty estimates quantify binomial sampling uncertainty for the recorded predictions; they do not account for model-training variability because repeated independent training runs were not available. For the synthetic fusion experiment, confidence intervals are descriptive only and should not be interpreted as patient-level clinical uncertainty.

Experiments were conducted in Python with PyTorch in Google Colab using an NVIDIA A100 GPU and approximately 12.7 GB of system memory. Exact package versions, CUDA version, random seeds, mixed-precision settings, checkpoint-selection rules, and the number of independent repetitions were not preserved in the experimental record. The reported values therefore represent a single experimental realization and require reproducible rerunning with

fixed seeds and deposited split manifests before definitive external validation.

III. RESULTS

The final binary collection was markedly imbalanced by modality. Mammography contributed 71.8% of all images, histopathology 23.1%, and ultrasound 5.0%. Malignant images represented 58.2% of the complete collection, largely because the histopathology and mammography cohorts contained more malignant than benign observations. The selected 80/10/10 partition retained representation of all modality-label combinations and produced the smallest reported chi-square deviation from the original collection. These distributions are important when interpreting cross-modality comparisons because each specialist was trained on a different number of observations and on a different case mix.

The ultrasound learning curves showed progressive improvement across 20 epochs. Training and validation accuracy rose together, while training and validation losses declined. Validation loss fluctuated more strongly than training loss, which is expected for the smallest modality cohort, but the final curves did not show the persistent divergence that would indicate severe overfitting. The validation summary reported accuracy close to 0.89, F1 near 0.90, and AUROC around 0.95 before final test evaluation.

The mammography specialist converged rapidly and benefited from the largest training cohort. Training loss decreased from approximately 0.33 to below 0.05, and validation loss remained comparatively low. Training and validation accuracy approached the upper 0.90 range. Precision, recall, F1 score, and AUROC stabilized at high levels, while the cosine learning-rate schedule decayed smoothly. The narrow gap between training and validation curves suggested stable optimization within the selected split, although independent external data would be required to determine whether the model generalized beyond the consolidated repositories.

The histopathology model was trained for 15 epochs. Accuracy improved during early optimization and remained in the mid-to-high 0.80 range during training, while validation accuracy fluctuated around higher values. Training loss decreased consistently. The adaptation coefficient rose smoothly toward its cap of 0.3, allowing the diagnostic branch to establish class separation before the adversarial objective reached its full strength. The reported curves were compatible with stable optimization, but the undefined domain labels prevent a precise interpretation of what acquisition variability was actually suppressed.

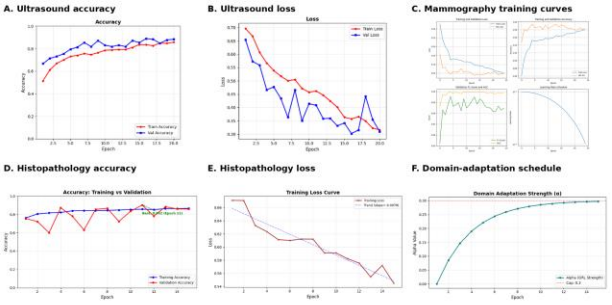

Figure 7: Optimization behavior of the modality specialists. Panels summarize ultrasound accuracy and loss, the four-panel mammography training record, histopathology accuracy and loss, and the gradual increase of the domain-adaptation coefficient.

Final test performance was derived from the detailed confusion matrices. The ultrasound matrix contained 91 true negatives, 7 false positives, 9 false negatives, and 67 true positives. The model correctly classified 158 of 174 images, giving 90.8% accuracy. Sensitivity was 88.2%, specificity 92.9%, precision 90.5%, F1 score 89.3%, and negative predictive value 91.0%. The reported AUROC was 0.946. The 16 errors were divided relatively evenly between missed malignant lesions and benign lesions incorrectly classified as malignant, indicating a balanced but non-negligible error profile.

The mammography matrix contained 1,040 true negatives, 48 false positives, 50 false negatives, and 1,317 true positives. These counts correspond to 2,455 classified test images and yield 96.0% accuracy, 96.3% sensitivity, 95.6% specificity, 96.5% precision, 96.4% F1 score, and 95.4% negative predictive value; the recorded AUROC was 0.970. A separate split summary listed 2,459 mammography

test records, leaving a four-image discrepancy that cannot be reconciled from the available files. Accordingly, all recalculated mammography metrics in this manuscript use the 2,455 observations represented by the confusion matrix. The near equality of false-positive and false-negative counts indicates a balanced operating point. The strong result is consistent with the large training cohort and the visibility of architectural and calcification patterns, although the mixture of multiple repositories raises the possibility that source-specific cues contributed to performance.

The histopathology matrix contained 258 true negatives, 24 false positives, 19 false negatives, and 519 true positives. Accuracy was 94.8%, sensitivity 96.5%, specificity 91.5%, precision 95.6%, F1 score 96.0%, and negative predictive value 93.1%. The reported AUROC was approximately 0.992. Histopathology achieved the highest sensitivity and AUROC but produced more false positives than false negatives, suggesting an operating point that favored malignant-case detection. Such a bias may be acceptable for screening or triage but would need calibration and threshold analysis before clinical use.

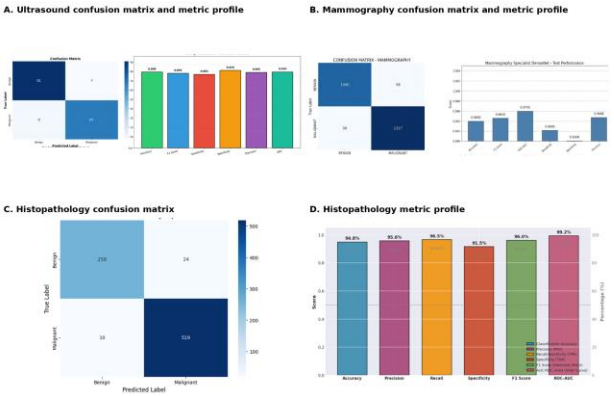

Figure 8: Independent specialist test performance. Ultrasound, mammography, and histopathology confusion matrices are accompanied by the corresponding evaluation profiles.

Across independent specialists, mammography achieved the highest overall accuracy and the most balanced sensitivity-specificity combination. Histopathology provided the highest sensitivity and AUROC, while ultrasound showed the lowest accuracy but still exceeded 90% despite the smallest cohort and the substantial acquisition variability associated with sonography. These differences reflect both modality information content and dataset composition; they should not be interpreted as a direct comparison of the intrinsic diagnostic value of the clinical examinations.

Table 4. Test performance recalculated from the detailed confusion matrices.

Model	n	Accuracy %	Sensitivity %	Specificity %	Precision %	F1 %	NPV %	AUROC
Ultrasound	174	90.8	88.2	92.9	90.5	89.3	91.0	0.946
Mammography	2,455	96.0	96.3	95.6	96.5	96.4	95.4	0.970
Histopathology	820	94.8	96.5	91.5	95.6	96.0	93.1	0.992
Highest observed fixed-weight fusion	171	98.8	98.7	99.0	98.7	98.7	99.0	1.000

AUROC values were taken from the recorded evaluation plots or fusion tables. Fusion observations were assembled from unmatched cohorts and do not represent same-patient clinical performance. Mammography threshold-dependent metrics use n=2,455 because that is the number of predictions represented by the detailed confusion matrix. A separate split summary listed 2,459 mammography test records; the four-image difference could not be reconciled from the available files and is disclosed as a source-record discrepancy.

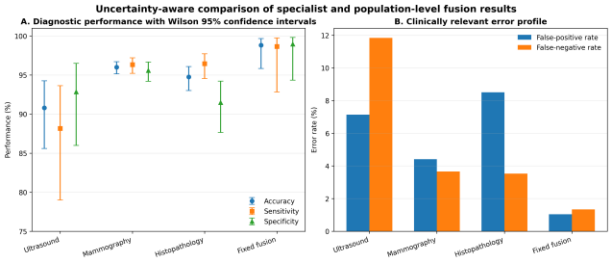

Figure 9: Uncertainty-aware comparison of the three independent specialists and the highest observed fixed-weight population-level fusion configuration. (A) Accuracy, sensitivity, and specificity with Wilson 95% confidence intervals derived from confusion-matrix counts. (B) False-positive and false-negative rates. The fusion interval is descriptive because the modality cohorts were not patient linked.

The uncertainty analysis clarifies the effect of unequal test-set sizes. Ultrasound accuracy was 90.8% (95% CI, 85.6-94.3%), sensitivity 88.2% (79.0-93.6%), and specificity 92.9% (86.0-96.5%). Mammography yielded 96.0% accuracy (95.2-96.7%), 96.3% sensitivity (95.2-97.2%), and 95.6% specificity (94.2-96.7%). Histopathology yielded 94.8% accuracy (93.0-96.1%), 96.5% sensitivity (94.6-97.7%), and 91.5% specificity (87.6-94.2%). The highest observed fixed-weight population-level fusion configuration yielded 98.8% accuracy (95.8-99.7%), 98.7% sensitivity (92.8-99.8%), and 99.0% specificity (94.3-99.8%). Balanced accuracy/MCC were 90.5%/0.813 for ultrasound, 96.0%/0.919 for mammography, 94.0%/0.883 for histopathology, and 98.8%/0.976 for the best fixed-weight fusion. The wider ultrasound intervals primarily reflect its much smaller test cohort, while the error-rate panel shows that histopathology had the largest false-positive rate and ultrasound the largest false-negative rate among the independent specialists.

The histopathology ablation showed approximately 92% test accuracy for DenseNet-121 without the adversarial branch and approximately 94% after adding DANN; the detailed final DANN confusion matrix yielded 94.8% accuracy. The direction of change is consistent with a beneficial regularization effect within the selected split. However, because the source-target patient-domain assignment was not preserved and the experiment was based on one training realization, the result should not be interpreted as proof of general stain, scanner, or institutional invariance.

Grad-CAM++ maps showed modality-dependent attention patterns in ultrasound and mammography. Ultrasound activation was concentrated around lesion margins, internal echogenicity, and posterior acoustic regions, whereas mammography maps highlighted dense masses and surrounding structural regions. These patterns are qualitatively plausible but are not equivalent to localization accuracy. Histopathology explanations are not interpreted because the available panel could not be verified as H&E tissue.

Figure 10 therefore presents only the ultrasound and mammography explanation outputs retained for qualitative auditing. No microscopic histopathology localization claim is made from the unavailable or unverified heat-map output.

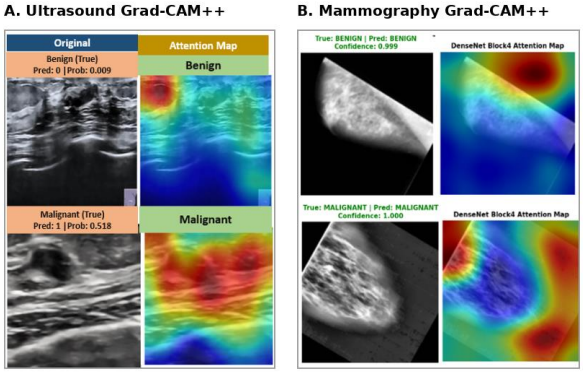

Figure 10: Grad-CAM++ examples for the ultrasound and mammography specialists. Warmer colors indicate stronger class-related activation. The maps are presented for qualitative auditing and are not interpreted as quantitative lesion-localization evidence.

The attention-based feature-fusion experiment concatenated the three 1,024-dimensional embeddings and trained a 3,072→1,024→512→2 classifier. Training accuracy approached 100%, while validation accuracy stabilized around the low-to-mid 0.90 range. The fused confusion matrix contained 95 true negatives, 1 false positive, 9 false negatives, and 66 true positives. Accuracy was 94.2%, sensitivity

88.0%, specificity 99.0%, precision 98.5%, F1 score approximately 92.9%, and AUROC 0.970. The large difference between specificity and sensitivity indicates that the model was conservative in assigning malignancy. The learned modality weights were 0.173 for ultrasound, 0.397 for mammography, and 0.430 for histopathology, suggesting that the attention module placed greatest emphasis on microscopic morphology and mammographic structure.

The feature-fusion result should be interpreted as an analytical benchmark rather than a clinical multimodal score because embeddings were aligned by class and index rather than by patient identity. The experiment remains useful because it quantifies the relative contribution of the modality representations and demonstrates that naive concatenation is not automatically superior to the strongest individual specialist. Its 94.2% accuracy was lower than the independent mammography accuracy, and its sensitivity was substantially lower than that of mammography or histopathology.

Eight fixed decision-weight combinations were evaluated as a sensitivity analysis. Equal weighting and a 0.30/0.40/0.30 ultrasound/mammography/histopathology allocation each produced 98.42% recorded accuracy, F1 score 0.9833, and AUROC 0.9985. A 0.40/0.30/0.30 allocation and a histopathology-emphasized 0.20/0.30/0.50 allocation each produced 98.25% accuracy. The highest observed configuration used 0.20/0.50/0.30, emphasizing mammography, and produced the confusion matrix TN=95, FP=1, FN=1, TP=74. Directly from these counts, accuracy was 98.8%, sensitivity 98.7%, specificity 99.0%, precision 98.7%, and F1 score 98.7%. Because a separate weight-selection set was not documented, this maximum is reported descriptively rather than as an independently selected optimum.

Majority voting produced 96 true negatives, no false positives, 3 false negatives, and 72 true positives. Accuracy was 98.2%, sensitivity 96.0%, specificity 100%, precision 100%, and F1 score approximately 98.0%. Compared with the highest observed fixed-weight rule, majority voting eliminated false positives but produced two additional false negatives.

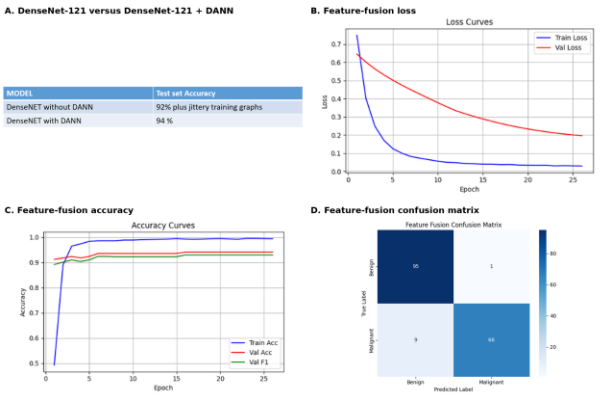

Figure 11: Histopathology and feature-fusion analyses. (A) DANN ablation. (B-C) Feature-fusion loss and accuracy curves. (D) Feature-fusion confusion matrix.

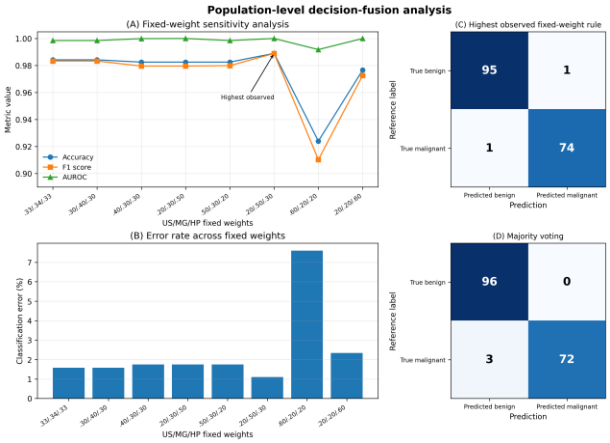

Figure 12: Population-level decision-fusion analysis. (A-B) Accuracy/F1/AUROC and error-rate sensitivity across the eight fixed weight vectors. (C) Confusion matrix for the highest observed fixed-weight configuration. (D) Confusion matrix for majority voting.

Table 5. Attention-based feature-fusion findings.

Feature-fusion result	Value
Attention weights	Ultrasound 0.173; mammography 0.397; histopathology 0.430
Confusion matrix	TN 95; FP 1; FN 9; TP 66
Accuracy	94.2%
Sensitivity / specificity	88.0% / 99.0%
Reported AUROC	0.970

The three embeddings were aligned synthetically by class and index and did not represent the same patients.

Table 6. Fixed decision-fusion weight sensitivity analysis.

US	MG	HP	Accuracy	F1 score	AUROC
0.33	0.34	0.33	0.9842	0.9833	0.9985
0.30	0.40	0.30	0.9842	0.9833	0.9985
0.40	0.30	0.30	0.9825	0.9796	0.9999
0.20	0.30	0.50	0.9825	0.9796	1.0000
0.50	0.30	0.20	0.9825	0.9799	0.9985
0.20	0.50	0.30	0.9890	0.9889	1.0000
0.60	0.20	0.20	0.9240	0.9103	0.9918
0.20	0.20	0.60	0.9766	0.9726	1.0000

US, ultrasound; MG, mammography; HP, histopathology. The table summarizes the eight fixed-weight configurations evaluated on the balanced 171-instance population-level prediction set. For the 0.20/0.50/0.30 configuration, the tabulated source summary gives F1=0.9889, whereas direct recalculation from TN=95, FP=1, FN=1, TP=74 gives F1=0.9867; confusion-matrix-derived values are used in the prose. No separate validation-only weight-selection stage was documented, so the table is interpreted as a sensitivity analysis.

Both reproducibly documented decision-level rules - fixed weighting and majority voting - exceeded the individual specialist accuracies within the balanced population-level simulation. The largest numerical gain was observed relative to ultrasound, but this does not imply an equivalent benefit for an individual patient because the fusion instances were assembled from unmatched modality cohorts. The principal methodological value of late fusion is preservation of independent specialist outputs; a clinically valid implementation would require calibrated probabilities from the same patient and diagnostic episode.

Published multimodal breast-imaging studies report a wide range of high accuracies, but direct numerical ranking is inappropriate because datasets, class definitions, magnifications, validation protocols, and fusion assumptions differ substantially [18-31]. The present work is distinguished by explicitly separating independent specialist evaluation from exploratory fusion and by not interpreting unmatched public cohorts as same-patient multimodal evidence.

IV. DISCUSSION

This study developed a common analytical framework for three fundamentally different breast-imaging modalities while preserving their independence during specialist training and testing. A shared ImageNet-initialized DenseNet-121 backbone provided a consistent feature-extraction foundation, whereas preprocessing, augmentation, classifier heads, class weights, and domain adaptation were adapted to modality characteristics. Mammography produced the highest overall accuracy, histopathology produced the highest sensitivity and AUROC, and ultrasound achieved a balanced result despite a substantially smaller and more heterogeneous cohort. Population-level decision fusion yielded the highest numerical accuracy, but the absence of patient linkage across modalities prevents interpretation as a clinical multimodal diagnostic score.

The modular design is an important strength. Independent specialists preserve the evidence produced by each modality and are compatible with settings in which only a subset of examinations is available. At the same time, the present data assembly highlights a central reproducibility issue: image repositories can contain multiple views, repeated cases, derivative datasets, and incomplete patient identifiers. In particular, CBIS-DDSM is derived from DDSM, and the available manifests did not permit a definitive cross-source case-overlap audit. The mammography and ultrasound results should therefore be interpreted as image-level results for records lacking recoverable patient identifiers, while BreakHis grouping was explicitly patient based. A future definitive evaluation should deduplicate all source repositories and enforce patient-level grouping before any train-test split.

The ultrasound result reflects both the value and the difficulty of sonographic classification. Sensitivity of 88.2% and specificity of 92.9% indicate that the

network learned useful lesion morphology, but the nine false-negative and seven false-positive cases show that difficult lesions remained. Sonographic appearance changes with equipment, gain, probe orientation, compression, and lesion depth. Median filtering and CLAHE improved local consistency, yet overly aggressive enhancement can alter subtle texture. Future evaluation should therefore include multiple institutions and devices and should compare preprocessing alternatives under a fixed patient-level split.

Mammography benefited from the largest training cohort and from previously enhanced source images. Its false-positive and false-negative counts were almost identical, indicating a balanced decision threshold. The performance is encouraging, but the consolidated dataset combined several repositories with different acquisition eras, formats, resolutions, and preprocessing histories. Without source-stratified external testing, the network may partly exploit dataset signatures. The use of uncropped full images preserved clinically important context, but it also allowed labels, borders, and acquisition artifacts to remain. Explanation maps provide a preliminary audit, but a more rigorous analysis should quantify attention inside breast tissue and lesion annotations. Histopathology achieved high sensitivity and AUROC but lower specificity than mammography. The network correctly detected most malignant images and generated comparatively more false-positive than false-negative predictions. In a diagnostic triage setting, a malignant-sensitive threshold may be reasonable, but clinical deployment would require probability calibration and decision-threshold analysis. The use of all BreakHis magnifications exposes the model to both cellular and architectural information; it also increases within-patient correlation, making the patient grouping strategy essential.

The DANN ablation is promising but not conclusive. The approximately two-percentage-point improvement in rounded accuracy is compatible with a useful adversarial regularization effect, but the precise rule used to define source and target patient domains was not preserved, no unseen external domain was evaluated, and repeated-run uncertainty was unavailable. The result is therefore best interpreted as an exploratory architectural ablation rather than evidence of stain or institution invariance. A stronger domain-generalization study should define domain labels prospectively and evaluate on a held-out acquisition domain.

The Grad-CAM++ analysis improves transparency but must be described conservatively. Ultrasound and mammography maps highlighted plausible lesion and architectural regions, supporting qualitative auditing of whether the radiological specialists relied exclusively on irrelevant background. However, heat maps are spatially coarse and can be sensitive to layer choice and implementation. Because a verifiable H&E explanation panel was not available, histopathology localization is not claimed. Future work should regenerate microscopic explanations, include correct and incorrect predictions, and assess localization with expert review or annotation-based metrics.

The feature-fusion experiment provided information about relative modality contribution. Histopathology and mammography received the largest attention weights, consistent with their stronger independent results. Nevertheless, the feature vectors were not patient linked, and training accuracy approaching 100% while validation accuracy remained near 93% suggests a degree of overfitting. The fused sensitivity was also lower than that of the strongest specialists. These observations reinforce the decision not to present the feature result as a clinically valid multimodal model.

Decision fusion produced the highest numerical values in the exploratory balanced population-level experiment. The highest observed fixed weighting emphasized mammography, while majority voting achieved perfect specificity at the cost of additional false negatives. These comparisons are useful for characterizing how different aggregation rules trade sensitivity against specificity, but they are not an unbiased estimate of an optimized clinical ensemble because the modality predictions were not patient linked and no separate fixed-weight selection set was documented. Late fusion remains methodologically attractive because it preserves

each modality's output and can support missing-modality workflows once same-patient calibrated probabilities are available.

Probability calibration is central to future fusion. A probability of 0.8 should have comparable empirical meaning across specialists before weighted averaging is interpreted diagnostically. Reliability diagrams, Brier scores, expected calibration error, and temperature scaling should therefore accompany any patient-linked fusion experiment. Calibration should be performed on a validation cohort and evaluated on an untouched external test cohort.

Reproducibility remains the principal limitation of the present retrospective study. A definitive replication should release patient/case-level split manifests, cross-source duplicate checks, preprocessing parameters, random seeds, package versions, checkpoint-selection rules, and prediction files. The archived mammography records contain a four-image difference between the split summary (2,459) and the detailed confusion-matrix total (2,455), and the current analysis conservatively uses the latter. Repeated model training with fixed seeds and bootstrap confidence intervals would quantify training and sampling variability more completely.

The study is limited by retrospective public repositories, incomplete patient identifiers in parts of the mammography and ultrasound collections, possible unresolved DDSM/CBIS-DDSM case overlap, absence of patient alignment across modalities, a single reported development split, incomplete source-target documentation for DANN, and lack of calibration or prospective validation. No lesion-localization metric, subgroup analysis, reader study, or decision-curve analysis was performed. The fusion experiment used balanced synthetic cross-cohort samples and did not document an independent weight-selection stage; it therefore does not estimate real-world disease prevalence or same-patient multimodal benefit. These limitations define the boundary between an experimental research framework and a clinically validated tool.

The next phase should assemble a patient-linked cohort containing mammography and ultrasound, with histopathology available for biopsied lesions. The specialists should first be externally evaluated and calibrated. Fusion should then operate on probabilities belonging to the same patient, with explicit handling of missing modalities. Performance should be reported by breast density, lesion type, age group, acquisition center, scanner, and histopathology subtype. A prospective reader study could assess whether explanations and fused probabilities improve sensitivity, specificity, reading time, or decision confidence beyond standard clinical interpretation.

Table 7. Contextual comparison with selected multimodal breast-imaging studies.

Study	Modalities	Reported performance	Important design consideration
Atrey et al. [18]	Mammography + ultrasound	98.84% combined	Feature-level fusion; moderate dataset
Habib et al. [19]	Mammography + ultrasound	94.0% combined	Joint neural analysis
Deb et al. [20]	Three modalities	99.96% combined	Potential cross-patient fusion concern
Nawaz et al. [21]	Three modalities	98.40%-99.12%	Explainable attention; selected histology magnification
Pacal and Attallah [22]	Multimodal	High reported accuracy	Transformer-based representation
Present study	Three independent specialists	US 90.8%; MG 96.0%; HP 94.8%; fusion 98.8%	Independent specialist evaluation; exploratory population-level fusion

V. CONCLUSION

This study presents an explainable framework for breast-cancer classification across mammography, ultrasound, and histopathology using modality-specific DenseNet-121 specialists. From the available detailed confusion matrices, ultrasound achieved 90.8% accuracy, mammography 96.0%, and histopathology 94.8%, with complementary sensitivity-specificity profiles. The histopathology DANN branch showed an encouraging exploratory ablation effect, while Grad-CAM++ enabled qualitative auditing of ultrasound and mammography attention patterns.

Attention-based feature fusion assigned the largest representation-level weights to histopathology and mammography but remained a class/index-aligned experiment rather than patient-level fusion. Fixed-weight and majority-vote decision rules produced high population-level values; the highest observed fixed-weight configuration yielded 98.8% confusion-matrix-derived accuracy. Because modalities came from unmatched cohorts and the fixed-weight sweep lacked a documented independent selection set, these figures are interpreted as exploratory ensemble sensitivity results rather than clinical multimodal performance.

The framework provides a useful foundation because it is modular and explicitly separates specialist performance from cross-cohort fusion. Its next stage should prioritize data provenance rather than additional architectural complexity: patient-level and duplicate-free splitting, reconciliation of the mammography evaluation count, explicit DANN domain construction, reproducible code and manifests, probability calibration, repeated training, quantitative explainability assessment, and external patient-linked validation. Only after these steps should fusion be evaluated as a clinical decision-support strategy.

Author Contributions: Muhammad Hamza; Conceptualization, Investigation, Methodology, Writing – Original Draft, Writing – Review & Editing, Sufiyan Hamza; Conceptualization, Supervision, Project administration, Resources, Writing – Review & Editing.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

REFERENCES

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229-263.
2. American Cancer Society. Breast Cancer Facts & Figures. Atlanta: American Cancer Society; 2024.
3. Zhang Y, Ji Y, Liu S, et al. Global burden of female breast cancer. J Natl Cancer Cent. 2025.
4. Gerami R, Joni SS, Akhondi N, et al. A literature review on imaging methods for breast cancer: mammography, histopathology, ultrasound, and MRI. Int J Physiol Pathophysiol Pharmacol. 2022;14(3):171.
5. Rehman H, Ahmad I, Rashid S, Mukhtar M, Khan AA, Khaliq H. Comparison of diagnostic accuracy of ultrasound and mammography in detecting breast cancer in radiographically dense breasts. Cureus. 2025.
6. Devolli-Disha E, Manxhuka-Kerliu S, Ymeri H, Kutllovci A. Comparative accuracy of mammography and ultrasound in women with breast symptoms according to age and breast density. Bosn J Basic Med Sci. 2009;9(2):131-136.
7. Tenajas R, Miraut D, Illana CI, et al. Recent advances in artificial intelligence-assisted ultrasound scanning. Appl Sci. 2023;13(6):3693.
8. Tseng LJ, Matsuyama A, MacDonald-Dickinson V. Histology and histopathology as the diagnostic standard for breast cancer. Can Vet J. 2023;64(4):389.
9. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. Med Image Anal. 2021;67:101813.
10. Cooper M, Ji Z, Krishnan RG. Machine learning in computational histopathology: challenges and

opportunities. Genes Chromosomes Cancer. 2023;62(9):540-556.
11. Veta M, Pluim JP, Van Diest PJ, Viergever MA. Breast cancer histopathology image analysis: a review. IEEE Trans Biomed Eng. 2014;61(5):1400-1411.
12. Hanby AM. The pathology of breast cancer and the role of the histopathology laboratory. Clin Oncol. 2005;17(4):234-239.
13. Rahman MA, et al. Advancements in breast cancer detection: global trends, risk factors, imaging modalities, machine learning, and deep learning approaches. BioMedInformatics. 2025;5(3):46.
14. Arravalli T, Chadaga K, Muralikrishna H, et al. Detection of breast cancer using machine learning and explainable artificial intelligence. Sci Rep. 2025;15.
15. Lin Q, Tan WM, Ge JY, et al. Artificial intelligence-based diagnosis of breast cancer by mammography microcalcification. Fundam Res. 2025;5(2):880-889.
16. Yao X, Wang X, Wang SH, Zhang YD. A comprehensive survey on convolutional neural networks in medical image analysis. Multimed Tools Appl. 2020;81(29):41361-41405.
17. Aitazaz T, Tubaishat A, Al-Obeidat F, Shah B, Zia T, Tariq A. Transfer learning for histopathology images: an empirical study. Neural Comput Appl. 2022;35(11):7963-7974.
18. Atrey K, Singh BK, Bodhey NK. Multimodal classification of breast cancer using feature-level fusion of mammogram and ultrasound images. Multimed Tools Appl. 2023;83(7):21347-21368.
19. Habib G, Kiryati N, Sklair-Levy M, et al. Automatic breast lesion classification by joint neural analysis of mammography and ultrasound. In: Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures. Cham: Springer; 2020. p.125-135.
20. Deb D, Dash R, Mohapatra DP. MMHBC-Net: a multimodal hybrid approach for breast cancer classification. Neural Comput Appl. 2025.
21. Nawaz U, Saeed Z, UbaidUllah HM, Mirza F, Muzzamil M. Explainable attention-enhanced approach for multimodal breast cancer diagnosis across diverse imaging modalities. Int J Imaging Syst Technol. 2025;35(6).
22. Pacal I, Attallah O. InceptionNeXt-Transformer: a multi-scale deep feature learning architecture for multimodal breast cancer diagnosis. Biomed Signal Process Control. 2025;110:108116.
23. Atrey K, Singh BK, Bodhey NK. Integration of ultrasound and mammogram for multimodal classification of breast cancer using hybrid residual neural network and machine learning. Image Vis Comput. 2024:104987.
24. Alom MR, et al. An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images. Sci Rep. 2025;15.
25. Al-Tam RM, Al-Hejri AM, Alshamrani SS, Al-Antari MA, Narangale SM. Multimodal breast cancer hybrid explainable computer-aided diagnosis using medical mammograms and ultrasound images. J Appl Biomed. 2024;44(3):731-758.
26. Sahu A, Das P, Meher S. An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound images. Biomed Signal Process Control. 2024;87:105377.
27. Wang YM, et al. CNN-based cross-modality fusion for enhanced breast cancer detection using mammography and ultrasound. Tomography. 2024;10(12):2038-2057.
28. Bouallegue B, El-Latif YMA, El-Sofany H, et al. An enhanced approach for predicting breast cancer using deep learning and explainable AI techniques in an IoT environment. Int J Intell Syst. 2025.
29. Devkota R, et al. Evaluation of breast mass by mammography and ultrasonography with histopathological correlation. J Nepal Health Res Counc. 2021;19(3):487-493.
30. Cong J, Wei B, He Y, Yin Y, Zheng Y. A selective ensemble classification method combining mammography and ultrasound images for breast cancer diagnosis. Comput Math Methods Med. 2017;2017:1-7.
31. Heath M, et al. Current status of the Digital Database for Screening Mammography. In: Computational Imaging and Vision. 1998. p.457-460.
32. Hatoum FT, Tomek RM, Whitney HM, Giger ML. Importance of stratified sampling in the development of medical-imaging AI training and

test sets. Proc SPIE Med Imaging. 2025;13407:210-217.
33. Gadzicki K, Khamsehashari R, Zetzsche C. Early versus late fusion in multimodal convolutional neural networks. In: 2020 IEEE 23rd International Conference on Information Fusion. IEEE; 2020. p.1-6.
34. Zhang Y, Chen H, Wei Y. Deep microscopy adaptation network for histopathology cancer image classification. In: Medical Image Computing and Computer-Assisted Intervention. Springer; 2019.
35. Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M. Domain-adversarial neural networks. arXiv:1412.4446; 2014.
36. Raghavan K, B S, V K. Attention-guided Grad-CAM: an improved explainable artificial intelligence model for breast cancer detection. Multimed Tools Appl. 2024;83:57551-57578.
37. Dwivedi R, Dave D, Naik H, et al. Explainable AI: core ideas, techniques, and solutions. ACM Comput Surv. 2023;55(9):1-33.
38. Angelov PP, Soares EA. Explainable artificial intelligence: an analytical review. WIREs Data Min Knowl Discov. 2021;11(5):e1424.
39. Talukder MA. An improved XAI-based DenseNet model for breast cancer detection using reconstruction and fine-tuning. Results Eng. 2025;26:104802.
40. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. Acad Radiol. 2012;19(2):236-248. doi:10.1016/j.acra.2011.09.014.
41. Suckling J, Parker J, Dance DR, Astley S, Hutt I, Boggis CRM, et al. The Mammographic Image Analysis Society digital mammogram database. Exerpta Medica International Congress Series. 1994;1069:375-378.
42. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data. 2017;4:170177. doi:10.1038/sdata.2017.177.
43. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief. 2020;28:104863. doi:10.1016/j.dib.2019.104863.
44. BUS-UCLM Breast Ultrasound Dataset. Kaggle. https://www.kaggle.com/datasets/orvile/bus-uclm-breast-ultrasound-dataset. Accessed 2026.
45. BUS-UC Breast Ultrasound Dataset. Kaggle. https://www.kaggle.com/datasets/orvile/bus-uc-breast-ultrasound. Accessed 2026.
46. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng. 2016;63(7):1455-1462. doi:10.1109/TBME.2015.2496264.
47. Sajjadnia Z, Khayami R, Moosavi MR. Preprocessing breast cancer data to improve data quality and diagnosis. Cancer Inform. 2020;19:1176935120917955.
48. George MJ, Dhas DAS. Preprocessing filters for mammogram images: a review. In: Conference on Emerging Devices and Smart Systems. IEEE; 2017. p.1-7.
49. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2261-2269. doi:10.1109/CVPR.2017.243.
50. Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision. 2018:839-847. doi:10.1109/WACV.2018.00097.