International Journal of Artificial Intelligence and Digital Innovation
https://journals.hu.edu.so/index.php/ijaidi

Hybrid Hard-Negative Learning for Breast Cancer Risk Classification in Mammography with Explainable AI

Yusuf Abdisalam Yusuf, Samia Abdullahi Mohamed, Mohamed Abdi Yusuf, Abdinasir Omar Mohamed, Yusuf Mohamed Abas*
Department of Data Science, Faculty of Computer Science and Information Technology, Hormuud University, Mogadishu, Somalia

*Corresponding author's email: Yuzerzef2@gmail.com

Abstract
Whole-image mammography models must identify small abnormalities within predominantly normal tissue, while their predictions remain difficult to audit. This paper evaluates a Hybrid Hard-Negative learning pipeline for binary mammographic risk classification with dual-level Grad-CAM explanations. The VinDr-Mammo dataset was mapped operationally as BI-RADS 1-2 (negative/benign assessment) versus BI-RADS 4-5 (suspicious/high-risk assessment); BI-RADS 3 was excluded. The balanced subset contained 1,976 images, including a held-out test set of 296 images. Training used lesion, lesion-context, full-image, ordinary negative-tissue, and model-mined hard-negative patches. At inference, 512 × 512 windows were scanned at stride 128 and combined using 0.75 times the mean of the three highest patch probabilities plus 0.25 times the full-image probability. Seven CNN backbones were compared, and an attention-based Scanner-MIL configuration served as a weakly supervised baseline. MobileNetV3Large achieved the strongest observed scan-level performance (ROC-AUC 0.7877, sensitivity 0.7055, F1-score 0.7128; 43 false negatives among 146 BI-RADS 4-5 images) with 3.18 million parameters. The matched Scanner-MIL baseline achieved sensitivity 0.5616. When lesion locations were supplied, pretrained backbones reached ROC-AUC 0.969-0.973, whereas automatic scan-level performance was 0.7321-0.7877 (0.6414 for the non-pretrained custom CNN). The complete Hybrid pipeline produced 12 fewer false negatives than the whole-image MobileNetV3Large configuration. Qualitative Grad-CAM examples were consistent with localization-related failure but were not a quantitative localization test. The results support candidate localization as a major performance constraint and position the system as a research prototype rather than an autonomous diagnostic tool.

Keywords: Mammography, BI-RADS, Hard-negative learning, Risk classification, Explainable AI

1. Introduction

Breast cancer is the most frequently diagnosed cancer among women. In 2022, an estimated 2.3 million women were diagnosed and roughly 670,000 died from the disease [1]. Mammography remains the standard screening tool because it can reveal masses, calcifications, and tissue distortion before symptoms appear. The image is difficult to read, however. Low contrast, overlapping tissue, and subtle abnormalities can hide malignancy or make normal tissue appear suspicious [2]. A false positive can lead to extra imaging or biopsy. A false negative delays investigation of a cancer that is already present.

Deep learning has improved automated mammographic analysis. Convolutional neural networks learn image features directly and now support classification, detection, and screening workflows that were previously built around handcrafted rules [3]. Clinical studies also show that AI-supported reading can maintain detection while reducing workload in screening programs [4], [5]. Those results do not remove the structural difficulty of the task. A mammogram is large, but a lesion can occupy only a small part of it. A model trained on one label for the whole image receives a weak signal. Most pixels describe normal tissue, yet the label depends on a small abnormal region.

This attention-dilution problem shaped the first part of the study. A whole-image classifier can learn broad correlations without learning where the lesion is. It can also raise the score of benign structures that resemble a mass. Patch-based learning narrows the field of view, but easy benign patches do not teach the model how to handle the difficult negative regions that lie close to the malignant decision boundary. The Hybrid Hard-Negative pipeline therefore combines lesion-centred patches, surrounding context, full-image information, ordinary benign tissue, and benign patches that the intermediate model itself confuses with malignancy.
Opacity creates the second obstacle. A probability alone does not show whether the model looked at a lesion, dense tissue, a burnt-in label, or the edge of the image. Grad-CAM exposes the spatial evidence behind a convolutional model by converting class gradients into a heatmap [6]. In this study, Grad-CAM is generated twice because the final score has two sources: the suspicious scan patches and the full-image branch. Agreement between the two views provides a direct check on whether the decision rests on one coherent anatomical region.

The paper makes a controlled comparison rather than claiming that one architecture solves mammographic screening. The primary experiment compares seven backbones inside the same Hybrid pipeline. A separate experiment compares the Hybrid strategy with attention-based Multiple Instance Learning (MIL). MIL was included even though it ultimately performed worse because it answers a practical annotation question. It learns from image-level labels and asks the model to discover the lesion itself, while the Hybrid pipeline uses lesion annotations during patch construction. Holding the dataset, study-level partitions, preprocessing, and test protocol constant puts a number on the performance gained from stronger supervision.

The main empirical result is not simply that MobileNetV3Large ranked first. When the lesion location was supplied, every pretrained backbone reached a ROC-AUC between 96.9% and 97.3%. Once the same networks had to search the full image, performance dropped by 19 to 24 points and the spread between models widened to 14.63 points. MobileNetV3Large fell from 97.3% to 78.77%. The experiment therefore separates two tasks that are often reported together: recognizing lesion tissue and finding that tissue in a full mammogram. The second task caused most of the loss.

1.1 Research Objective

The overall objective was to determine whether hybrid lesion-focused and hard-negative patch learning improves BI-RADS risk classification and interpretability relative to weaker-supervision and whole-image alternatives. The specific objectives were to: (1) compare the Hybrid pipeline with Scanner-MIL on identical partitions; (2) compare seven CNN backbones under the Hybrid protocol; (3) quantify the performance difference between ROI-assisted, scan-level, and whole-image evaluation; and (4) use dual-level Grad-CAM for bounded qualitative error analysis and demonstrate the selected model in a research prototype.

2. Literature Review

2.1 Deep Learning for Mammographic Classification

Deep learning replaced much of the manual feature engineering used in earlier computer-aided diagnosis. CNNs learn edges, textures, shapes, and higher-level image patterns through stacked convolutional layers. Transfer learning made this approach practical for medical datasets that are much smaller than natural-image collections. Reviews published in 2024 describe CNNs as the dominant family for breast imaging and identify hybrid systems, ensembles, and explainability as the main directions of current work [2], [3].

Recent studies also moved beyond retrospective benchmark accuracy. The MASAI trial compared AI-supported screen reading with standard double reading and reported a clinical safety analysis under real screening conditions [4]. Lauritzen et al. [5] examined early indicators after AI was introduced into screening. These studies support AI as decision support, but they do not imply that a stand-alone classifier is ready to replace a reader. The operating point matters, and sensitivity remains the metric that carries the greatest clinical weight.

Work on VinDr-Mammo shows why task definition matters. Kebede et al. [7] combined dual-view classification with a detection branch and reported an AUC of 0.83 and F1 of 0.87. The detector recovered sensitivity when the classifier missed a lesion. Petrini and Kim [8] reported an AUC of 0.85 using multi-view transfer learning. Both systems use information that a single-image scan-level classifier does not. Their results are useful benchmarks, but they are not directly interchangeable with lesion-level, patch-level, or single-view image-level results.

2.2 CNN Backbone Architectures

Backbone choice changes the feature representation available to the classifier. ResNet uses residual connections to train deep networks without losing the gradient signal [9]. DenseNet connects each layer to later layers and encourages feature reuse [10]. EfficientNet balances width, depth, and input resolution through compound scaling, while EfficientNetV2 modifies the design for faster training [11], [12]. MobileNetV3 uses neural architecture search and lightweight blocks for efficient inference [13]. ConvNeXt updates a convolutional network with design choices influenced by vision transformers [14].

Many studies select one of these networks and then attribute the final result to the full method. That makes it difficult to tell whether the learning strategy or the backbone produced the gain. This study holds the patch set, curriculum, aggregation rule, threshold-selection procedure, and test split constant while changing only the feature extractor. A shallow custom CNN trained from random initialization provides a lower bound. Its purpose is to measure what transfer learning contributes, not to compete on model size alone.

2.3 Patch-Based Learning

Patch learning responds directly to the scale mismatch between a lesion and a full mammogram. The network sees a small region where tissue detail is preserved instead of a downsampled image dominated by background and normal anatomy. Pasini et al. [15] described a fully automatic patch pipeline for masses and microcalcifications, reporting an accuracy of 81% and AUC of 0.94 while using Grad-CAM for interpretation. Shen et al. [16] showed that patch size and resolution affect detection and that combining scales can improve the result.

A patch pipeline still depends on how patches are sampled. A lesion crop teaches local appearance. A wider crop teaches the relation between the lesion and nearby tissue. A full-image view prevents the representation from becoming detached from breast anatomy. Benign patches teach normal variation. The fifth category in this study, mined hard negatives, focuses training on benign regions that the model scores as malignant. The patch taxonomy described

in Section 3.3 makes these roles explicit rather than treating every crop as an interchangeable training example.

2.4 Hard-Negative Mining

A random benign patch is often easy after the first stages of training. It contributes little because it lies far from the decision boundary. A dense overlap, fibroadenoma-like structure, or other benign region that receives a high malignant score carries more information. Hard-negative mining searches for those failures and returns them to training. Huang et al. [17] used the same general idea in whole-slide image classification, where difficult negatives sharpened the learned representation more efficiently than repeated random sampling.

The mammography use case is narrower. The aim is not only to lower false positives. A few high-scoring benign patches can dominate an image-level top-k aggregation rule. Suppressing those patches changes the final mammogram score even when average patch ranking changes little. This explains why the Hybrid curriculum can show patch-level overfitting while still improving the scan-level result. The deployed decision is made at image level, so the curriculum must be judged there.

2.5 Multiple Instance Learning

MIL replaces instance labels with a bag label. A mammogram becomes a bag of patches, and a malignant bag is assumed to contain at least one malignant instance. Zhu et al. [18] applied deep multi-instance learning to whole mammograms. Ilse et al. [19] introduced attention-based pooling that learns a weight for each instance before producing the bag representation. The weights provide a coarse indication of which patches influenced the prediction.

MIL is attractive because lesion boxes are expensive. Bobowicz et al. [20] reported an AUC near 0.90 with a weakly supervised attention system for breast lesions. Mourão et al. [21] extended attention MIL across scales and reported improvement over a single-scale design. The trade-off is that the network must learn representation, localization, and bag aggregation from a weaker label. The Scanner-MIL baseline in this study tests that trade-off on the same images used by the supervised pipeline. Its fixed EfficientNetV2B0 backbone keeps the comparison focused on the learning strategy. It does not support a general claim that every MIL design will behave the same way.

2.6 Explainable AI in Mammography

Grad-CAM is widely used because it produces a region-level explanation with one backward pass and works with standard convolutional backbones [6]. Alternatives such as raw saliency maps can be noisy, while perturbation methods require many model evaluations. Ameen [22] paired EfficientNetV2 with Grad-CAM++ for mammogram analysis. Ahmmed et al. [23] combined a lightweight CNN with SHAP and Grad-CAM. Camurdan et al. [24] went further by measuring how closely activation maps aligned with annotated lesions. Shifa et al. [25] reviewed XAI methods in mammography and emphasized that an attractive heatmap is not evidence by itself.

The present study uses explanations as an error-analysis tool. A correct heatmap should concentrate on plausible breast tissue. A false positive and a false negative should not be assumed to share one mechanism. The first may locate a real structure and classify it incorrectly. The second may never attend to the lesion. This distinction matters because the first calls for better discrimination, while the second calls for better localization or preprocessing.

2.7 Synthesis and Research Gap

The reviewed work leaves three connected gaps. Patch systems usually emphasize lesion crops but do not add model-mined benign failures. MIL systems reduce annotation cost but are rarely compared with a supervised lesion-focused system on identical partitions. Backbone choice is often fixed before the experiment, so architecture and learning strategy remain mixed. Explainability is commonly displayed after prediction, but less often used to identify which component failed.

This study joins those questions in one protocol. The Hybrid pipeline combines five patch types and one hard-negative stage. Seven backbones are compared within that pipeline. Scanner-MIL tests the cost of weaker supervision. ROI-assisted and scan-level evaluation separate recognition from localization. Grad-CAM then checks whether the observed errors follow the same explanation suggested by the numerical results. The research gap concerns how supervision, scanning, architecture choice, and explanation contribute to the final scan-level result. It is a controlled account of where performance is lost and what the annotation, scanning, and transfer-learning choices contribute.

3. Data and Method

3.1 Dataset Construction and Operational Labels

VinDr-Mammo contains 5,000 four-view examinations, equivalent to 20,000 mammographic images, with breast-level BI-RADS assessments and finding-level bounding boxes [26]. The study defined two operational classes from the breast-level assessments (Table 1). BI-RADS 1-2 images were assigned to the benign class, while BI-RADS 4-5 images were assigned to the malignant class. BI-RADS 3 was excluded because it represents a probably benign but clinically indeterminate assessment [27]. VinDr-Mammo does not provide pathology-confirmed outcomes. Therefore, malignant refers only to the BI-RADS 4-5 operational class and should not be interpreted as biopsy-confirmed malignancy. For readability, benign and malignant are used as shorthand for these two operational classes throughout the remainder of the paper.

Table 1. Operational BI-RADS Mapping Used for Binary Risk Classification

BI-RADS category	**Clinical interpretation**	**Assigned class**
1	Negative	BI-RADS 1-2
2	Benign finding	BI-RADS 1-2
3	Probably benign (indeterminate)	Excluded
4	Suspicious abnormality	BI-RADS 4-5
5	Highly suggestive of malignancy	BI-RADS 4-5

All 762 BI-RADS 4 and 226 BI-RADS 5 images were retained (n = 988). The BI-RADS 1-2 pool contained 13,406 BI-RADS 1 and 4,676 BI-RADS 2 images; stratified sampling retained 490 and 498, respectively. The balanced analysis subset therefore contained 988 BI-RADS 1-2 and 988 BI-RADS 4-5 images (Figure 1).

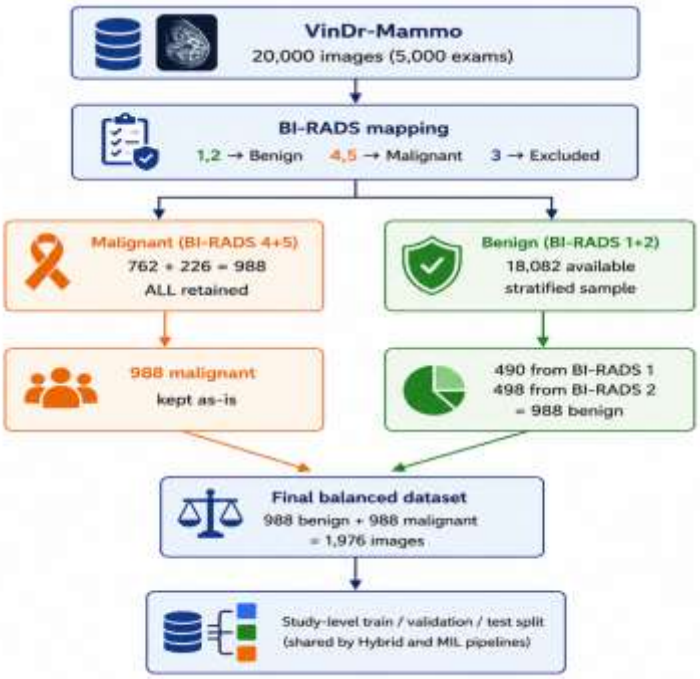


Figure 1. Dataset Construction Funnel

The held-out test set contained 296 images: 150 BI-RADS 1-2 and 146 BI-RADS 4-5. The remaining images formed training and validation partitions using study-level separation, so views from one examination remained in one partition. Exact train/validation class counts were not recorded. Both pipelines used the same retained partitions.

3.2 Preprocessing

All images passed through one preprocessing path before either pipeline used them. DICOM files were converted to 8-bit PNG images, and MONOCHROME1 images were inverted so that intensity followed one convention. A tissue mask was formed by thresholding and retaining the largest connected component. Cropping to that component removed most black background, embedded labels, and scanner artifacts. The remaining breast image was normalized with a 0.5th to 99.5th percentile stretch, followed by gamma correction toward a target median brightness. Images and extracted patches were resized to 512 × 512 pixels.

The shared preprocessing is part of the experimental control. If the Hybrid pipeline received cleaner images than Scanner-MIL, the comparison would measure two changes at once. The same masking, normalization, resolution, and study partitions were therefore used before the learning strategies diverged.

3.3 Hybrid Hard-Negative Pipeline

The primary pipeline learns from five patch types. ROI lesion patches crop the annotated abnormality. ROI context patches include tissue around the lesion. Full-image patches preserve global structure. Benign tissue patches provide straightforward negative examples. Hard negatives are benign regions that the intermediate model scores as malignant. Figure 2 shows the taxonomy.

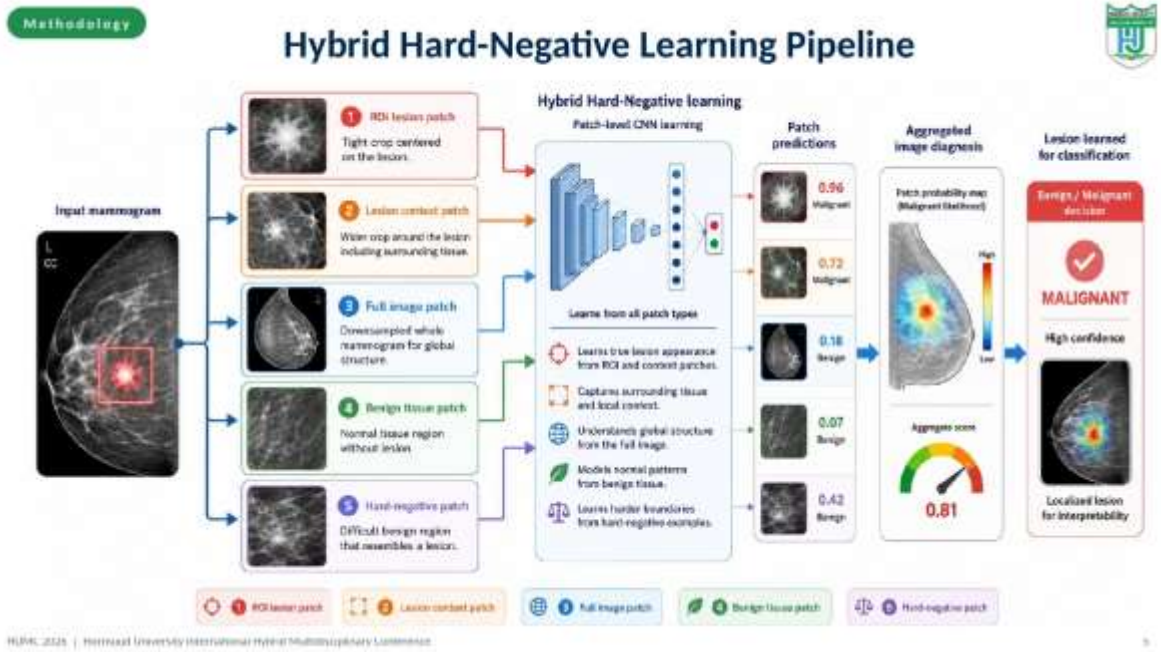


Figure 2. Hybrid Hard-Negative learning pipeline and the five patch types used for training and scan-level prediction.

Training used a curriculum rather than mining from the first epoch (Table 2). The first stage froze the pretrained backbone and trained only the classification head. The second stage partially unfroze the upper backbone layers and fine-tuned the network on the original patch set. The resulting model then scanned benign training images. Benign patches receiving a malignant probability at or above the mining threshold were retained, up to a capped total. The third stage added those mined patches and fine-tuned the network again. Mining occurred once. This made the search computationally manageable and ensured that a hard negative was defined by a model that had already learned a reasonable boundary.

Table 2. Three-Stage Hybrid Training Curriculum

Stage	**Frozen components**	**Trainable components**	**Data change**
1. Frozen backbone	CNN backbone	Classification head	Original lesion, context, full-image, and benign patches
2. Partial fine-tuning	Lower backbone layers	Upper backbone layers and head	Same patch set; intermediate model produced
3. Hard-negative fine-tuning	Lower layers remain frozen	Unfrozen backbone layers and head	Mined benign false positives added

3.4 Scanner-MIL Baseline

Scanner-MIL treats one mammogram as a bag of candidate patches. The patches do not receive instance labels. EfficientNetV2B0 extracts a feature vector for each patch, gated attention assigns a weight, and the weighted vectors form one bag representation for benign versus malignant classification. The automatic scanner variant uses no lesion location at inference. An ROI-guided variant was also built to test how the method behaves when sampling is biased

toward known lesion regions, but the main strategy comparison uses the automatic scanner result.

MIL was included because it removes the lesion-box requirement. That is a real advantage in datasets where only image-level assessments exist. Its disadvantage is equally clear: the model must discover the relevant instance from the bag while learning the class boundary. Figure 3 places both approaches in one architecture. The figure also shows why the seven-backbone experiment belongs only to the Hybrid branch. The MIL branch stays fixed so that its result answers an annotation and learning-level question rather than a second architecture search. Table 3 summarizes the design of the two pipelines.

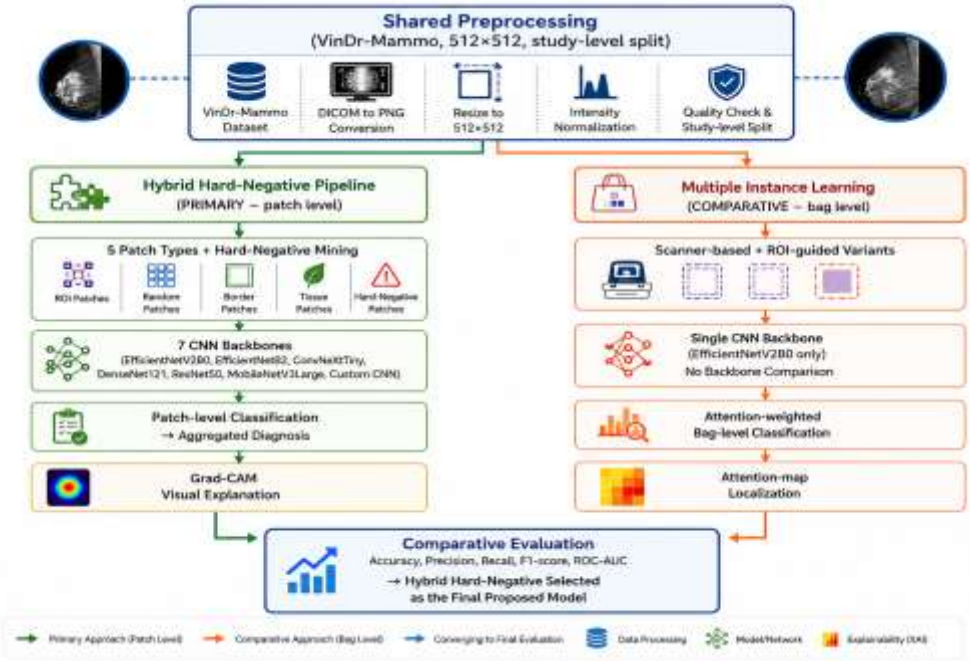


Figure 3. Two-Pipeline Architecture and Matched Evaluation Flow

Table 3. Side-by-Side Design of the Two Learning Pipelines

Design feature	**Hybrid Hard-Negative**	**Scanner-MIL**
Learning level	Individually labelled patches	Image-level bags of unlabeled patches
Supervision	Breast label plus lesion boxes for patch construction	Breast label only in automatic scanner mode
Aggregation	Top-3 patch mean combined with full-image probability	Gated attention pooling over patch features
Backbones evaluated	Seven	EfficientNetV2B0 fixed
Main question	Which supervised patch design and backbone work best?	Can automatic lesion discovery match stronger supervision?

3.5 Backbones, Training, and Model Selection

The Hybrid comparison covered EfficientNetV2B0, EfficientNetB2, ConvNeXtTiny, DenseNet121, ResNet50, MobileNetV3Large, and a shallow custom CNN. The six established networks used ImageNet pretraining; the custom CNN began with random weights. Each experiment used 512 × 512 inputs, mixed-precision training, weighted sampling where required, and a fixed seed within that run. Training was not repeated over a common seed set for every backbone.

Checkpoints and the decision threshold were selected using validation data; the final MobileNetV3Large threshold was 0.405. The held-out set was then used for comparative evaluation. MobileNetV3Large produced the strongest observed test-set result and was selected for prototype deployment after that comparison; the deployment choice is therefore not presented as independent pre-test model selection.

3.6 Scan-Level Inference and Aggregation

At scan-level inference, the masked image was tiled into 512 × 512 windows with stride 128. Any window containing less than 10% breast tissue was discarded. Each retained patch was classified with horizontal-flip test-time augmentation. The full resized mammogram was classified in parallel. The image probability was then computed as 0.75 times the mean of the three highest patch probabilities plus 0.25 times the full-image probability. A score at or above 0.405 was assigned to the malignant class.

The rule reflects the lesion geometry. Averaging every patch dilutes a small lesion with many normal regions. Taking only the maximum makes one accidental high score decisive. The top-three mean reduces dependence on one patch while still emphasizing suspicious regions. The full-image branch adds context, but receives the smaller weight because resizing the entire breast can suppress local detail.

3.7 Evaluation and Explainability

The test analysis reports accuracy, balanced accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, and PR-AUC. ROC-AUC ranks discrimination across thresholds, while sensitivity and the absolute false-negative count expose the errors that matter most in screening. ROI-assisted evaluation supplies the lesion crop and measures recognition with location known. Scan-level evaluation withholds lesion location and measures the deployed task.

Grad-CAM was computed from the last convolutional layer of MobileNetV3Large. One heatmap was produced for the highest-scoring scan patch and another for the full-image branch. Quantitative overlap with lesion boxes was planned, but the completed evaluation inspected six representative cases qualitatively and did not calculate intersection-over-union or activation mass across the full test set. This paper reports the completed analysis and treats the missing quantitative XAI evaluation as a limitation.

4. Results and Discussion

4.1 Hybrid Hard-Negative versus Scanner-MIL

Table 4 compares the strongest Hybrid configuration, MobileNetV3Large, with Scanner-MIL using EfficientNetV2B0. Both were evaluated on the same 296 images. The Hybrid pipeline beat the baseline on every metric. Its specificity was only 0.66 percentage points higher, so benign recognition was nearly the same. The decisive difference was sensitivity: 70.55% against 56.16%, a gap of 14.39 points.

Table 4. Hybrid Hard-Negative versus Scanner-MIL on the Test Set (n = 296)

Metric	Hybrid	Scanner-MIL	Difference
Accuracy	0.7196	0.6453	+7.43 points
Balanced accuracy	0.7194	0.6442	+7.52 points
Precision	0.7203	0.6667	+5.36 points
Sensitivity	0.7055	0.5616	+14.39 points
Specificity	0.7333	0.7267	+0.66 points
F1-score	0.7128	0.6097	+10.31 points
ROC-AUC	0.7877	0.7308	+5.69 points
PR-AUC	0.8010	0.7516	+4.94 points

Scanner-MIL missed 64 of the 146 malignant cases. The Hybrid pipeline missed 43. The 21-case difference explains the sensitivity and F1 gaps more clearly than a general statement about higher performance. The two methods were similarly able to reject benign images, but the supervised patch pipeline identified more malignant cases. On this dataset, lesion-level annotation bought 14.39 sensitivity points over automatic attention-based discovery.

This result does not show that MIL is inherently unsuitable for mammography. The baseline used one backbone and one attention design. A different encoder, pooling function, or training set could improve it. The controlled conclusion is narrower: EfficientNetV2B0 Scanner-MIL did not match the Hybrid pipeline under this protocol. The value of including MIL is that the annotation trade-off becomes measurable rather than assumed.

4.2 ROI-Assisted versus Automatic Scan-Level Classification

Table 5 reports the easier condition. Each network received an ROI crop around a known lesion. The pretrained backbones clustered between 0.969 and 0.973 ROC-AUC, a spread of 0.004. The shallow custom CNN reached 0.920. Once the lesion was placed in view, architecture choice had little effect among the pretrained models.

Table 5. ROI-Assisted Performance on the Test Set

Backbone	**Accuracy**	**Precision**	**Recall**	**F1**	**Specificity**	**ROC-AUC**
ResNet50	0.956	0.978	0.932	0.954	0.980	0.973
MobileNetV3Large	0.946	1.000	0.890	0.942	1.000	0.973
EfficientNetB2	0.959	0.993	0.925	0.957	0.993	0.972
DenseNet121	0.959	0.986	0.932	0.958	0.987	0.972
EfficientNetV2B0	0.939	0.964	0.911	0.937	0.967	0.971
ConvNeXtTiny	0.932	0.920	0.945	0.932	0.920	0.969
Shallow Custom CNN	0.936	0.985	0.884	0.931	0.987	0.920

Table 6 reports the deployed condition. The network scanned the full mammogram and had to choose which regions mattered. MobileNetV3Large led with 0.7877 ROC-AUC, but the custom CNN fell to 0.6414. The spread widened from 0.004 in the ROI condition to 0.1463 at scan level. Every model lost between 18 and 28 ROC-AUC points when lesion location was withdrawn.

Table 6. Scan-Level Hybrid Performance on the Test Set

Backbone	**Accuracy**	**Precision**	**Recall**	**F1**	**Specificity**	**ROC-AUC**
MobileNetV3Large	0.7196	0.7203	0.7055	0.7128	0.7333	0.7877
ResNet50	0.6757	0.7273	0.5479	0.6250	0.8000	0.7803
ConvNeXtTiny	0.6655	0.6822	0.6027	0.6400	0.7267	0.7649
DenseNet121	0.6723	0.7475	0.5068	0.6041	0.8333	0.7472
EfficientNetB2	0.6554	0.6618	0.6164	0.6383	0.6933	0.7465
EfficientNetV2B0	0.6622	0.6716	0.6164	0.6429	0.7067	0.7321
Shallow Custom CNN	0.6047	0.6179	0.5205	0.5651	0.6867	0.6414

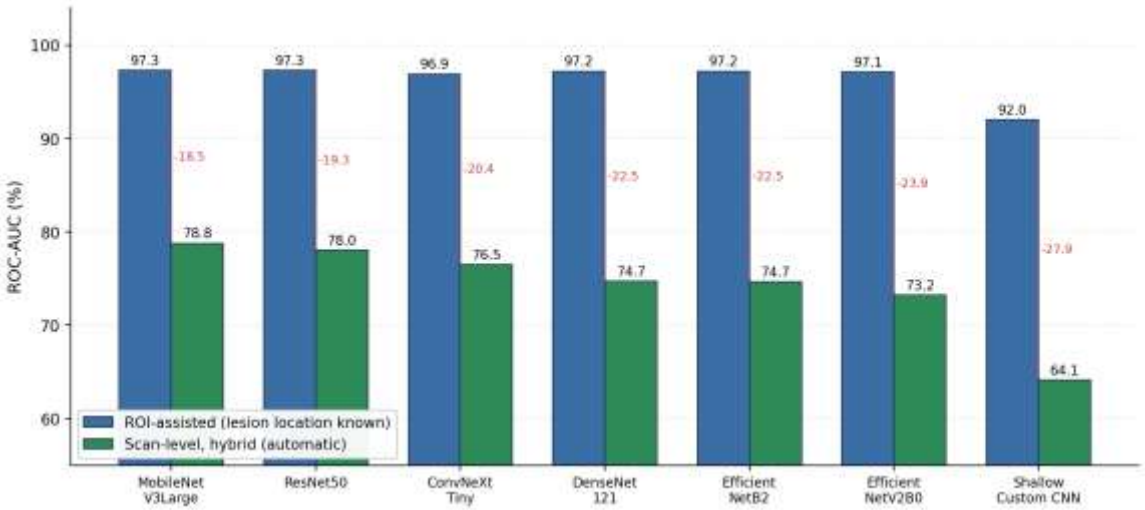


Figure 4. ROI-assisted versus automatic scan-level ROC-AUC for the seven evaluated backbones.

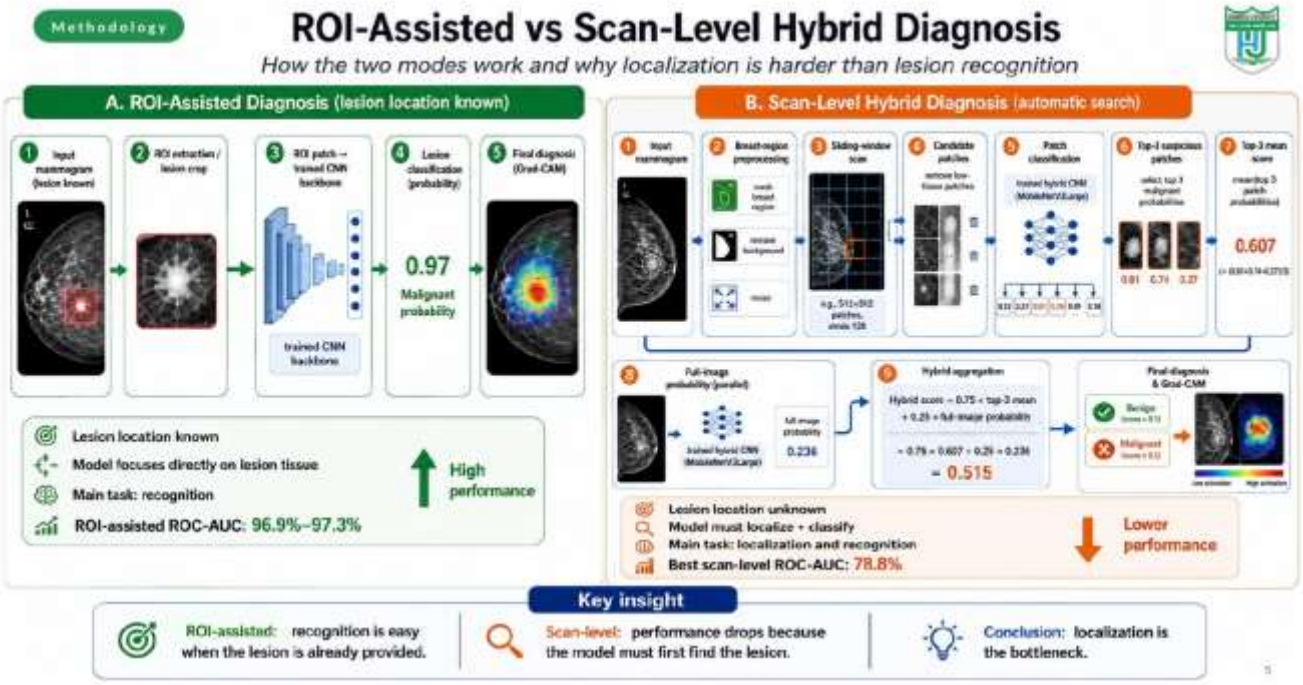


Figure 5. Comparison of ROI-assisted classification and automatic scan-level Hybrid classification.

Figures 4 and 5 show a consistent gap between the ROI-assisted and automatic scan-level conditions. MobileNetV3Large declined from 0.973 to 0.7877 ROC-AUC, and the custom CNN from 0.920 to 0.6414. Because the two conditions also differ in available context and aggregation, the gap is best interpreted as evidence that automatic candidate localization and scan-level aggregation are major constraints, not as a pure causal estimate of localization alone. The custom CNN clarifies what transfer learning contributes. It learned a useful lesion representation from scratch, reaching 0.920 ROC-AUC on cropped ROIs. It failed when exposed to the variety of normal tissue encountered during scanning. ImageNet pretraining therefore contributed less to judging an isolated lesion than to separating lesion-like regions from the many normal patterns traversed by the window.

4.3 Backbone Selection and Class-Specific Behaviour

MobileNetV3Large was selected because its class-specific behaviour was stronger, not because it won one aggregate metric. It reached sensitivity of 0.7055 and specificity of 0.7333, a gap of 2.8 points. It was the only backbone above 70% sensitivity and produced the fewest false negatives, 43. DenseNet121 reached the highest specificity, 0.8333, and the highest precision, 0.7475, but detected only 0.5068 of malignant cases and produced 72 false negatives. High specificity obtained by avoiding high-risk calls is not a strong screening result.

ResNet50 came close in ROC-AUC at 0.7803, only 0.0074 below MobileNetV3Large, but its sensitivity was 0.5479 and it missed 66 malignant cases. The operating threshold exposes a difference that ROC-AUC alone does not. MobileNetV3Large achieved the best accuracy, sensitivity, F1-score, and ROC-AUC at the selected operating point, while keeping the two class recalls close.

Table 7. Model Size, ROI Performance, and Scan Performance

Backbone	**Parameters (M)**	**ROI accuracy**	**ROI ROC-AUC**	**Scan accuracy**	**Scan ROC-AUC**	**AUC drop**
ResNet50	23.98	0.956	0.973	0.676	0.780	-19.3 points
MobileNetV3Large	3.18	0.946	0.973	0.720	0.788	-18.6 points
EfficientNetB2	8.13	0.959	0.972	0.655	0.746	-22.6 points
DenseNet121	7.23	0.959	0.972	0.672	0.747	-22.5 points
EfficientNetV2B0	6.17	0.939	0.971	0.662	0.732	-23.9 points
ConvNeXtTiny	28.02	0.932	0.969	0.666	0.765	-20.4 points
Shallow Custom CNN	1.32	0.936	0.920	0.605	0.641	-27.9 points

Model size did not predict performance (Table 7). ConvNeXtTiny used 28.02 million parameters, while MobileNetV3Large used 3.18 million. The smaller network produced the higher scan-level ROC-AUC. ResNet50 used 23.98 million and also ranked below it. The balanced subset contained 1,976 images, which likely left the larger models with more capacity than the data could constrain. The validation-to-test results support that reading. ConvNeXtTiny fell by 4.4 points and EfficientNetV2B0 by 6.8, while MobileNetV3Large increased by 0.5 points. It was the only backbone with no test degradation.

The selection therefore favored the smallest pretrained network in the comparison, not despite its size but because it combined efficiency with the most stable test behaviour. Choosing ConvNeXtTiny from its 0.809 validation ROC-AUC would have produced a weaker test result of 0.7649. Choosing DenseNet121 from precision or specificity would have produced 72 false negatives. MobileNetV3Large is supported by the combined evidence: highest scan ROC-AUC, highest sensitivity, highest F1-score, smallest class gap, fewest false negatives, least ROI-to-scan loss, and stable generalization.

4.4 Aggregation and Whole-Image Ablation

The aggregation experiment compared full-image probability, maximum patch score, mean patch score, top-three mean, and the 0.75/0.25 hybrid rule. The mean over all patches was consistently weak because normal patches outnumbered lesion patches. The maximum and top-three rules recovered between 0.02 and 0.11 ROC-AUC relative to weaker aggregation. The hybrid rule was applied to every backbone without per-model tuning and gave the best or near-best result in each case.

Table 8 removes scanning entirely. Each backbone classified one resized 512 × 512 mammogram. Every model degraded. MobileNetV3Large fell from 0.7877 to 0.738 ROC-AUC, and false negatives rose from 43 to 55. The complete Hybrid configuration, combining patch scanning, lesion-focused training, hard-negative learning, and hybrid aggregation, produced 12 fewer false negatives than the whole-image MobileNetV3Large configuration; the design does not isolate the independent effect of hard-negative mining.

Table 8. Full-Image Classification Without Patch Scanning

Backbone	**Accuracy**	**Precision**	**Recall**	**F1**	**Specificity**	**ROC-AUC**	**FN**
ResNet50	0.669	0.726	0.527	0.611	0.807	0.743	69
MobileNetV3Large	0.682	0.700	0.623	0.659	0.740	0.738	55
ConvNeXtTiny	0.622	0.637	0.541	0.585	0.700	0.708	67
EfficientNetV2B0	0.608	0.621	0.527	0.570	0.687	0.693	69
DenseNet121	0.591	0.662	0.349	0.457	0.827	0.684	95
EfficientNetB2	0.615	0.610	0.610	0.610	0.620	0.678	57
Shallow Custom CNN	0.527	0.520	0.534	0.527	0.520	0.546	68

DenseNet121 illustrates why the ablation matters. Its full-image specificity was 0.827, but recall collapsed to 0.349 and it missed 95 of 146 malignant images. ResNet50 reached the highest full-image ROC-AUC, 0.743, yet MobileNetV3Large had higher recall and produced 14 fewer false negatives. The extra scan stage is not a cosmetic addition. It changes which tissue reaches the decision and protects sensitivity.

4.5 Final Model, Confusion Matrix, and Error Pattern

At the validation-tuned threshold of 0.405, MobileNetV3Large correctly classified 110 of 150 benign images and 103 of 146 malignant images. It produced 40 false positives and 43 false negatives (Figure 6). Accuracy was 0.7196, balanced accuracy 0.7194, precision 0.7203, sensitivity 0.7055, specificity 0.7333, F1-score 0.7128, ROC-AUC 0.7877, PR-AUC 0.8010, negative predictive value 0.7190, and Matthews correlation coefficient 0.439.

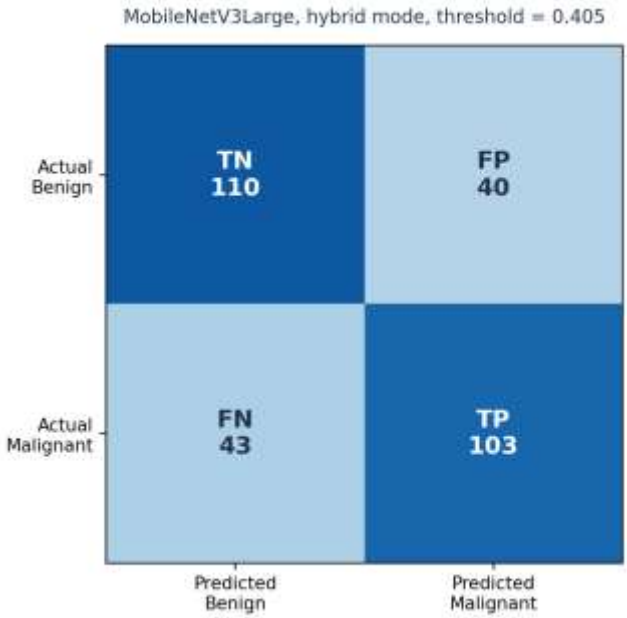


Figure 6. Confusion Matrix of the Final MobileNetV3Large Hybrid Model

The seven backbones agreed on the correct answer for only 99 of 296 images and agreed on the wrong answer for 21. They disagreed on the remaining 176, showing that the networks learned different decision boundaries. Fifteen of the 21 universally missed images were malignant. Every one of those fifteen was BI-RADS 4, and none was BI-RADS 5. The common failures were suspicious but not definitive lesions rather than the clearest malignant cases.

Dense tissue also appeared more often in the hardest group. Category D represented 9.5% of the 21 images missed by all models, compared with 4.1% of the test set. This pattern supports a specific interpretation. Residual error is concentrated in borderline BI-RADS 4 abnormalities whose margins are obscured by dense tissue. That is where candidate localization is hardest and where human readers also face overlapping anatomy.

4.6 Grad-CAM Explains Two Different Failure Mechanisms

Correct malignant predictions produced compact activations. The patch heatmap and full-image heatmap focused on the same fibroglandular region rather than spreading across the breast. Correct benign cases showed the opposite pattern, with weak or dispersed activation and no dominant focus. These explanations fit the intended decision rule: a malignant score should be supported by a localized abnormal region.

The error cases were more informative. The false positive was a BI-RADS 2 image with category C density and a hybrid probability of 0.640. Both heatmaps focused on a plausible dense region inside the breast. The network located tissue that could resemble malignancy and classified it incorrectly. That is a discrimination error and the type of case hard-negative mining was designed to reduce.

The false negative was a BI-RADS 4 image assigned a probability of 0.014. Its heatmaps concentrated near the image border and did not settle on the lesion-bearing parenchyma. The network did not inspect the lesion and decide it was benign. It failed to place the lesion among the regions that controlled the score. Figure 7 contrasts the two mechanisms.

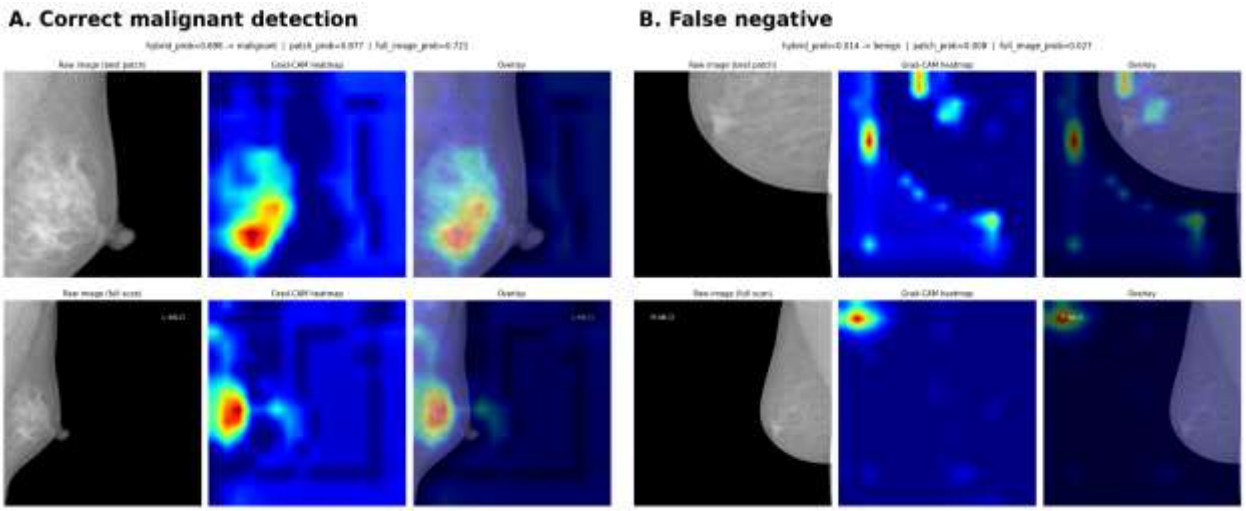


Figure 7. Dual-level Grad-CAM comparison of a correct malignant classification and a false negative.

This asymmetry independently supports the ROI-versus-scan result. With location supplied, every pretrained backbone exceeded 0.969 ROC-AUC. Without location, all lost 0.18 to 0.28. The false-negative heatmap shows how that loss can occur in one image: the top-scoring patch comes from the wrong place, and aggregation never receives a strong lesion score. The confusion matrix records only one false negative. Grad-CAM shows that the failed component was localization.

The XAI evidence remains qualitative. Six representative cases were inspected, and this study did not compute overlap between activation maps and lesion boxes across the full malignant test subset. The examples can diagnose plausible mechanisms, but they cannot establish how frequently each mechanism occurs or rank backbones by explanation quality.

4.7 Deployed System

The selected model was served through a FastAPI back end and a React front end styled with Tailwind CSS. The server loads MobileNetV3Large at startup, decodes an uploaded grayscale image, builds a tissue mask, applies the same scan and aggregation rule, and returns the predicted class, hybrid probability, risk level, inference time, patch statistics, and six Grad-CAM renderings. The interface provides Home, About, Analyze, and Help views. It accepts PNG or JPEG mammograms up to ten megabytes and displays a medical disclaimer.

The served pipeline reproduces the evaluated computation. It uses 512 × 512 windows, stride 128, the 10% tissue filter, horizontal-flip test-time augmentation, the top-three mean, the 0.75 and 0.25 weights, and the 0.405 threshold. Tissue masking also suppresses peripheral non-breast regions and burnt-in view labels. The masked deployment was not re-evaluated on the held-out test set, so the reported test sensitivity remains 0.7055. No improved figure is claimed from the additional deployment constraint.

The worked example in Figure 8 used a 1024 × 1024 right craniocaudal scan. The scan produced fourteen retained patches. The strongest patch scored 0.794, the top-three mean was 0.765, and the full-image probability was 0.685. The final score was 0.75(0.765) + 0.25(0.685) = 0.745. It exceeded 0.405 and was reported as malignant at 74.5% confidence. Inference took 24.71 seconds. The report displayed the classification beside the patch-level and full-image explanations so that the user could inspect the evidence with the score.

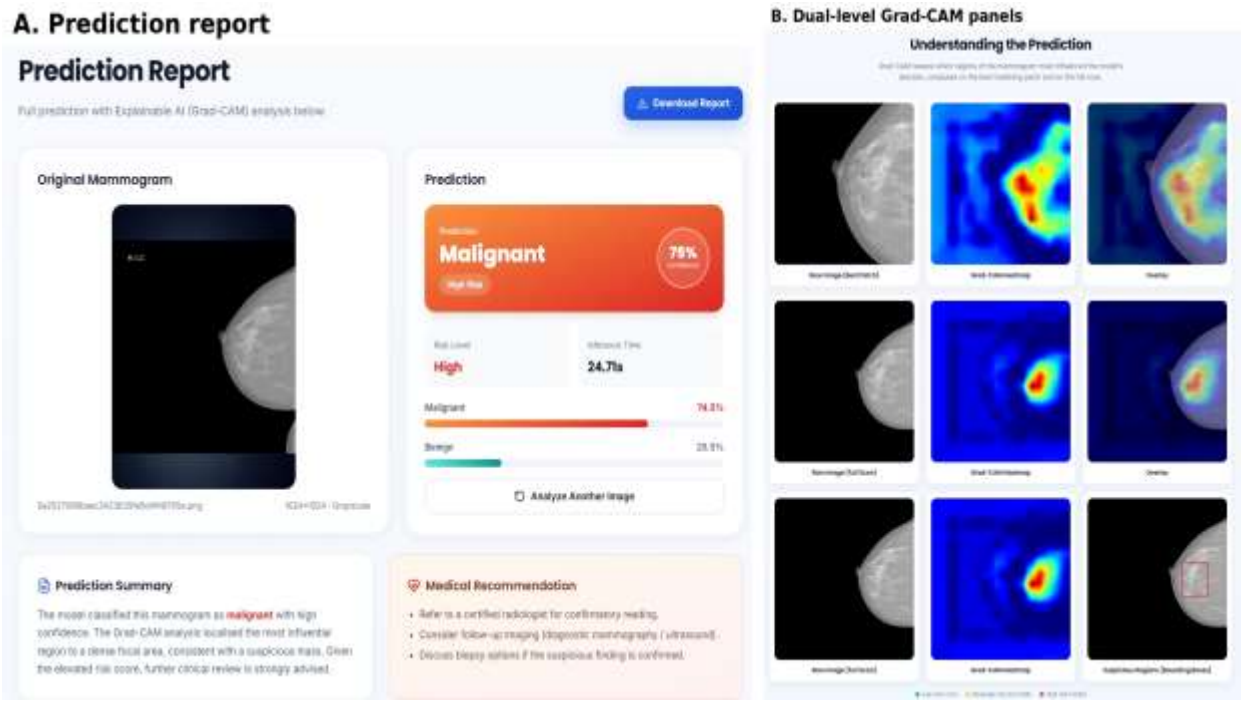


Figure 8. Deployed Prediction Report and Dual-Level Grad-CAM Panels

Functional tests checked file validation, score arithmetic, and production of the six explanation images. The application was not tested with radiologists. No reader study measured whether the output improves or harms human interpretation. Latency was not optimized and was not an objective. The 24.71-second example should therefore be read as an observed prototype runtime, not a deployment benchmark.

4.8 Limitations

The study used a balanced subset of 1,976 images from one dataset and did not include external validation. The backbone results are completed individual experimental runs rather than means across repeated runs with identical seeds. Confidence intervals and statistical comparison tests were not computed; small differences between architectures must therefore be interpreted cautiously.

The scan-level sensitivity of 0.7055 does not support autonomous clinical use. Localization remains unsolved. The MIL result covers one EfficientNetV2B0 configuration, not the full class of attention-based methods. The task is binary and uses one image at a time. It does not combine craniocaudal and mediolateral oblique views, predict lesion subtype or BI-RADS category, or include history, prior imaging, genetics, or other modalities.

Grad-CAM alignment was inspected qualitatively rather than measured across the test set. The deployed tissue mask was not re-evaluated offline, so its effect on the 43 false negatives is unknown. The application received functional testing only. There was no clinical user test and no reader study. Inference latency was never optimized. These limitations should be considered when interpreting the reported results.

5. Conclusion

The Hybrid Hard-Negative pipeline outperformed Scanner-MIL on every reported test metric under a matched protocol. Its main gain was sensitivity, 0.7055 against 0.5616. MobileNetV3Large was the strongest backbone at scan level, with 0.7877 ROC-AUC, 0.7055 sensitivity, 0.7128 F1-score, 43 false negatives, and a 2.8-point sensitivity-specificity gap. It

achieved that result with 3.18 million parameters and was the only backbone that did not degrade from validation to test.

A consistent gap separated ROI-assisted recognition from automatic scan-level classification: pretrained backbones achieved 0.969-0.973 ROC-AUC with lesion location supplied but 0.7321-0.7877 at scan level. The complete Hybrid MobileNetV3Large configuration produced 12 fewer false negatives than its whole-image counterpart. Together with the qualitative Grad-CAM examples, these findings identify candidate localization and aggregation as priority areas for improvement; they do not prove that hard-negative mining alone caused the observed gain.

Future work should use detection-style region proposals or box-supervised attention, repeat each backbone across identical seeds with confidence intervals, validate externally without retraining, and quantify explanation overlap against lesion boxes. Multi-view breast-level modeling and a prospective reader study are also required. Until those steps are completed, the system should remain a research prototype for decision-support investigation, not an autonomous diagnostic system.

Data and Reproducibility Statement

VinDr-Mammo is publicly available for research subject to its data-use terms [26]. The parameter values necessary to reproduce the reported inference rule are stated in Section 3.6. The analysis notebook documents the principal workflow; exact train/validation class counts and repeated-run uncertainty estimates were not recorded.

Conflict of Interest

The authors confirm that there is no conflict of interest related to this manuscript.

References

[1] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229–263, 2024.

[2] A. Carriero, L. Groenhoff, E. Vologina, P. Basile, and M. Albera, "Deep learning in breast cancer imaging: State of the art and recent advancements in early 2024," Diagnostics, vol. 14, no. 8, art. no. 848, 2024.

[3] J.-R. Pérez-Núñez, C. Rodríguez, L.-J. Vásquez-Serpa, and C. Navarro, "The challenge of deep learning for the prevention and automatic diagnosis of breast cancer: A systematic review," Diagnostics, vol. 14, no. 24, art. no. 2896, 2024.

[4] K. Lång, V. Josefsson, A.-M. Larsson, et al., "Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): A clinical safety analysis," The Lancet Oncology, vol. 24, no. 8, pp. 936–944, 2023.

[5] A. D. Lauritzen, M. Lillholm, E. Lynge, M. Nielsen, N. Karssemeijer, and I. Vejborg, "Early indicators of the impact of using AI in mammography screening for breast cancer," Radiology, vol. 311, no. 3, art. no. e232479, 2024.

[6] 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," International Journal of Computer Vision, vol. 128, no. 2, pp. 336–359, 2020.
[7] S. R. Kebede, F. G. Waldamichael, T. G. Debelee, M. Aleme, W. Bedane, B. Mezgebu, and Z. C. Merga, "Dual view deep learning for enhanced breast cancer screening using mammography," Scientific Reports, vol. 14, art. no. 3839, 2024.
[8] D. G. P. Petrini and H. Y. Kim, "Optimizing breast cancer detection in mammograms: A comprehensive study of transfer learning, resolution reduction, and multi-view classification," arXiv preprint arXiv:2503.19945, 2025.
[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[10] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
[11] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), vol. 97, 2019, pp. 6105–6114.
[12] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in Proc. 38th Int. Conf. Machine Learning (ICML), vol. 139, 2021, pp. 10096–10106.
[13] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, and Q. V. Le, "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.
[14] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986.
[15] G. Pasini, N. Lauciello, A. Finti, G. Russo, F. Marinozzi, F. Bini, and A. Stefano, "A fully automatic patch-based deep learning pipeline for mass and microcalcification detection in mammography: A preliminary study," in Image Analysis and Processing – ICIAP 2025 Workshops, vol. 16169, 2026, pp. 157–165, doi: 10.1007/978-3-032-11317-7_13.
[16] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," Scientific Reports, vol. 9, art. no. 12495, 2019.
[17] W. Huang, X. Hu, S. Abousamra, P. Prasanna, and C. Chen, "Hard negative sample mining for whole slide image classification," in Medical Image Computing and Computer Assisted Intervention (MICCAI 2024), Lecture Notes in Computer Science, vol. 15004, Springer, 2024.
[18] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in Medical Image Computing and Computer Assisted Intervention (MICCAI 2017), 2017, pp. 603–611, Springer.
[19] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in Proc. 35th Int. Conf. Machine Learning (ICML), vol. 80, 2018, pp. 2127–2136.
[20] M. Bobowicz, M. Rygusik, J. Buler, R. Buler, M. Ferlin, A. Kwasigroch, E. Szurowska, and M. Grochowski, "Attention-based deep learning system for classification of breast lesions: Multimodal, weakly supervised approach," Cancers, vol. 15, no. 10, art. no. 2704, 2023.
[21] M. Mourão, J. C. Nascimento, C. Santiago, and M. Silveira, "Multi-scale attention-based multiple instance learning for breast cancer diagnosis," in Medical Image Computing and Computer Assisted Intervention (MICCAI 2025), Lecture Notes in Computer Science, vol. 15974, 2025, pp. 364–374, Springer.

[22] M. Ameen, "Explainable mammogram analysis with EfficientNetV2 and Grad-CAM++ for robust cancer diagnosis," Diagnostics, vol. 16, no. 1, art. no. 105, 2026.
[23] J. Ahmmed, F. Ahmed, M. A. Kabir, M. T. Ahad, M. A. Jadoon, A. ur Rehman, and A. Bermak, "LBNet: An optimized lightweight CNN for mammographic breast cancer classification with XAI-based interpretability," Scientific Reports, vol. 16, art. no. 2013, 2026.
[24] O. Camurdan, T. Tanyel, E. Aktufan Cerekci, D. Alis, E. Meltem, N. Denizoglu, M. E. Seker, I. Oksuz, and E. Karaarslan, "Annotation-efficient, patch-based, explainable deep learning using curriculum method for breast cancer detection in screening mammography," Insights into Imaging, vol. 16, art. no. 60, 2025.
[25] N. Shifa, M. Saleh, Y. Akbari, and S. Al-Maadeed, "A review of explainable AI techniques and their evaluation in mammography for breast cancer screening," Clinical Imaging, vol. 123, art. no. 110492, 2025.
[26] H. T. Nguyen, H. Q. Nguyen, H. H. Pham, K. Lam, L. T. Le, M. Dao, and V. Vu, "VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography," Scientific Data, vol. 10, art. no. 277, 2023.
[27] M. S. Newell, S. V. Destounis, J. W. T. Leung, W. B. DeMartini, C. H. Lee, and P. R. Eby, ACR BI-RADS® v2025 Manual. Reston, VA, USA: American College of Radiology, 2025.