JADIMI:
JOURNAL OF AI-DRIVEN INFORMATICS AND MANAGEMENT INFORMATION
Vol. 1 No. 1, August 2026, pp 20-33

Controlled Comparison of Multimodal Hybrid Deep Learning for Benign-Malignant Mammogram Classification on CBIS-DDSM

Marshall Rasendria Mahendra[1], Farhan Muamar Fawwaz[2], Naisya Aghis Nabila[3], Muhammad Rakha[4*]

[1] *School of Industrial Engineering, Telkom University, Bandung, Indonesia*
[2] *School of Computing, Telkom University, Bandung, Indonesia*
[3] *School of Electrical Engineering, Telkom University, Bandung, Indonesia*
[4] *Department of Information Systems, Telkom University, Bandung, Indonesia*
**Corresponding author: muhammadrakhamr@telkomuniversity.ac.id*

***Abstract*—Breast cancer is the leading cause of cancer death among women, and mammography remains the first line of screening. Many deep learning studies on CBIS-DDSM report accuracy above 95%, yet they partition the archive at image level, so views of one patient land in both training and test sets and the figures cannot be trusted. This study builds a controlled comparison of two image representations, the whole mammogram and the lesion region of interest, under one identical training recipe. Six models are trained on each representation: EfficientNet-B3, ConvNeXt-Tiny, a gated hybrid fusion of the two with the Convolutional Block Attention Module, two variants that fuse seven clinical descriptors through a perceptron, and a Mammo-CLIP backbone pretrained on mammograms. The official CBIS-DDSM partition is enforced at patient level and yields 2,148 training, 261 validation and 448 test mammograms from 1,104, 122 and 234 disjoint patients, and the BI-RADS assessment is dropped to avoid circular prediction. On the test set the strongest model, Hybrid Fusion+CBAM+Clinical, reaches an AUC of 0.856 at an accuracy of 78.1% on regions of interest, and an AUC of 0.845 at an accuracy of 78.4% on whole images. Clinical metadata is the largest single source of gain, close to six AUC points on both representations. After its BatchNorm adaptation is corrected, Mammo-CLIP recovers to an AUC of 0.789 and becomes the best image-only whole-image backbone. Grad-CAM shows that attention settles on the lesion rather than on artifacts, and the reported range is what remains once leakage is removed.**

Keywords—breast cancer; mammography; transfer learning; clinical metadata fusion; attention mechanism; explainable AI; patient-wise split

Manuscript received 11 Aug. 2026; revised 17 Aug. 2026; accepted 23 Aug. 2026. Date of publication 30 Aug. 2026.

I. INTRODUCTION

Breast cancer is a global health challenge for women and the leading cause of high mortality rates among women. Based on GLOBOCAN estimates, there are approximately 2.3 million new cases globally, with a death toll exceeding 680,000 [1], [2]. To reduce this mortality rate, clinical mammography screening is recognized as the gold standard for detecting microscopic lesions before clinical symptoms appear [3].

However, manual visual analysis by radiologists is prone to subjectivity and is time-consuming, particularly in dense breast tissue, which often obscures the cellular morphology of an anomaly [4]. To address these clinical challenges, Convolutional Neural Network (CNN) architectures are widely relied upon because they have proven robust at automatically and objectively extracting and recognizing pathological features [5].

The use of conventional CNNs in mammogram classification still carries an inherent weakness. The models often assign uniform attention weights to all image areas [6], which leaves them susceptible to distraction by background noise [7]. A second and more damaging problem sits outside the architecture. Many published studies split their datasets randomly by image

instead of by patient [8], so the two mammographic views of one breast, which share the same anatomy and the same acquisition artifacts, can be separated between training and testing.

This practice gives rise to data leakage. The model is rewarded for memorizing the anatomy of a specific patient and produces biased accuracy figures above 95% that do not transfer to the clinic [9]. In response to the limitations of single-architecture models, hybrid multimodal approaches have been introduced to guide the model towards lesion regions while combining the image evidence with the clinical context of the patient [10]. The gain is real, but it is paid for with architectural complexity that deepens the black-box character of the model and conflicts with the clinical requirement that a diagnostic suggestion be justified [11]. Explainable AI (XAI) is therefore needed to bridge mathematical accuracy and medical accountability, and this need for visual transparency points towards classification systems that produce a clinically readable justification alongside a score.

Motivated by this methodological gap, this study conducts a controlled comparison of hybrid deep learning architectures for benign and malignant classification on the CBIS-DDSM dataset under an evaluation protocol that is free of data leakage. The main contributions of this research are the following:

- Conduct a comparative evaluation between whole-image and Region of Interest (ROI) representations using a strict patient-wise split to ensure there is no data leakage.
- Determine the effectiveness of cross-modality fusion (clinical metadata) in a CBAM-based attention hybrid architecture compared to the use of a mammography-specific foundation model (Mammo-CLIP).
- Utilize Grad-CAM++ (Generalized Gradient-based Visual Explanations) to demonstrate that the accuracy of the model is validated by precise lesion localization, which supports transparent decision-making.

The remainder of this article is organized as follows. Section II reviews the related work on automated mammogram classification and locates the gap that this study addresses. Section III describes the materials and methods, covering the dataset, the two representations, the preprocessing chain, the patient-level split, the clinical metadata encoder, the six architectures, the training recipe and the evaluation protocol. Section IV presents and discusses the results, and Section V concludes the study.

II. RELATED WORK

Convolutional Neural Networks (CNNs) have been widely adopted in recent literature for the automated classification of mammogram images [12]. Most of the published work concentrates on optimizing the architecture so that it copes with the visual difficulties that are specific to medical imaging, such as variation in breast tissue density and ambiguous lesion boundaries [13]. The survey below follows four lines of development and closes with the methodological gap that motivates this study.

In the early stage of development, research relied entirely on a single deep learning architecture without incorporating other methods or data modalities. Several studies implemented backbones such as ResNet, VGG or EfficientNet end-to-end on whole mammogram images [7]. A parallel line of work extracts features strictly from Region of Interest (ROI) segments with an isolated single model [14], which removes the background problem at the cost of the surrounding tissue context. Both families share the same ceiling. A single backbone struggles to produce robust diagnostic features when the lesion is small, low-contrast or embedded in dense tissue, and neither family reports what the model actually looked at.

To lift that ceiling, later studies combined more than one feature extractor. Fusing two backbones with complementary inductive biases captures local texture and global shape at the same time [15], and attention modules are commonly inserted so that the fused representation is re-weighted towards suspicious regions instead of being averaged uniformly over the film. A second family of hybrids crosses the modality boundary. Multimodal studies show that fusing the mammogram with the tabular clinical record of the patient improves the predictive metrics well beyond what pixels alone support [16], because descriptors such as breast density or lesion margin carry diagnostic information that is not recoverable from a single view. What these studies rarely report is how much of the reported gain comes from the extra modality and how much comes from the larger model, since the image-only and the multimodal variants are seldom trained under one identical recipe.

As the models grew more complex, Explainable AI (XAI) was introduced to keep the diagnostic reasoning inspectable. Recent studies apply Gradient-weighted Class Activation Mapping (Grad-CAM) to the attention maps of a CNN to show that the network responds to the lesion and not to random background pixels [17]. The visualization is convincing when the underlying evaluation is sound, but a heatmap that points at the correct lesion says nothing about whether the reported accuracy was measured on truly unseen patients. Several authors have warned that saliency evidence is being used as a substitute for validation rather than as a complement to it [18].

Studies that impose strict methodological control report a different reality. When the partition is performed at patient level, the measured performance of mammogram classification drops sharply and empirical

accuracy settles in the range of 75% to 79% [19]. The gap between that range and the 95% to 99% claimed elsewhere is largely explained by leakage, because two views of the same breast placed on opposite sides of the split let the network recognize the patient instead of the pathology. The effect has been quantified outside mammography as well, where an image-level split in brain MRI classification inflated the reported accuracy by a wide margin [8].

Hybrid architectures, clinical data fusion and Grad-CAM have therefore all been explored, but almost always in isolation and rarely under an evaluation protocol that survives scrutiny. No prior study places the two image representations, the six architectures and the two threshold policies inside a single controlled experiment on the official patient-level CBIS-DDSM partition. That is the gap this study fills. Training the image-only and the multimodal variants under one identical recipe is what makes the contribution of each component measurable rather than assumed.

III. MATERIALS AND METHODS

The complete research workflow is presented in Fig. 1. It runs from dataset curation through dual-representation preprocessing and patient-wise splitting to the training of six models and finally to evaluation and interpretation. Each stage is applied to the two representations with the same code path and the same hyperparameters, so that any difference in the final metrics can be attributed to the representation and not to the recipe.

Reading the diagram from the top, the curated archive is first turned into two parallel inputs, the whole mammogram and the lesion crop. The patient-level split is applied once and before any augmentation, so the training, validation and test partitions are fixed and identical in construction for both representations. Augmentation and model fitting act on the training partition alone, the validation partition is used for checkpoint selection and for choosing the decision threshold, and the test partition is read only at the end.

A. Dataset

The dataset used in this study is the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) [20], a re-curated and standardized release of the Digital Database for Screening Mammography. The collection covers two abnormality types, masses and calcifications, each acquired in the Craniocaudal (CC) and the Mediolateral Oblique (MLO) view. Every case is distributed in three forms: the full mammogram, a binary lesion mask and a cropped patch centered on the lesion. The benign or malignant ground truth comes from biopsy rather than from a reader opinion, which makes the label objective. Alongside the pixel data the archive provides per-lesion clinical annotations, namely breast density, subtlety rating, abnormality type, mass shape, mass margins, calcification type and calcification distribution, together with the BI-RADS assessment. Each lesion is explicitly mapped to its source patient, and it is that mapping which makes a leak-free partition possible.

The pathology field of the archive carries three values, and the binarization applied here is explicit about all of them. A record is labeled malignant when its pathology string contains MALIGNANT and benign otherwise, so the BENIGN_WITHOUT_CALLBACK category is folded into the benign class instead of being discarded. Keeping those cases is the conservative choice for a screening task. They are findings that a radiologist saw and judged not to warrant a recall, which makes them the benign examples closest to the decision boundary, and removing them would leave an artificially easy negative class and an optimistic estimate of separability. The same rule is applied to the whole-image and to the crop inventory, so the two representations carry identical label semantics and any difference measured between them comes from the pixels rather than from the labeling.

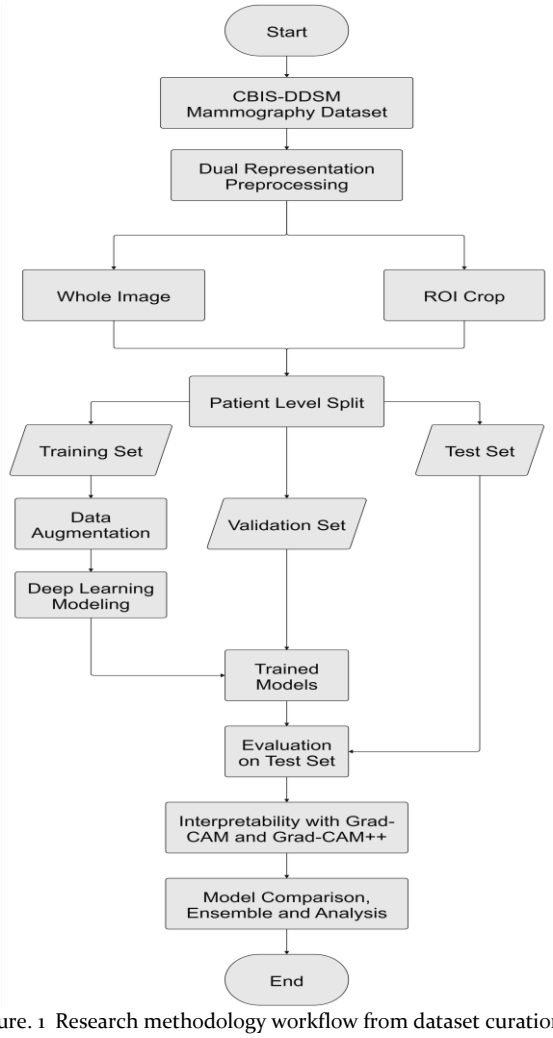

Figure. 1 Research methodology workflow from dataset curation to evaluation and interpretability

Records were assembled by reading the mass and the calcification case description tables, normalizing the column names, and joining them to the image inventory through the patient identifier, the breast side and the view. Only the series marked as full mammogram images were retained for the whole-image branch, so that lesion masks and cropped patches could not enter it by accident. Because a single full mammogram may carry more than one annotated lesion, the clinical annotation of a view is aggregated from the lesion that determines the label, that is the malignant lesion when one is present and the first lesion otherwise. The procedure mapped 2,857 mammograms, all of which received a per-view label, so no per-patient fallback and no unmatched image remained. Fig. 2 shows one benign and one malignant case per view, each as the original scan, the whole-image input and the ROI crop input of the same lesion, so the two representations can be compared directly.

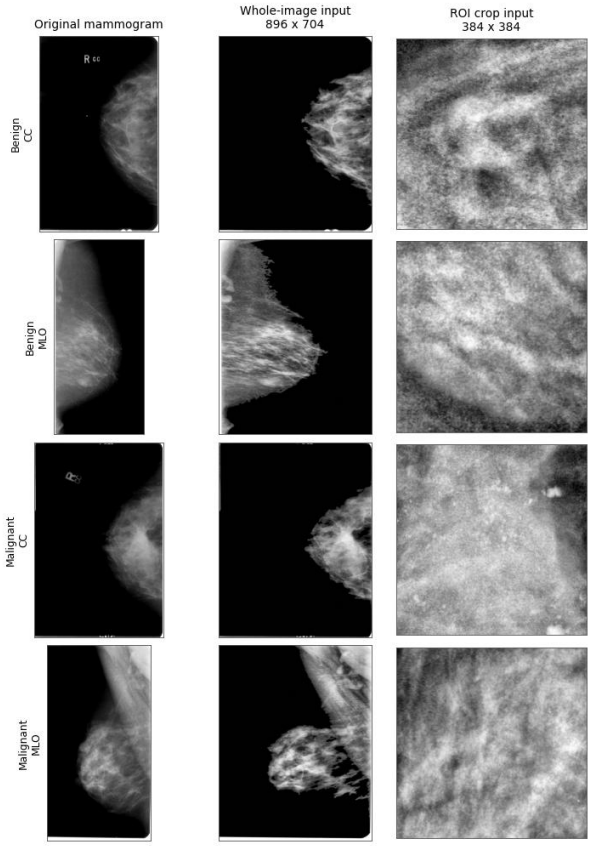

Figure. 2 Test-split samples of the two representations. Each row is one case: the original mammogram, the whole-image network input, and the ROI crop input of the same lesion

B. Image Representation

Two representations are compared side by side. The whole-image representation keeps the full mammogram, so the model sees the lesion together with the surrounding parenchyma, the tissue density and the position of the finding inside the breast, even though the lesion itself occupies only a small fraction of the pixels. The ROI representation takes the opposite route and keeps only the crop around the lesion that was delineated during dataset curation. The target becomes large and unambiguous, and the class balance of the crop set is more even, but the global context disappears and the model can no longer judge a finding relative to the rest of the breast. This tension between detail and context, and its effect on discriminability, is exactly what the controlled comparison is designed to measure. The distinction also matters in deployment, since a whole-image model runs on a raw scan while an ROI model presupposes that the lesion has already been marked.

C. Preprocessing

Whole mammograms pass through a four-step chain. The breast is first isolated: Otsu thresholding separates tissue from the black film background, a morphological opening with a 25 × 25 elliptical kernel deletes scanner labels and adhesive tape, and only the largest connected component is kept and cropped to its bounding box. The step is reverted whenever the surviving region falls below 5% of the original area, which protects the rare scans where the threshold fails.

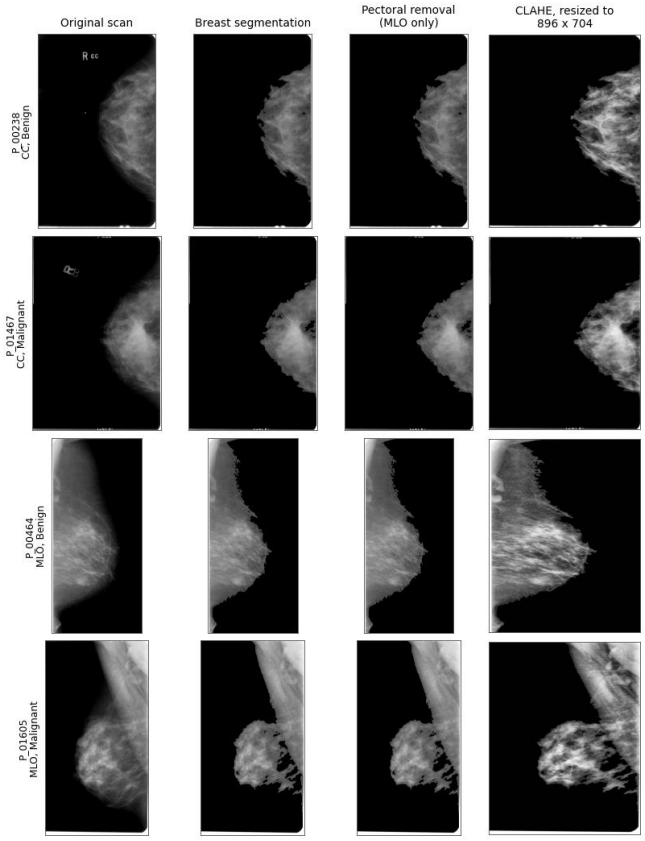

Figure. 3 Whole-image preprocessing chain: original scan, breast segmentation, pectoral muscle removal on the MLO views, and the CLAHE-enhanced network input

On MLO views the pectoral muscle is then removed, because its dense and homogeneous texture is an easy shortcut that carries no diagnostic value [21]. The chest wall side is detected by comparing the intensity mass of

the left and the right image quarter, the image is flipped so that the muscle sits in the top left corner, pixels above the 88th intensity percentile are extracted and opened with a 9 × 9 elliptical kernel, and the convex hull of the component that touches the corner is erased. Two guards protect the breast: a candidate larger than 35% of the frame is not accepted, and a filled hull larger than 40% of the frame is discarded. CC views skip this step entirely.

Contrast is then enhanced with Contrast Limited Adaptive Histogram Equalization (CLAHE) [22] using a clip limit of 2.5 and an 8 × 8 tile grid, which raises local lesion contrast without amplifying global film noise. Finally the grayscale result is replicated to three channels and resized with area interpolation to 985 × 774 pixels, a 10% margin above the network input of 896 × 704 that leaves room for random cropping during training. Lesion crops follow the same chain without the segmentation steps, since a crop already contains only the lesion: CLAHE with identical parameters, channel replication, and resizing to 384 × 384 pixels. Both branches are normalized with the ImageNet channel statistics, and every statistic used at validation or test time is derived from the training partition alone. Fig. 3 traces the four stages on four cases, two CC and two MLO, and shows that the pectoral column changes only the MLO rows.

D. Patient-Level Data Splitting

The partition follows the official CBIS-DDSM split, which is encoded in the patient identifier of every record, and it is enforced at patient level rather than at image level. All images belonging to one patient are constrained to a single subset. A patient who appears in the mass training list and in the calcification test list at the same time is forced entirely into the test set. That direction is the conservative one, because it can only shrink the training pool and can never contaminate the evaluation. Ten percent of the remaining training patients, drawn with a fixed seed, are held out as the validation set, which is used for checkpoint selection and for choosing the decision threshold. The three patient sets are then asserted to be pairwise disjoint, and the assertion reports zero overlap for both representations.

TABLE I
DATASET COMPOSITION AFTER THE PATIENT-LEVEL SPLIT

Subset	Benign	Malignant	Images	Patients
Whole image				
Train	1,164	984	2,148	1,104
Validation	155	106	261	122
Test	245	203	448	234
Total	1,564	1,293	2,857	1,460
ROI crop				
Train	1,511	1,037	2,548	1,096
Validation	124	108	232	121
Test	475	312	787	349
Total	2,110	1,457	3,567	1,566

Table I reports the resulting composition. The whole-image branch contains 2,148 training, 261 validation and 448 test mammograms from 1,104, 122 and 234 patients. The same rule applied to the lesion crops gives 2,548 training, 232 validation and 787 test regions of interest from 1,096, 121 and 349 patients. The crop inventory is the larger of the two, 3,567 items against 2,857, because one mammogram can carry more than one annotated lesion and a few patients enter the crop inventory without a matching full mammogram. The malignant share runs from 40.6% to 45.8% across the whole-image subsets and from 39.6% to 46.6% across the crop subsets, so no subset is markedly easier or harder than the one the models were fitted on.

E. Clinical Metadata Encoding

The variants that use clinical data fuse seven descriptive attributes: breast density, subtlety, abnormality type, mass shape, mass margins, calcification type and calcification distribution. Breast density and subtlety are ordinal and are standardized by z-score, with an explicit indicator channel that flags a missing value instead of silently imputing a mean. The remaining five attributes are categorical and are one-hot encoded over a vocabulary that reserves a dedicated bucket for values that are missing or that were never observed during training. The vocabulary and the ordinal statistics are fitted on the training partition only, so no information from validation or test records reaches the encoder.

Coverage of the seven attributes on the whole-image training split is uneven by construction. Breast density, subtlety and abnormality type are present on every record, while mass shape and mass margins are filled on 52.1% and 50.6% of the records and calcification type and distribution on 47.4% and 40.9%. The two groups are close to complementary because a mass lesion has no calcification descriptor and a calcification has no mass descriptor. This is why the missing bucket is an explicit vocabulary entry rather than an imputed value: the absence of a mass margin is itself evidence about the lesion type and the model is allowed to use it.

The resulting sparse vector is projected by a small perceptron of two linear layers with batch normalization, GELU activation and dropout of 0.3 into a 128-dimensional embedding, which is concatenated with the pooled image feature before the classifier. When MixUp is active the metadata vector is interpolated with the same coefficient as the image and the label, so the three stay consistent. The BI-RADS assessment and the pathology field are deliberately excluded from the feature set. The first is the suspicion verdict of the radiologist and would make the prediction circular, and the second is the label itself.

F. Model Architectures

Six models are trained with identical configurations so that the comparison stays fair. The image-only baselines are ConvNeXt-Tiny [23], which produces a 768-dimensional feature map, and EfficientNet-B3 [24], which produces a 1536-dimensional map. The Hybrid Fusion+CBAM model integrates both. Each branch is reduced to 512 channels by a 1 × 1 convolution and refined by a Convolutional Block Attention Module (CBAM) [6], whose channel branch pools the map by average and by maximum and passes both through a shared bottleneck with reduction 8, and whose spatial branch concatenates the channel-wise mean and maximum and convolves them with a 7 × 7 kernel. The refined maps are globally averaged and merged by a gate that is learned from their concatenation, so that the contribution of each backbone is decided per sample rather than fixed in advance, following the gated multimodal fusion formulation of Arevalo et al. [25], as expressed in (1) and (2).

$$g = softmax\left(W_g[f_e; f_c]\right) \tag{1}$$

$$f_{fus} = [g_1 f_e; g_2 f_c] \tag{2}$$

Here f_e and f_c are the EfficientNet and the ConvNeXt features and g is the gate. The fused vector of 1024 dimensions is classified by a two-layer head with dropout of 0.4 and GELU activation. Because gradients that reach a deep backbone through a single fused head tend to be diluted, an auxiliary linear classifier is attached to each branch following the deeply supervised network scheme [26] and its loss is added with a fixed weight α of 0.4, which keeps both backbones learning discriminative features of their own, as expressed in (3). The full layout, including the clinical branch, is drawn in Fig. 4.

$$\mathcal{L} = \mathcal{L}_{main} + \alpha\left(\mathcal{L}_{aux}^{eff} + \mathcal{L}_{aux}^{cnx}\right) \tag{3}$$

The sixth model replaces the ImageNet initialization with Mammo-CLIP [27], a vision language foundation model whose EfficientNet-B5 image encoder was pretrained on mammograms. The encoder weights are loaded from the public checkpoint, the text branch is discarded, and the feature map is reduced by generalized-mean pooling [28] with a learnable exponent p initialized at 3, which interpolates between average and maximum pooling and suits the small bright structures of a mammogram better than plain averaging, as expressed in (4).

$$f^{(k)} = \left(\frac{1}{|X|}\sum_{x \in X} x^p\right)^{1/p} \tag{4}$$

G. Training Strategy and Configuration

Training runs in two transfer learning phases. In the first phase the backbone is frozen and only the newly initialized modules are optimized for three epochs at a learning rate of 1×10^{-3}, which prevents the random head from destroying the pretrained filters. In the second phase the whole network is unfrozen for at most 25 epochs with discriminative rates, 3×10^{-4} for the new modules and 5×10^{-5} for the backbone. Inside each backbone the rate is scaled block by block, at 0.25 for the first third of the blocks, 0.5 for the middle third and the full rate for the last third, so that the generic early filters move far less than the task-specific late ones. Both phases use AdamW [29] with a weight decay of 1×10^{-2} under the cosine annealing schedule of Loshchilov and Hutter [30], preceded by a linear warmup over the first 10% of the steps, as expressed in (5).

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + cos\frac{\pi t}{T}\right) \tag{5}$$

Regularization is applied in several layers because the training set is small for a network of this size. Label smoothing of 0.1 [31] softens the hard targets, MixUp with α of 0.2 [32] is applied to half of the batches and interpolates image, metadata and label together, and dropout of 0.4 together with weight decay limits the effective capacity. Class imbalance is handled by weights inversely proportional to class frequency inside a soft cross-entropy loss that accepts the mixed soft targets, as expressed in (6).

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c\, \tilde{y}_{n,c} \log p_{n,c} \tag{6}$$

The resolution of the whole-image branch forces a physical batch of 2, which is restored to an effective batch of 32 by accumulating gradients over 16 steps. Gradient checkpointing keeps the activations within memory and the BatchNorm statistics are frozen, since batch statistics estimated from two samples are unusable. The ROI branch runs a physical batch of 16 with 2 accumulation steps and therefore needs neither checkpointing nor frozen normalization. Every other hyperparameter is shared. Gradients are clipped to a norm of 1.0 and the whole run uses automatic mixed precision.

One adjustment applies to Mammo-CLIP alone. Its BatchNorm statistics are kept adaptive instead of frozen, because the freezing that small batches normally require prevents the pretraining statistics from adapting to the distribution of CLAHE-processed input [33]. In the first run of this study that single flag was the difference between a near-random model and a competitive one, which is reported in Section IV.

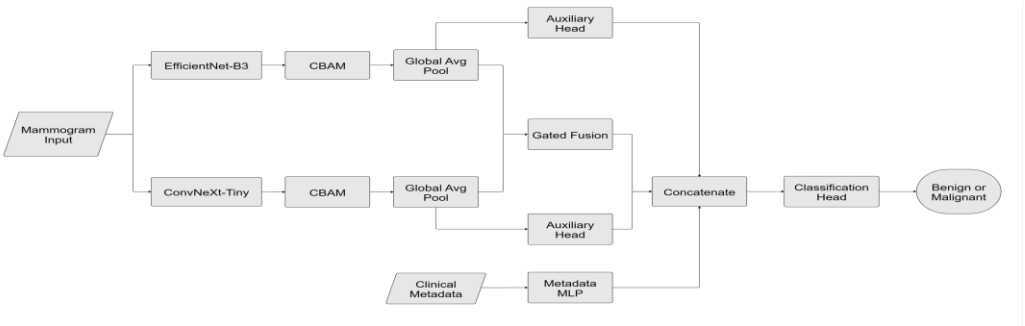

Figure. 4 Hybrid Fusion+CBAM+Clinical architecture with gated fusion and deep supervision

An exponential moving average of the weights with a decay of 0.9997 is maintained through the second phase. At the end of training the best-validation-AUC checkpoint and the averaged weights are both scored on the validation set and the better of the two is kept, and training stops early when the validation AUC has not improved for six epochs. The random seed is fixed at 42 for every model. The complete configuration, identical across all six architectures except where the table states otherwise, is listed in Table II.

TABLE II
TRAINING CONFIGURATION SHARED BY ALL SIX MODELS

Parameter	Value
Weight initialization	ImageNet for EfficientNet-B3 and ConvNeXt-Tiny; Mammo-CLIP checkpoint for EfficientNet-B5
Network input	896 × 704 (whole image); 384 × 384 (ROI)
Physical batch × accumulation	2 × 16 (whole image); 16 × 2 (ROI)
Effective batch size	32
Optimizer	AdamW, weight decay 1e-2
Learning-rate schedule	Cosine annealing, linear warmup over 10% of steps, floor 0.02
Learning rate	Head 1e-3; fine-tune head 3e-4; backbone 5e-5
Layer-wise rate multiplier	0.25 / 0.50 / 1.00 over the three backbone thirds
Epochs	3 head + 25 fine-tune, early stopping patience 6 on validation AUC
Gradient clipping	Max norm 1.0
Precision and memory	AMP; gradient checkpointing and frozen BatchNorm on whole image only
Augmentation	RandomResizedCrop scale 0.85–1.00, horizontal flip, rotation ±12°, brightness and contrast jitter 0.12
Regularization	Label smoothing 0.1, MixUp α = 0.2 applied with probability 0.5, dropout 0.4
Class weights	Inverse frequency inside a soft cross-entropy
Deep supervision	Auxiliary branch weight α = 0.4
Weight averaging	EMA decay 0.9997; best checkpoint and EMA compared on validation
Inference	Horizontal-flip TTA; soft-voting ensemble over the image-only models
Random seed	42

H. Evaluation Protocol and Explainability

Performance is measured with accuracy, precision, recall, F1 score and AUC. The first four are computed from the confusion matrix following the standard definitions of Sokolova and Lapalme [34], as expressed in (7) - (10).

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (7)$$

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (8)$$

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (9)$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \times 100\% \quad (10)$$

AUC is treated as the primary metric because it does not depend on a particular threshold and therefore measures the discriminative power of the model rather than the calibration of one operating point. The threshold itself is always selected on the validation set and never on the test set. The first operating point maximizes the Youden index [35], that is the difference between the true positive rate and the false positive rate, as expressed in (11).

$$t^* = \arg\max_t (\mathrm{TPR}(t) - \mathrm{FPR}(t))) \quad (11)$$

The second operating point is clinical. It is the lowest threshold that still keeps validation sensitivity at or above 90%, on the argument that missing a malignant lesion costs far more than an unnecessary recall. Both points are reported so that the trade-off is visible instead of hidden inside a single number [36]. Test-time augmentation averages the prediction of the image and its horizontal mirror, and the three image-only models are additionally combined by soft-voting over their probability outputs.

Because both test partitions are of moderate size, every AUC is reported together with a 95% confidence interval rather than as a point estimate alone. The interval is obtained from the Hanley–McNeil estimator [37], which derives the standard error of the statistic from the number of malignant and of benign cases the partition actually contains, so the precision of each estimate is bounded by the evidence available to measure it. These intervals are marginal rather than paired. They do not exploit the fact that every model is scored on the same cases, and they are therefore wider than the interval that a paired procedure such as the DeLong test [38] would place on the difference between two models. They are used here to bound the precision of each estimate and to decide which of the observed gaps are large enough to carry an interpretation, not to certify one architecture as superior to another.

For interpretation, Grad-CAM [39] and Grad-CAM++ [40] produce heatmaps of the regions that drive the prediction. The class-discriminative weight of each feature channel is obtained by backpropagating the target logit to the last convolutional map, and the weighted sum is passed through a rectifier, as expressed in (12).

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}, L_{\text{CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \quad (12)$$

The target layer is the last feature block for the single-backbone models, the EfficientNet CBAM output for the hybrid models, and the head convolution for Mammo-CLIP. For the variants that consume clinical data the metadata vector is set to zero during the backward pass, so that the resulting saliency reflects the image evidence alone and is not contaminated by the tabular branch.

IV. RESULTS AND DISCUSSION

A. Discriminative Performance

Table III and Table IV report the test performance at the Youden threshold for the ROI and the whole-image representation. Every number is measured on patients that no model has seen during training or threshold selection, and the row in bold is the one with the highest AUC in that table.

Each row is one architecture and the columns report accuracy, precision, recall, F1 and AUC in that order. The two tables carry the same eight rows, so one model can be followed across the representations by reading the same line in both. The first four columns are threshold-dependent and, therefore, move together whenever the operating point moves, whereas AUC does not, which is why AUC carries the comparison, whereas the remaining columns are read as the description of one particular point on the curve.

TABLE III
TEST PERFORMANCE ON THE ROI REPRESENTATION, YOUDEN THRESHOLD

Model	Acc	Prec	Rec	F1	AUC
EffNet-B3	71.66%	69.10%	51.60%	59.08%	0.778
ConvNeXt-T	72.43%	67.79%	58.01%	62.52%	0.795
Hybrid+CBAM	73.57%	76.53%	48.08%	59.06%	0.797
EffNet-B3+Clin	76.11%	69.25%	71.47%	70.35%	0.828
Hybrid+CBAM+Clin	**78.14%**	**74.31%**	**68.59%**	**71.33%**	**0.856**
Mammo-CLIP	73.82%	67.21%	66.35%	66.77%	0.802
Ensemble-SV	72.30%	65.99%	62.18%	64.03%	0.802
Ensemble+Clin	76.62%	71.05%	69.23%	70.13%	0.836

TABLE IV
TEST PERFORMANCE ON THE WHOLE-IMAGE REPRESENTATION, YOUDEN THRESHOLD

Model	Acc	Prec	Rec	F1	AUC
EffNet-B3	70.76%	69.78%	62.56%	65.97%	0.784
ConvNeXt-T	67.63%	63.43%	67.49%	65.39%	0.743
Hybrid+CBAM	72.99%	78.08%	56.16%	65.33%	0.785
EffNet-B3+Clin	70.98%	71.35%	60.10%	65.24%	0.817
Hybrid+CBAM+Clin	78.35%	73.04%	82.76%	77.60%	0.845
Mammo-CLIP	71.43%	68.12%	69.46%	68.78%	0.789
Ensemble-SV	68.97%	64.95%	68.47%	66.67%	0.788
Ensemble+Clin	**74.11%**	**66.93%**	**84.73%**	**74.78%**	**0.848**

The clearest and most consistent pattern across both representations is that adding clinical metadata produces the largest gain in AUC of any intervention tested here. On the ROI representation the AUC rises from 0.797 for Hybrid Fusion+CBAM to 0.856 once the metadata branch is attached, and on whole images it rises from 0.785 to 0.845. The improvement of roughly six points appears on both representations and on both backbone families. The comparison is meaningful precisely because the image-only and the multimodal variants differ in nothing but the metadata branch.

The split between precision and recall is where the models differ most, even where their accuracy is close. On the ROI, Hybrid Fusion+CBAM records the highest precision of the image-only group at 76.5% but the lowest recall at 48.1%, so at the Youden point it misses more than half of the malignant crops. Attaching the metadata branch moves the same architecture to 74.3% precision and 68.6% recall, and lifts the F1 score from 59.1% to 71.3%. The pattern repeats on whole images, where recall rises from 56.2% to 82.8% and F1 from 65.3% to 77.6%. Since the AUC rises with them, the clinical descriptors are not merely moving the decision threshold along a fixed curve.

The highest AUC on both representations is recorded by Hybrid Fusion+CBAM+Clinical, which reaches 0.856 at an accuracy of 78.1% on the ROI and 0.845 at an accuracy of 78.4% on the whole image, although its margin over the other multimodal configurations falls inside the confidence intervals reported below. The

distance between the two representations is small. The ROI is marginally ahead because the lesion dominates the frame without background interference, and the whole image follows closely because the surrounding tissue supplies context that the crop discards. The practical conclusion is that the choice of representation matters far less than is usually assumed, and that the effort is better spent on the modality than on the crop. These values also sit in a realistic band for CBIS-DDSM once leakage is prevented, which refutes the 98% to 99% claims that follow from image-level splitting and confirms recent empirical evaluations of what CNNs actually achieve in a clinical setting [41],[42].

The two image-only backbones do not agree on which representation suits them. ConvNeXt-Tiny leads EfficientNet-B3 on the ROI, at an AUC of 0.795 against 0.778, and trails it on whole images, at 0.743 against 0.784. The gated hybrid matches or beats the better of the two on both, at 0.797 and 0.785, which is the behavior the gate is meant to produce, since it can weight the branch that suits the sample instead of committing to one backbone in advance. Soft voting over the three image-only models adds little on top, reaching 0.802 on the ROI and 0.788 on whole images, which is consistent with three members that were fitted on the same partition and therefore carry much of the same error.

Table V places those margins in context by reporting every AUC with its 95% confidence interval. The intervals extend between 0.029 and 0.046 on either side of the estimate, which is wider than most of the gaps that separate the leading configurations. The three highest values obtained in this study, 0.856 for Hybrid Fusion+CBAM+Clinical on the ROI, 0.848 for Ensemble+Clinical on whole images and 0.845 for Hybrid Fusion+CBAM+Clinical on whole images, carry intervals that overlap over almost their entire length, so the ordering among them is not statistically meaningful and none of the three can be declared the better model on this test set. The same caution applies inside the image-only group, where ConvNeXt-Tiny, Hybrid Fusion+CBAM, Mammo-CLIP and the soft-voting ensemble occupy a band from 0.795 to 0.802 on the ROI that a single one of these intervals covers in full.

What does survive the interval is the contribution of the clinical modality. On the ROI the interval of Hybrid Fusion+CBAM ends at 0.830 and the interval of its multimodal counterpart begins at 0.827, so the two barely meet, and on both representations every image-only variant records a point estimate below the lower bound of the best multimodal configuration. The claims of this study are restricted accordingly. The gain attributable to clinical metadata is large relative to the precision of the experiment and is therefore reported as a finding, whereas the ranking among the top three configurations is not and is reported as a tie. Establishing the metadata effect formally would require a paired test on the per-case scores, which is stated as a limitation in Section V. A useful way to read the intervals is to note what the test set would have to look like for them to tighten: halving the width of the interval on the ROI would require roughly four times as many test cases, which is far beyond what the official partition of CBIS-DDSM makes available. The precision reported here is therefore a property of the benchmark rather than of the training procedure, and it is the reason this study frames its conclusion around the size of the metadata effect rather than around the identity of the single best architecture.

TABLE V
AUC WITH 95% CONFIDENCE INTERVALS ON BOTH REPRESENTATIONS

Model	ROI AUC [95% CI]	Whole image AUC [95% CI]
EffNet-B3	0.778 [0.744–0.812]	0.784 [0.741–0.827]
ConvNeXt-T	0.795 [0.762–0.828]	0.743 [0.697–0.789]
Hybrid+CBAM	0.797 [0.764–0.830]	0.785 [0.742–0.828]
EffNet-B3+Clin	0.828 [0.797–0.859]	0.817 [0.777–0.857]
Hybrid+CBAM+Clin	0.856 [0.827–0.885]	0.845 [0.807–0.883]
Mammo-CLIP	0.802 [0.769–0.835]	0.789 [0.746–0.832]
Ensemble-SV	0.802 [0.769–0.835]	0.788 [0.745–0.831]
Ensemble+Clin	0.836 [0.806–0.866]	0.848 [0.811–0.885]

The Receiver Operating Characteristic curves in Fig. 5 for the ROI and Fig. 6 for the whole image show the same hierarchy across the whole operating range and not only at one threshold.

The curve of Hybrid Fusion+CBAM+Clinical dominates the others, and the margin is widest in the low false-positive region that matters most for screening. The soft-voting ensemble that includes the clinical models reaches an AUC of 0.848 on whole images, which is on par with the 0.8483 reported by an end-to-end two-view EfficientNet on the same official split [7], so the protocol used here reproduces a credible external benchmark rather than an inflated one.

Mammo-CLIP deserves a separate note because its behavior changed completely after a single correction. With its BatchNorm statistics frozen, as the small whole-image batch normally demands, the backbone was unable to adapt its pretraining statistics to CLAHE-processed input and collapsed to a near-random AUC of 0.52. Keeping the statistics adaptive lifted it to 0.789 and made it the strongest purely image-based backbone for whole mammograms. On the ROI, where the batch is large enough that the question never arises, Mammo-CLIP reached an AUC of 0.802 and beat the ImageNet-initialized baselines. The two results together support the argument that mammography-specific pretraining is worth its cost, on the condition that its normalization layers are allowed to follow the domain-specific preprocessing.

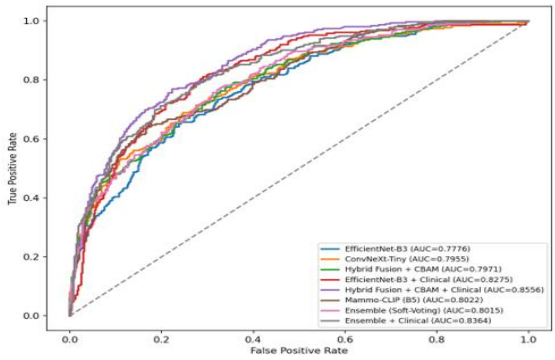

Figure. 5 ROC curves of the test data for the ROI representation.

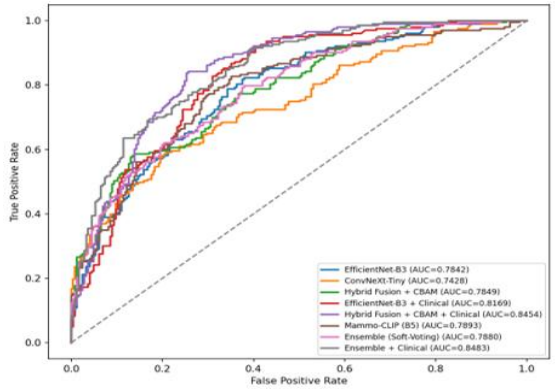

Figure. 6 ROC curves of the test data for the whole-image representation.

B. Clinical Operating Points

Table VI and Table VII report the clinical operating points, obtained by lowering the threshold until validation sensitivity reaches at least 90%.

TABLE VI
HIGH-SENSITIVITY OPERATING POINTS (VALIDATION SENSITIVITY ≥ 90%), ROI

Model	Acc	Prec	Rec	F1	AUC
EffNet-B3	54.76%	46.54%	94.87%	62.45%	0.778
ConvNeXt-T	58.70%	48.91%	93.27%	64.17%	0.795
Hybrid+CBAM	58.83%	48.99%	93.27%	64.24%	0.797
EffNet-B3+Clin	70.78%	58.91%	86.86%	70.21%	0.828
Hybrid+CBAM+Clin	**73.06%**	**61.26%**	**87.18%**	**71.96%**	**0.856**
Mammo-CLIP	60.86%	50.35%	93.27%	65.39%	0.802
Ensemble-SV	60.48%	50.09%	93.27%	65.17%	0.802
Ensemble+Clin	65.44%	53.73%	92.31%	67.92%	0.836

TABLE VII
HIGH-SENSITIVITY OPERATING POINTS (VALIDATION SENSITIVITY ≥ 90%), WHOLE IMAGE

Model	Acc	Prec	Rec	F1	AUC
EffNet-B3	56.25%	50.91%	96.55%	66.67%	0.784
ConvNeXt-T	56.47%	51.09%	92.12%	65.73%	0.743
Hybrid+CBAM	60.49%	53.65%	94.09%	68.34%	0.785
EffNet-B3+Clin	72.54%	63.33%	93.60%	75.55%	0.817
Hybrid+CBAM+Clin	74.33%	65.38%	92.12%	76.48%	0.845
Mammo-CLIP	58.26%	52.16%	95.07%	67.36%	0.789
Ensemble-SV	60.49%	53.59%	95.57%	68.67%	0.788
Ensemble+Clin	**64.73%**	**56.41%**	**97.54%**	**71.48%**	**0.848**

At these thresholds recall on the test set rises consistently while precision falls, which is the expected trade-off. The best whole-image model holds a recall of 92.1% at an accuracy of 74.3%, and the best ROI model a recall of 87.2% at an accuracy of 73.1%. The whole-image Ensemble+Clinical passes 97.5% recall, although its precision drops to 56.4%. The Youden index remains the right summary when the question is overall balance, but for a triage tool that must not miss a malignant lesion the high-sensitivity point is the one to deploy, and reporting both makes the cost of that choice explicit.

Expressed in cases rather than in percentages the choice becomes concrete. The whole-image test set holds 245 benign and 203 malignant mammograms. At the Youden threshold the best model returns 168 true positives, 35 false negatives, 62 false positives and 183 true negatives. At the high-sensitivity threshold it returns 187 true positives and 16 false negatives, so nineteen fewer malignant lesions are missed, while the false positives rise from 62 to 99. The ROI test set holds 475 benign and 312 malignant crops, where the same shift carries the model from 214 to 272 true positives and from 98 to 40 false negatives against a rise from 74 to 172 false positives. A service adopting the high-sensitivity point therefore pays about two additional recalls for every additional malignancy it finds, on either representation.

C. Model Attention and Interpretability

Fig. 7 and Fig. 8 show the Grad-CAM and Grad-CAM++ maps of the best model on each representation, arranged so that a correct benign case, a false positive and a correct malignant case can be compared directly.

Both methods produce similar maps, with Grad-CAM++ drawing slightly sharper lesion boundaries. On correctly classified cases the activation falls on the lesion itself. A benign mass with a well-defined margin triggers a compact and centered response, while a malignant lesion with a spiculated pattern spreads the activation outward along the tissue contours, which is consistent with the morphological criterion a radiologist would apply. On whole images the heatmap concentrates inside the breast and ignores the film background and the pectoral muscle, which is direct evidence that the segmentation and muscle-removal steps of the preprocessing chain did their job. The difference between the two representations is visible in the maps as well. On the ROI the activation is compact because the crop leaves the network little else to attend to, whereas on the whole mammogram the same model must first locate the finding inside a far larger field and the response is correspondingly broader before it settles on the lesion. Neither representation places activation on the film edge, the scanner label or the adhesive markers, which are precisely the shortcut features that an unsegmented input would leave available to the network.

Read together with the fact that the metadata vector is zeroed during the backward pass, the maps describe the image branch on its own, and they indicate that the gain measured in Section IV.A is not produced by a network attending to acquisition artifacts that happen to correlate with the label.

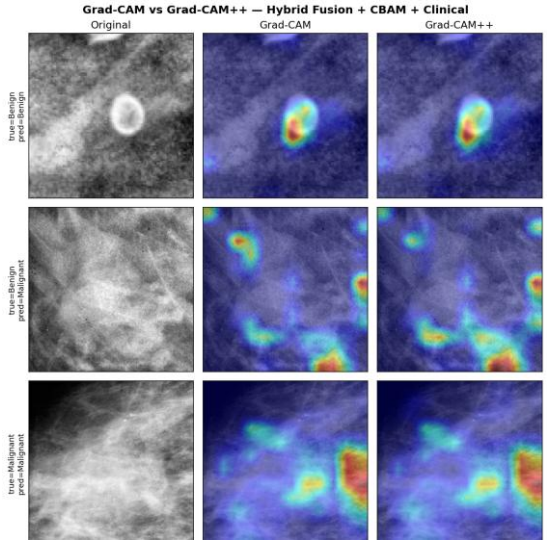

Figure. 7 Grad-CAM and Grad-CAM++ maps of the best model on the ROI. Top row: true benign; middle row: false positive; bottom row: true malignant.

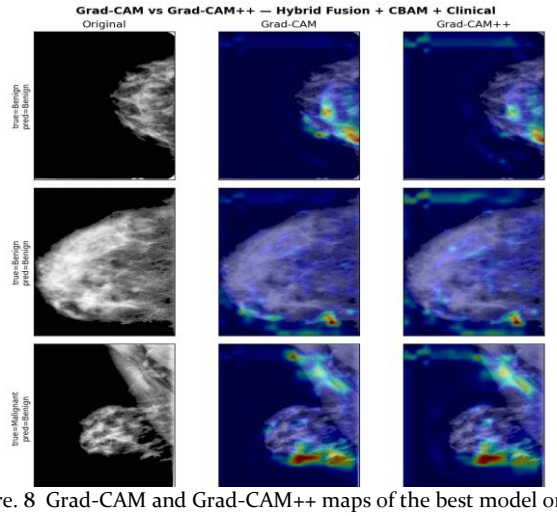

Figure. 8 Grad-CAM and Grad-CAM++ maps of the best model on the whole image, with the same row arrangement.

The misclassified cases are just as informative. A false positive on the ROI arises where dense tissue shows bright foci that mimic a genuine lesion and pull the activation towards them. A false negative on the whole image produces a diffuse map with no clear center, which is the signature of a low-contrast finding that the network never localized in the first place. Both patterns indicate that the decisions rest on clinically relevant visual evidence rather than on shortcut features in the background.

V. CONCLUSION

This study compared whole-image and ROI representations for benign and malignant classification on CBIS-DDSM using six identically configured models and the official patient-level split. Three results stand out. Clinical metadata is the largest single lever, and its effect is stable at roughly six AUC points on both representations. The difference between the two representations is small enough that the choice is not the critical design decision it is often taken to be. Mammo-CLIP, once its BatchNorm adaptation was corrected, recovered from a near-random result to become the best image-only backbone for whole mammograms. The strongest configuration, Hybrid Fusion+CBAM+Clinical, reached an AUC of 0.856 on the ROI and 0.845 on the whole image, and the whole-image ensemble with clinical metadata reached 0.848, which is in line with credible published benchmarks on the same split. Those three values lie inside one another's 95% confidence intervals and are therefore reported as a tie rather than as a ranking.

One caveat belongs with those numbers. The clinical descriptors are unevenly populated, with mass shape and mass margins present on 52.1% and 50.6% of the training records and calcification type and distribution on 47.4% and 40.9%, so part of the metadata gain rests on the informative absence of a descriptor rather than on its value. A site that records all seven attributes for every case would not reproduce it unchanged, and the gain reported here should be read as the value of the descriptors as this archive actually stores them.

It is necessary to specifically address other methodological shortcomings in addition to the unequal distribution of descriptors. First, a single fixed random seed was used to train each model. Future research should assess performance stability across several seeds because deep networks of this scale may show non-trivial run-to-run variance due to the dataset's relatively limited size (e.g., only 2,148 training samples for the whole-image branch). Second, the Mammo-CLIP model's near-total failure until its adaptive BatchNorm statistics were corrected emphasizes how sensitive foundation models are to particular implementation specifics. Despite the universal training recipe used in this work, this discovery raises the prospect that the other examined models might be equally sensitive to architectural subtleties or unexplained hyperparameter selections. Third, the confidence intervals reported in Table V are marginal intervals derived from the Hanley–McNeil estimator, and they are conservative for comparisons between models that were scored on the same cases. A paired procedure such as the DeLong test, or a bootstrap over the paired per-case scores, would decide the remaining comparisons more sharply than the present analysis can, and is the natural next step. The evidence assembled here

supports the size of the metadata effect but stops short of a formal test of it.

The numbers reported here are lower than much of the CBIS-DDSM literature, and that is the point. They are what remains once patient-level separation is enforced, and they are therefore usable as a reference. Three directions follow naturally. Fusing the CC and MLO views of the same breast would add the context that a single view cannot supply. Probability calibration [43] would make the output readable as a risk rather than as a score. External validation on an independent archive such as INbreast [44] would test whether the conclusions survive a change of acquisition hardware. The intended endpoint remains a web-based decision-support tool that returns a calibrated probability together with the Grad-CAM map that justifies it, so that a clinician can inspect the evidence rather than accept a number.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest, financial or non-financial, related to this research.

AUTHOR CONTRIBUTIONS

Marshall Rasendria Mahendra designed the experimental protocol and ran the training campaign, covering the patient-level split, the preprocessing chain, the clinical metadata fusion and the final architecture and configuration, and wrote the initial draft. Farhan Muamar Fawwaz conducted the literature review, framed the research gap and revised the manuscript. Naisya Aghis Nabila prepared the Grad-CAM and Grad-CAM++ visualizations, compiled the evaluation tables at both operating points and finalized the formatting. Muhammad Rakha supervised the study, validated the methodological design and the evaluation protocol, and reviewed and approved the final manuscript. All authors have read and agreed to the submitted version.

ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DATA AND CODE AVAILABILITY

The CBIS-DDSM archive analysed in this study is openly available from The Cancer Imaging Archive at https://www.cancerimagingarchive.net/collection/cbis-ddsm/ and no new patient data were collected for this work. The complete source code is openly available at https://github.com/mxslr/cbis-ddsm-controlled-comparison, covering the curation and joining of the mass and calcification case description tables, the two preprocessing chains, the patient-level splitting routine, the clinical metadata encoder, the definitions of the six architectures, the shared training recipe, and the evaluation, threshold-selection and Grad-CAM scripts.

The repository also releases the patient identifier lists of the training, validation and test partitions, so the exact partition used here can be rebuilt from the public archive without redistributing any image data, and the confidence intervals reported in Table V can be recomputed from the class counts of those lists. The trained checkpoints are available from the corresponding author on reasonable request, as their size exceeds the limits of the repository.

The repository is organised as two notebooks, one for each representation, that run end to end from the raw archive to the tables and figures reported here, together with the environment specification needed to execute them. Each notebook fixes the same random seed, writes the per-epoch validation metrics to disk, and stores the selected decision thresholds alongside the checkpoint that produced them, so that the values in Table III to Table VII can be traced back to the run that generated them rather than taken on trust. The preprocessing chain is exposed as a separate module so that the segmentation, the pectoral-muscle removal and the CLAHE step can be inspected or replaced independently of the training code, this being the part of the pipeline most likely to differ between implementations and therefore the part whose release is most useful.

REFERENCES

[1] H. Sung et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” CA. Cancer J. Clin., vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/CAAC.21660.

[2] F. Bray et al., “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA. Cancer J. Clin., vol. 74, no. 3, pp. 229–263, May 2024, doi: 10.3322/CAAC.21834.

[3] A. Saber, M. Sakr, O. M. Abo-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.

[4] Y. Shen et al., “An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization,” Med. Image Anal., vol. 68, p. 101908, Feb. 2021, doi: 10.1016/J.MEDIA.2020.101908.

[5] X. Wang et al., “Intelligent Hybrid Deep Learning Model for Breast Cancer Detection,” Electronics, vol. 11, no. 17, p. 2767, Sep. 2022, doi: 10.3390/ELECTRONICS11172767.

[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[7] 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.

[8] E. Yagis et al., “Effect of data leakage in brain MRI classification using 2D convolutional neural networks,” Sci. Reports, vol. 11, no. 1, p. 22544, Nov. 2021, doi: 10.1038/s41598-021-01681-w.

[9] M. A. Karagöz et al., "Deep learning-based breast cancer diagnosis with multiview of mammography screening to reduce false positive recall rate," Turkish J. Electr. Eng. Comput. Sci., vol. 32, no. 3, pp. 382–402, May 2024, doi: 10.55730/1300-0632.4076.

[10] S. Bai, S. Nasir, R. A. Khan, A. Meyer, and H. Konik, "Breast Cancer Diagnosis: A Comprehensive Exploration of Explainable Artificial Intelligence (XAI) Techniques," IEEE Access, vol. 13, pp. 205197–205223, Sep 2025, doi: 10.1109/ACCESS.2025.3639184.

[11] H. Zhao et al., "An end-to-end interpretable machine-learning-based framework for early-stage diagnosis of gallbladder cancer using multi-modality medical data," BMC Cancer, vol. 25, no. 1, p. 1178, Jul 2025, doi: 10.1186/S12885-025-14462-9.

[12] S. ur Rehman et al., "BRMI-Net: Deep Learning Features and Flower Pollination-Controlled Regula Falsi-Based Feature Selection Framework for Breast Cancer Recognition in Mammography Images," Diagnostics, vol. 13, no. 9, p. 1618, May 2023, doi: 10.3390/DIAGNOSTICS13091618.

[13] M. A. Aldawsari, S. J. Aldosari, A. Ismail, and M. M. Emam, "A deep learning framework for breast cancer diagnosis using Swin Transformer and Dual-Attention Multi-scale Fusion Network," Sci. Reports, vol. 16, no. 1, p. 8941, Mar. 2026, doi: 10.1038/s41598-026-37969-y.

[14] O. Camurdan et al., "Annotation-efficient, patch-based, explainable deep learning using curriculum method for breast cancer detection in screening mammography," Insights into Imaging, vol. 16, no. 1, p. 60, Mar. 2025, doi: 10.1186/S13244-025-01922-W.

[15] M. M. Altaf, "A hybrid deep learning model for breast cancer diagnosis based on transfer learning and pulse-coupled neural networks," MBE, vol. 18, no. 5, pp. 5029–5046, 2021, doi: 10.3934/mbe.2021256.

[16] C. Ben Rabah, A. Sattar, A. Ibrahim, and A. Serag, "A Multimodal Deep Learning Model for the Classification of Breast Cancer Subtypes," Diagnostics, vol. 15, no. 8, p. 995, Apr. 2025, doi: 10.3390/DIAGNOSTICS15080995.

[17] F. M. Talaat, S. A. Gamel, R. M. El-Balka, M. Shehata, and H. ZainEldin, "Grad-CAM Enabled Breast Cancer Classification with a 3D Inception-ResNet V2: Empowering Radiologists with Explainable Insights," Cancers, vol. 16, no. 21, p. 3668, Oct. 2024, doi: 10.3390/CANCERS16213668.

[18] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," Lancet Digit. Heal., vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.

[19] N. Urr Rehman et al., "A Patient-Split, Multi-Centre Study of CNN–Transformer Synergy for Explainable Breast-MRI Diagnosis," IEEE Trans. Consum. Electron., vol. 72, no. 2, pp. 4641–4653, May 2026, doi: 10.1109/TCE.2026.3678174.

[20] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "Data Descriptor: A curated mammography data set for use in computer-aided detection and diagnosis research," Sci. Data, vol. 4, no. 1, pp. 170177-, Dec. 2017, doi: 10.1038/SDATA.2017.177.

[21] R. Fusco et al., "Engineering the Image Representation for Deep Learning in Contrast-Enhanced Mammography: A Systematic Analysis of Preprocessing and Anatomical Masking," Bioeng., vol. 13, no. 3, p. 322, Mar. 2026, doi: 10.3390/BIOENGINEERING13030322.

[22] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in Graphics Gems IV, 1994, doi: 10.5555/180895.180940.

[23] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022. Accessed: Jul. 24, 2026. [Online]. Available: https://github.com/facebookresearch/ConvNeXt

[24] M. Tan and Q. V Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," May 24, 2019, PMLR. Accessed: Jul. 24, 2026. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[25] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal networks," Neural Comput. Appl., vol. 32, no. 14, pp. 10209–10228, Jan 2020, doi: 10.1007/S00521-019-04559-1.

[26] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-Supervised Nets," Feb. 21, 2015, PMLR. Accessed: Jul. 29, 2026. [Online]. Available: https://proceedings.mlr.press/v38/lee15a.html

[27] S. Ghosh, C. B. Poynton, S. Visweswaran, and K. Batmanghelich, "Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) , vol. 15012 LNCS, pp. 632–642, 2024, doi: 10.1007/978-3-031-72390-2_59.

[28] F. Radenovic, G. Tolias, and O. Chum, "Fine-Tuning CNN Image Retrieval with No Human Annotation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1655–1668, Jul. 2019, doi: 10.1109/TPAMI.2018.2846566.

[29] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2019. [Online]. Available: https://arxiv.org/abs/1711.05101.

[30] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," Feb. 06, 2017. Accessed: Jul. 29, 2026. [Online]. Available: https://github.com/loshchil/SGDR

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[32] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., Oct. 2017, Accessed: Jul. 29, 2026. [Online]. Available: https://arxiv.org/pdf/1710.09412

[33] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive Batch Normalization for practical domain adaptation," Pattern Recognit., vol. 80, pp. 109–117, Aug. 2018, doi: 10.1016/J.PATCOG.2018.03.005.

[34] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/J.IPM.2009.03.002.

[35] W. J. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, no. 1, pp. 32–35, Jan. 1950, doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[36] M. E. Seker et al., "Diagnostic capabilities of artificial intelligence as an additional reader in a breast cancer screening program," Eur. Radiol., vol. 34, no. 9, pp. 6145–6157, Feb. 2024, doi: 10.1007/S00330-024-10661-3.

[37] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/RADIOLOGY.143.1.7063747.

[38] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, Sep. 1988, doi: 10.2307/2531595.

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," 2017. Accessed: Jul. 29, 2026. [Online]. Available: http://gradcam.cloudcv.org

[40] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," Proc. - 2018 IEEE Winter Conf. Appl. Comput.

Vision, WACV 2018, vol. 2018-January, pp. 839–847, May 2018, doi: 10.1109/WACV.2018.00097.

[41] A. E. Oyekanmi, Y. Sun, and J. O. Olamijuwon, "Deep Learning-Based Breast Cancer Classification in Mammography: A Lesion-Specific Approach with Hybrid Ensemble," IEEE Access, 2026, doi: 10.1109/ACCESS.2026.3709667.

[42] T. Mahmood, J. Li, Y. Pei, and F. Akhtar, "An Automated In-Depth Feature Learning Algorithm for Breast Abnormality Prognosis and Robust Characterization from Mammography Images Using Deep Transfer Learning," Biol., vol. 10, no. 9, p. 859, Sep. 2021, doi: 10.3390/BIOLOGY10090859.

[43] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," Jul. 17, 2017, PMLR. Accessed: Jul. 29, 2026. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[44] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a Full-field Digital Mammographic Database," Acad. Radiol., vol. 19, no. 2, pp. 236–248, Feb. 2012, doi: 10.1016/J.ACRA.2011.09.014.