Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

BREAST LESION CLASSIFICATION: LEAKAGE-FREE PATCH-BASED TRANSFER LEARNING WITH CROSS-DATASET EVALUATION ON CBIS-DDSM, MIAS AND INBREAST

Hitesh Kag[1], Dinesh Kumar[2] and Anjali Diwan[3]

(Received: 3-Jul.-2026, Revised: 23-Aug.-2026 and 27-Aug.-2026, Accepted: 28-Aug.-2026)

ABSTRACT

High reported accuracies on small mammographic datasets are frequently inflated by data leakage, most often from augmentation applied before the train/test split. Changing only the split order on whole-image Mini-MIAS moves a deep-feature plus Support Vector Machine (SVM) pipeline from 0.976 accuracy to 0.513, which is equivalent to the random chance. The predictive signal survives at the lesion-patch scale. Our contribution is protocol-level rather than architectural: a reproducible leakage audit; the first quantification and repair of the reported patient overlap in the CBIS-DDSM official split, which we trace to its mass and calcification partitions being drawn independently (31 patients, 8.5% of test lesions); an architecture sweep that tests rather than assumes the performance ceiling; and a label-free procedure for transferring the decision threshold to a new institution. A lightweight EfficientNet patch pipeline trained on CBIS-DDSM under a strict patient-wise protocol obtains ROC-AUC 0.745 (95% CI 0.703–0.794); without retraining, it attains 0.734 (0.640–0.826) on MIAS and 0.877 (0.804–0.946) on INbreast and a DeLong test cannot distinguish the MIAS result from the internal one ($p = 0.82$). Results are stable over ten seeds ($\sigma \leq 0.011$) with clustered bootstrap intervals. Across 11 configurations spanning six architecture families, test AUC spans only 0.707–0.777. For small datasets, leakage control and external validation matter more than headline accuracy.

KEYWORDS

Breast lesion classification, Data leakage, Transfer learning, External validation, Mammography.

1. INTRODUCTION

Breast cancer is the most common cancer and a leading cause of cancer-related death among women worldwide and early detection remains the single most effective means of reducing mortality [1]. X-ray mammography, the principal population-screening modality, is difficult to interpret, because lesions are often small, of low contrast and partly obscured by overlapping fibro-glandular tissue; sensitivity therefore depends heavily on the reader's workload and experience [2]. These challenges have driven decades of computer-aided detection and diagnosis (CAD) research, a fast-moving field that has accelerated further with the advent of deep learning [3].

Convolutional neural networks (CNNs) currently dominate research on mammographic-image analysis. Because well-annotated mammograms are scarce, the prevailing approach is transfer learning, in which a backbone pre-trained on a large corpus of natural images is used as a feature extractor or fine-tuned to the target task. A systematic review of recent medical-imaging studies confirms a strong trend toward feature-extraction and fine-tuning methods, but also notes inconsistent reporting of configurations [4]. In mammography specifically, end-to-end CNNs with tailored pre-processing and augmentation have been shown to classify pre-segmented masses as benign or malignant [5], while lightweight hybrids that combine CNN features with classical classifiers remain attractive for smaller datasets [6].

Despite this progress, reported performance on small public datasets is often difficult to replicate and, on closer inspection, optimistic. Mini-MIAS studies, for example, frequently report high accuracies—such as 93.9% for normal-*versus*-abnormal classification [7] and 96% sensitivity at AUC 0.99 for normal/benign/malignant classification [8]—yet these are typically whole-image results obtained without explicit control of data leakage. One commonly overlooked problem is the application of augmentation before the train/test split: augmented instances (e.g., rotations) of a single image end up

1. H. Kag is with the Department of Computer Eng. (AI, ML & DS), Marwadi Uni., Rajkot, India. Email: hitesh07kag@gmail.com
2. D. Kumar is with the School of Artificial Intell., Bennett Uni. (The Times Group), Greater Noida, India. Email: dineshkr@ieee.org
3. A. Diwan is with the Department of Computer Engineering, Marwadi University, Rajkot, India. Email: anjali.diwan@ieee.org

in both sets, so the images are no longer independent and the reported scores present an over-optimistic picture [9]. Reproducibility and test–retest robustness are likewise rarely assessed, even though small changes in the input can significantly alter CNN predictions [10] and external (cross-dataset) validation—the strongest evidence of generalizability in clinical settings— remains relatively scarce [11]. In this work, we take a different approach: rather than chasing a headline accuracy figure, we focus on methodologically sound evaluation and quantify what is genuinely generalizable. Our contributions are as follows.

- **A reproducible data-leakage demonstration.** We isolate augment-before-split leakage in a controlled experiment that changes one variable. Holding the films, labels, features, classifier and rotation count fixed, splitting the augmented pool at random yields 0.976 accuracy and 0.997 ROC-AUC on whole-image Mini-MIAS, whereas splitting by source film first yields 0.513 accuracy and 0.500 ROC-AUC — chance and below the majority-class rate. We release a single script that runs both arms.
- **A quantified and repaired leakage channel in the CBIS-DDSM official split.** Zafari et al. [12] report that patient overlap exists between the pre-defined CBIS-DDSM training and test sets and discard that split for the purpose of dataset harmonization. We confirm the overlap independently and extend it in four ways: we identify the mechanism; namely that the mass and calcification partitions were drawn independently, so 31 patients contribute a mass case to one side and a calcification case to the other; we quantify it as 60 of the 704 official test lesions (8.5%), with 19 of the 31 patients sharing the same breast and view; we measure its effect on a patch-based transfer-learning pipeline; and we supply a repaired, strictly patient-disjoint split that preserves the official test set, so that results remain comparable with published CBIS-DDSM numbers.
- **A whole-image *versus* lesion-patch analysis.** We show that, under a strict leakage-free protocol, whole-image transfer learning performs far worse than lesion-scale patches for the datasets, backbones and input resolutions evaluated here—now demonstrated on CBIS-DDSM as well as Mini-MIAS—whereas lesion-scale patches carry meaningful and statistically significant signals.
- **Leakage-free cross-dataset external validation.** We train benign-*versus*-malignant classifiers on CBIS-DDSM and test them, without retraining, on MIAS (cross-institution) and INbreast (cross-modality, full-field digital mammography), reporting clustered bootstrap 95% confidence intervals, DeLong tests between datasets and classifier stability across ten random seeds.
- **A tested rather than assumed performance ceiling.** Eleven configurations spanning six architecture families and 2.2–86 M parameters, including ViT-B/16 and Swin-T, are trained under one identical leakage-free protocol. Heavier backbones do help—ViT-B/16 is significantly ahead of the lightweight baseline—but the entire family spans only 0.07 AUC, far short of the accuracies reported under leaky evaluation. The same design also separates backbone from input resolution and, in doing so, corrects a claim in the submitted version.
- **A label-free procedure for transferring the operating threshold.** We find that dis- crimination survives a change of institution, but calibration does not, so a threshold tuned at the source site misbehaves at the target. We give a rank-based transfer that repairs the operating point using only un-labelled target scores and show that on the like-for-like external cohort it recovers the operating point an oracle threshold would choose. We also identify and report, the condition under which it should not be used.

The remainder of the paper is organized as follows. Section 2 reviews related works; Section 3 details the datasets; and Section 4 describes the methodology, including patch extraction, pre-processing, models, the leakage-free protocol, the statistical evaluation and the threshold-transfer procedure. Section 5 presents the results and discusses them. Section 6 details the released reproducibility package. Section 7 tackles the study limitations. Finally Section 8 addresses the conclusions and future work.

2. RELATED WORKS

Deep and transfer learning. Although traditional CAD techniques for mammography were developed over the preceding decade, deep learning has since emerged as the most successful approach, as recent surveys make clear [3]. Under data scarcity, the most common strategy is transfer learning from ImageNet-pre-trained CNNs: such models, combined with augmentation, classify pre-segmented masses as benign or malignant [4], [13] and hybrid approaches classify masses as benign or malignant

or distinguish soft-tissue from bone tumours [5]. On small datasets, however, combining a CNN with a classical classifier (e.g., SVM or dRVFL) still tends to outperform the alternatives [6, 14].

Benchmarks. CBIS-DDSM is an officially train/test-split sub-set of DDSM, curated by using biopsy outcomes to select the best cases of the original DDSM; it was created specifically, because earlier results were difficult to replicate for lack of a standard evaluation set [15]. INbreast provides full-field digital mammography (FFDM) with precise annotations and BI-RADS assessments [16], while larger resources, such as VinDr-Mammo, further extend the landscape [17]. Mini-MIAS, although digitized and relatively small, remains popular and frequently yields strong reported results [7, 8]; its apparent performance, however, depends heavily on the evaluation protocol and its rigour. Attention has also turned to the quality of the benchmarks themselves. Zafari et al. [12] harmonized CBIS-DDSM, TOMPEI-CMMD and VinDr-Mammo into a common format, quantified distributional shifts in breast density and abnormality prevalence between them and reported patient overlap between the pre-defined CBIS-DDSM training and test sets, which led them to discard that split. Their concern is cross-dataset harmonization; ours is what the overlap costs; a classifier that keeps the official split, which we address in Sub-section 5.3.

Patch-based *vs.* whole-image. Whole-image models generally require substantially larger datasets and/or detection-and-aggregation pipelines, whereas patch- or ROI-based classification concentrates the lesion signal and is more effective with limited data [5]. Patch size and resolution matter: a multi-scale design achieved an AUC of ~0.72 on CBIS-DDSM, compared with ≈0.70 for a single-scale design [18].

Augmentation and leakage. For small datasets, augmentation is the most common remedy [19], but when it is applied before splitting, transformed copies of a single image can appear in both partitions, inflating apparent performance. This leakage is rarely controlled for and seldom reported, even when very high accuracies are claimed (e.g., 99.8% on small mammogram sets [20]) [8, 9].

External validation and recent models. Cross-dataset generalization is under-reported, yet crucial: stand-alone models trained to match radiologists in-distribution can perform poorly at different sites or on different scanners, motivating domain-generalization methods [11]. In recent years (2024–2026), research has advanced beyond simple CNNs along three main directions. First, by capturing long-range dependencies, Vision Transformers (ViTs) and hybrid CNN–Transformer models have been applied to tasks, such as BI-RADS and lesion classification, though they typically require thousands of images or extensive pre-training to train reliably—for example, a custom ViT for BI-RADS assessment [21], dual-branch ViT fusion for micro-calcifications [22] and comparative CNN-*versus*-ViT studies for mammography [23]. Second, vision–language foundation models, pre-trained on large collections of mammogram–caption pairs, can improve data efficiency and robustness when data is limited [24], but they perform best on large downstream sets. Third, explainable AI has become an expectation, with methods, such as Grad-CAM [25] and gradient-based attribution [26] routinely used across breast-cancer pipelines and dedicated reviews discussing their use and assessment in mammography [27].

These approaches are generally data-hungry and comparatively heavy to train and deploy. Our focus is complementary: a lightweight, rigorously validated baseline the contribution of which lies in honest, externally validated evaluation rather than architectural novelty. Specifically, our work stands apart in that it is not optimized for peak accuracy, but instead provides (i) a reproducible leakage demonstration, (ii) an explicit whole-image-*versus*-patch comparison and (iii) statistically quantified cross-institution and cross-modality external validation.

3. DATASETS

We use three public mammography datasets (Table 1). CBIS-DDSM [15] is the principal training and testing dataset, providing biopsy-confirmed benign and malignant lesion ROIs with an official train/test partition. The submitted version described that partition as “patient-disjoint”; the word has been struck here, because Sub-section 5.3 shows that the property does not hold once mass and calcification cases are pooled. The leakage study is conducted on Mini-MIAS, which also serves as a cross-institution external test set; lesion centres and radii are read directly from its annotations. INbreast [16] provides full-field digital mammography (FFDM) images and is used as a cross-modality external test set. All labels are mapped to two classes. For CBIS-DDSM, BENIGN and BENIGN_WITHOUT_CALLBACK are mapped to benign and MALIGNANT to malignant, so that the benign-*versus*-malignant separation

follows directly from the biopsy-confirmed labelling. For IN- breast, labels are BI-RADS-derived (1–3 benign, 4–6 malignant) rather than biopsy-confirmed as in CBIS-DDSM—a caveat quantified in Section 5.10.

Table 1 reports the exact composition of every partition at the patient, image, lesion and patch levels. Because each annotated lesion contributes exactly one network input, the lesion and patch counts coincide throughout and only the former is tabulated.

Table 1. Exact composition of every data partition: patients, source images and annotated lesions. Each lesion yields one input patch. The validation fold is carved out of the CBIS-DDSM training split by patient and is re-drawn for every run, so its size is given as a range over the ten seeds.

Dataset	Role	Patients	Images	Lesions	Benign	Malig.
CBIS-DDSM	official train	1248	2458	2863	1682	1181
of which	validation fold	188	—	385–461	—	—
CBIS-DDSM	official test	349	399	704	428	276
Mini-MIAS	external test	104	111	119	68	51
INbreast	external test	50	107	132	43	89

Patient-level independence. Independence was verified programmatically rather than assumed and the audit script is released with the code. Across all ten runs, no patient is ever shared between a training fold and its validation fold. Mini-MIAS films are distributed as left/right pairs, so patient identity is recovered as refnum/2; INbreast redacts patient identifiers in the public release, so its 107 images map onto 50 anonymization tokens, which we treat as patients for clustered resampling. Both external sets are held out in their entirety and never seen in training, so no patient can cross that boundary by construction. The audit did, however, expose one genuine violation inside the CBIS-DDSM official split itself, which we quantify and repair in Sub-section 5.3.

The three datasets are deliberately chosen to probe complementary axes of generalization. CBIS-DDSM, the largest of the three, ships with an official train/test split and biopsy-confirmed labels, eliminating two of the most common confounds—label noise and split arbitrariness—and is therefore adopted as our primary training set. Mini-MIAS, acquired on different equipment, but in the same modality (digitized screen-film), provides a cross-institution stress test, since its site is held out entirely. A more difficult, cross-modality challenge is posed by INbreast: its full-field digital mammograms (FFDMs) have different dynamic ranges, bit depths and noise characteristics from digitized film, so a model that transfers here must cope with a fundamentally different image- formation process. In this setting, we train a single model on CBIS-DDSM and test it—without retraining—on each external set in turn. Statistically, this zero-adaptation protocol disentangles two questions that are conflated in single-dataset studies: robustness to site (the Mini-MIAS axis) and robustness to acquisition technology (the INbreast axis), the latter of which cannot be assessed when training and test data is drawn from the same distribution. Finally, all three datasets are publicly available, so the entire benchmark—the training data, both external test sets and our splits—can be reproduced by any reader, preserving the full reproducibility of the evaluation.

4. METHODOLOGY

The pipeline (Figure 1) is deliberately lightweight, allowing a clean separation of the experimental variables of interest: leakage control, image-*versus*-patch framing, backbone, resolution and cross-dataset transfer.

4.1 Patch Extraction

The network input is a patch rather than a whole mammogram—specifically, a square patch centred on a lesion. For CBIS-DDSM, cropped-ROI images are provided directly. For MIAS, the lesion centre and radius are loaded from the metadata; because MIAS uses a bottom-left origin, the row coordinate is flipped (row = H-y) before cropping, which is necessary to avoid missing the lesion. The crop side length is scaled according to the marked lesion area. For INbreast, mass contours are extracted from the

per-image XML and the resulting bounding box is enlarged by 30% to provide surrounding context. Representative patches are illustrated in Figure 2.

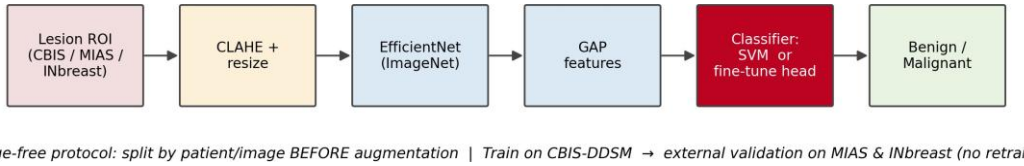


Figure 1. Leakage-free cross-dataset pipeline: lesion ROI→CLAHE→EfficientNet features→classifier, trained on CBIS-DDSM and externally validated on MIAS and INbreast.

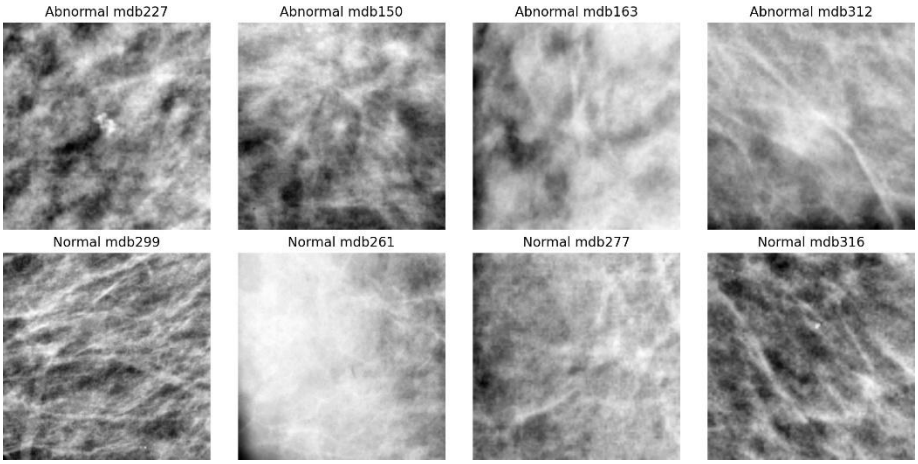


Figure 2. Representative lesion ROI patches after CLAHE enhancement.

4.2 Pre-processing

Each patch is converted to grayscale, resized to the network input size, contrast-enhanced with CLAHE (clip limit 2.0, 8 × 8 tiles) [28] and duplicated across three channels.

Normalization proceeds in three explicit stages, which we spell out here because the term is often used loosely in the literature. (i) Bit-depth windowing. Screen-film scans (CBIS-DDSM, Mini- MIAS) are already 8-bit and are used as-is. INbreast supplies 16-bit FFDM, so intensities are windowed to 8 bits by clipping at the 1st and 99th percentiles of each image and rescaling that interval linearly onto [0, 255]; percentile clipping rather than min–max is used because a handful of saturated detector pixels would otherwise compress the entire breast into a few grey levels. (ii) Local contrast enhancement. CLAHE is then applied to the resized patch. CLAHE is a local, histogram-equalization operator: it equalises the histogram within each 8×8 tile and clips the histogram at 2.0 times the mean bin height before equalizing, which limits noise amplification in the uniform fatty regions, after which tiles are recombined by bilinear interpolation. No global histogram equalization is applied. (iii) Intensity scaling. No z-score or min–max normalization is applied by us. Patches are handed to the network in the raw [0, 255] range, because the Keras EfficientNet models begin with their own Rescaling and Normalization layers that expect un- normalized input and internally apply the ImageNet channel statistics; pre-scaling the input would apply that transform twice. For the PyTorch backbones used in the extended architecture comparison of Sub-section 5.5, which carry no such internal layer, the equivalent transform is applied explicitly: division by 255 followed by subtraction of each model's ImageNet channel mean and division by its channel standard deviation.

4.3 Models and Training

The main pipeline uses three ImageNet-pretrained backbones (MobileNetV2, EfficientNetB0 and EfficientNetB3). These were chosen because compound scaling gives the EfficientNet family a markedly better accuracy-per-parameter trade-off than earlier CNNs at the scale of data available here [29], because the same family has been reported to work well on comparably small medical- imaging sets—including an improved EfficientNetB3 for melanoma lesion classification [30] — and because the whole study had to run on a single 6 GB consumer GPU. We acknowledge, however, that three members

of one efficiency-oriented family cannot by themselves establish that performance is limited by the data rather than by model capacity. Sub-section 5.5 therefore extends the comparison to ResNet-50 [31], DenseNet-121 [32], ConvNeXt-T [33], ViT-B/16 [34] and Swin-T [35] under one identical protocol, so that the claim is tested and not assumed.

We evaluate the three main backbones under two regimes. **Frozen + SVM:** the frozen backbone with global average pooling yields a fixed-length descriptor, which is standardized (zero mean, unit variance) and classified with an RBF-SVM using balanced class weights. **End-to-end fine-tuning:** a head (GAP → dropout 0.4 → sigmoid) is trained in two phases—phase 1 trains the head only (Adam, 10^{-3}, 5 epochs) and phase 2 unfreezes the top ∼40–50 layers and continues from a low learning rate (10^{-5}–3×10^{-5}) with early stopping on validation AUC. All batch-normalization layers are kept in inference mode, which stabilizes training with small batches. Training-data augmentation (rotation/flip) is applied on the fly.

4.4 Leakage-free Evaluation Protocol

The key to preventing leakage is that every train/test split is performed before augmentation and at the patient or imaging-unit level, never at the level of individual augmented patches. This guarantees that augmented copies derived from the same source image can never be separated across the split. Figure 3 contrasts the two orderings step by step. The exact splitting rule depends on the dataset. For CBIS-DDSM we split by patient using StratifiedGroupKFold (or GroupShuffleSplit) and use the dataset's official train/test split for the headline results. For Mini-MIAS, we split by image ID, keeping all rotated versions of an image within the same fold. Augmentation is applied only to the training folds, never to the held-out data. The external datasets, Mini-MIAS and INbreast, are never seen during training and are used solely for evaluation.

4.5 Statistical Evaluation

The primary metric is ROC-AUC, which is threshold-independent; accuracy, sensitivity (malignant recall) and specificity are additionally reported at the 0.5 operating point. Where no official split exists, we use five-fold cross-validation (patient-wise for CBIS-DDSM, image-wise for MIAS). We emphasize ROC-AUC as the primary endpoint precisely, because its threshold-independence allows comparison across datasets and operating points. Hyper-parameters are taken from the literature at standard default values; a leakage-safe search (with no use of the test set for selection) confirms that they are near-optimal (Sub-section 5.7).

Three distinct sources of uncertainty are separated, because conflating them is a common cause of overstated claims.

(i) Training variability is measured by repeating the entire two-phase procedure with ten random seeds, which govern both weight initialization of the classifier head and the patient-wise validation carve-out. We report the mean and standard deviation of the per-seed AUCs together with their range. Ten seeds replace the three used in the original submission; three runs bound the spread only very loosely.

(ii) Finite-sample uncertainty of the test set is quantified by percentile bootstrap over 2,000 re-samples. We state the resampling unit explicitly, since it materially changes the width of the interval. Patches drawn from one patient are not independent, so the headline intervals resample *clusters*—patients for CBIS-DDSM, source images for Mini-MIAS and INbreast—taking all patches belonging to each drawn cluster. Patch-level intervals are also reported for comparison and are, as expected, slightly narrower. All intervals are computed on seed-averaged predicted probabilities. We note that a bootstrap can only re-weight the cases actually collected; it cannot represent the between-cohort variability that a genuinely new external sample would exhibit, so for the two small external sets the reported intervals should be read as a lower bound on total uncertainty.

(iii) Differences between AUCs are assessed with DeLong's method [36] and not by inspecting whether an interval excludes chance, using the fast rank-based implementation of Sun and Xu [37]. For two models scored on the same cases-for example two backbones on the CBIS-DDSM test split-we use the paired DeLong test on correlated ROC curves. For two different case sets- for example internal CBIS-DDSM against external Mini-MIAS-the curves are independent and we combine the two DeLong variance estimates additively, $z = (A_1 - A_2)/\sqrt{\mathrm{var}_1 + \mathrm{var}_2}$. Only this comparison can support or refute a claim that external performance matches internal performance; a confidence interval lying

above 0.5 establishes merely that a model beats random guessing, which is a much weaker statement.

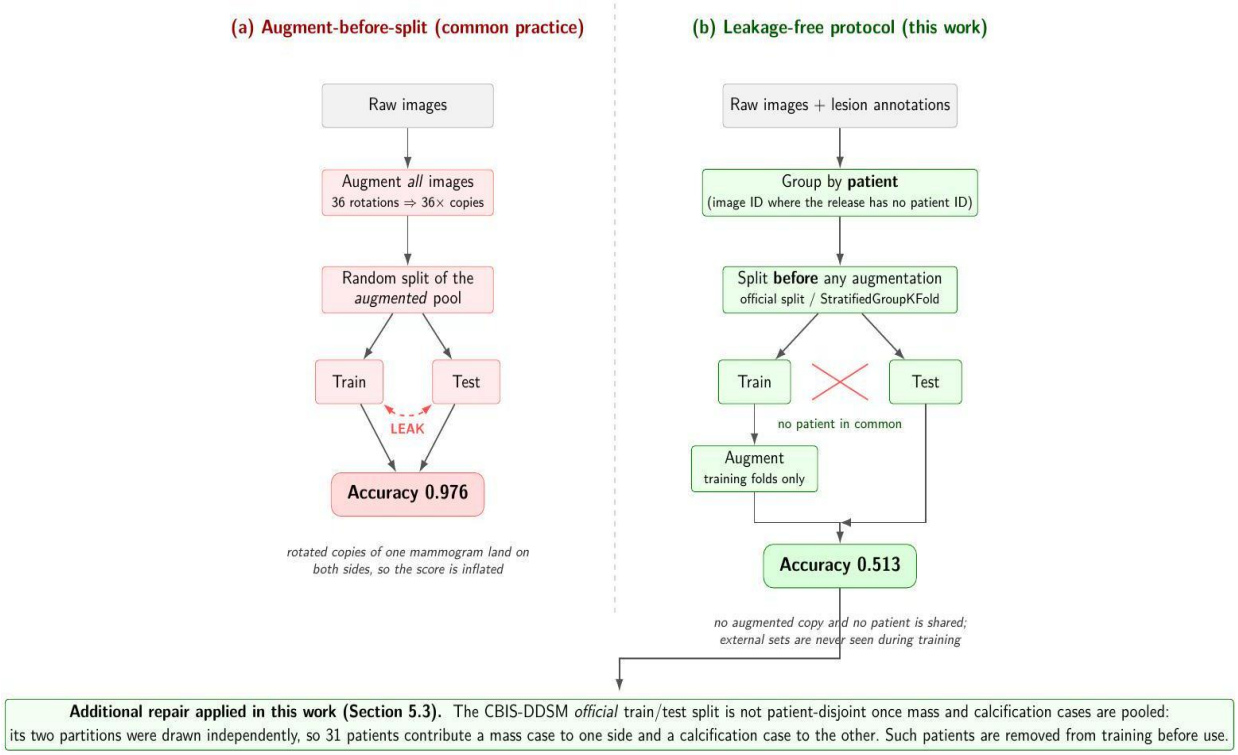


Figure 3. The leakage-free evaluation protocol, with the controlled whole-image Mini-MIAS experiment of Table 16 shown on each branch. (a) When augmentation precedes splitting, rotated copies of the same mammogram appear in both partitions and accuracy reaches 0.976. (b) Grouping by source film and splitting *before* augmenting, then augmenting only the training folds, removes the coupling; the identical pipeline then scores 0.513, which is chance. The lower panel records a further leakage channel, found during this revision, inside the CBIS-DDSM official split itself.

Calibration is reported separately from discrimination, using the Brier score, the expected calibration error over ten equal-width bins and reliability curves, following standard practice for assessing whether modern networks' confidences can be read as probabilities [38]. This distinction turns out to matter: Sub-section 5.9 shows that ranking transfers across domains while calibration does not.

4.6 Label-free Cross-domain Threshold Transfer

Our own results show that a model can rank lesions correctly on a new scanner while its probability scale drifts, so that a threshold tuned on the source institution no longer sits where it should sit. A deployed second reader would encounter exactly this failure mode and reporting AUC does not address it. We add a small procedure that repairs the operating point without requiring labelled data from the target site.

Let S_{src} be the scores the trained model assigns to the held-out *source* validation fold and let $t^{\star}$ be the threshold maximizing Youden's J [39] on that fold. Both are available before deployment. At the target site we observe only un-labelled scores S_{tgt}. We map each target score onto the source score distribution by rank,

$$\tilde{s} = F_{\text{src}}^{-1}\big(F_{\text{tgt}}(s)\big), s \in S_{\text{tgt}}, \tag{1}$$

where F_{tgt} is the empirical CDF of the target scores and F^{-1} the empirical quantile function of the source scores and then apply the unchanged source threshold $t^{\star}$ to $\tilde{s}$. Because the mapping in (1) is monotone, it leaves the ranking and therefore the ROC curve and AUC exactly unchanged; it moves only the operating point. The procedure needs no target labels, no retraining and no access to the source images—only the source score distribution, a few hundred numbers that can be shipped with the model.

Equation (1) implicitly assumes that source and target have comparable class prevalence, since matching quantiles matches the flagged fraction. Where that assumption fails—INbreast is 67% malignant against 41% for CBIS-DDSM—we also evaluate a prevalence-aware variant that rescales the flagged fraction by the ratio of target to source prevalence, which requires an estimate of the target prevalence, but still no case-

level labels. Both are compared in Sub-section 5.9 against a fixed 0.5 threshold and against an oracle that tunes Youden's *J* directly on the labelled target set, the latter serving as an upper bound that a deployed system could not attain.

5. RESULTS AND DISCUSSION

5.1 The Leakage Effect (Mini-MIAS)

The submitted version of this table compared an augment-before-split accuracy against leakage- free ROC-AUCs that had been produced by different scripts on a different task and asserted that "only the augmentation/split order differs". That was not accurate: between those rows, the label definition, the sample, the backbone and the rotation count all changed as well. We have replaced it with a properly controlled experiment.

Table 16 holds everything constant—the same 115 Mini-MIAS films, the same benign-*versus*-malignant labels, the same frozen MobileNetV2 features, the same RBF-SVM, the same 36 rotations — and changes exactly one thing: whether the split is taken before or after augmentation. Splitting the augmented pool at random gives accuracy 0.976 and ROC-AUC 0.997. Splitting by source film first and augmenting only the training side, with the identical pipeline, gives accuracy 0.513 and ROC-AUC 0.500—below the 0.556 majority-class rate and indistinguishable from chance. The entire apparent skill of the first configuration is an artefact of rotated copies of one film appearing on both sides of the split.

The mechanism is straightforward. With 36 rotations per film and a random split of the augmented pool, almost every test sample has a near-duplicate of itself in the training partition, so even a nearest-neighbor lookup would score highly. The effect is a specific instance of a phenomenon documented across machine-learning-based science [9], in which pre-processing or augmentation applied before partitioning couples the partitions and yields over-optimistic, irreproducible results. The safeguard is simple and frequently neglected: partition at the imaging-unit level, by image or preferably by patient, before any augmentation, is applied.

Under the leakage-free protocol, the whole-image task remains non-discriminative across configurations. Table 2 reports three further leakage-free variants on the related normal-*versus*-abnormal task — frozen features, end-to-end fine-tuning and 512 px inputs—none of which recovers any skill. We label these explicitly as a different task from Table 16, rather than tabulating them alongside the benign-*versus*-malignant numbers as though they were comparable.

Table 2. Leakage-free whole-image Mini-MIAS, *normal versus abnormal* (322 films). These represent a different task from Table 16 and are not comparable with it as numbers; they are reported to show that the absence of whole-image skill under a leakage-free protocol is not an artefact of one backbone or one input resolution. Majority-class accuracy for this task is 0.64.

Configuration	Metric	Score	Note
Frozen, 224 px	ROC-AUC	0.518 ± 0.083	chance
Fine-tuned, 224 px	ROC-AUC	0.471	chance
Frozen, 512 px	ROC-AUC	0.478 ± 0.023	chance

5.2 Lesion-patch Signal Is Real (Mini-MIAS)

Re-framing the problem at the lesion scale (image-wise, leakage-free, normal *versus* abnormal) recovers a well-separated discrimination (Figure 4A). The suspiciously high AUC of the naïve normal-patch sampling arises, because the normal patches are not necessarily lesion-sized or lesion-centred, whereas the abnormal patches are. When normal samples are resampled to match the size distribution of the abnormal patches and given a tissue-density-weighted location, the AUC drops by about 0.10 to 0.841 ± 0.026, which we term the *honest* AUC. The fact that a large margin above chance persists even under this stringent control indicates that the residual signal is genuinely pathological rather than an artefact of patch selection. The per-fold breakdown (Table 3) shows that discrimination is consistent across the five image-disjoint folds (AUC 0.806–0.868), with sensitivity for malignant/abnormal lesions (~0.67) being the limiting factor—as expected, given that only 119 abnormal lesions were available. This pattern, in which sensitivity rather than specificity is the bottleneck, recurs throughout our experiments and is discussed further below. We stress that this normal-*versus*-abnormal AUC (0.841) is a *different* task from the benign-*versus*- malignant

external validation reported in Sub-section 5.8 (MIAS AUC 0.734); the two numbers use different label sets and are not directly comparable.

Table 3. Mini-MIAS lesion-patch (size/density-matched), per-fold image-wise 5-fold CV.

Fold	Acc	AUC	Sens.	Spec.
1	0.803	0.868	0.680	0.878
2	0.846	0.864	0.739	0.905
3	0.779	0.828	0.615	0.881
4	0.828	0.837	0.565	0.976
5	0.794	0.806	0.727	0.829
Mean	**0.810**	**0.841**	**0.665**	**0.894**

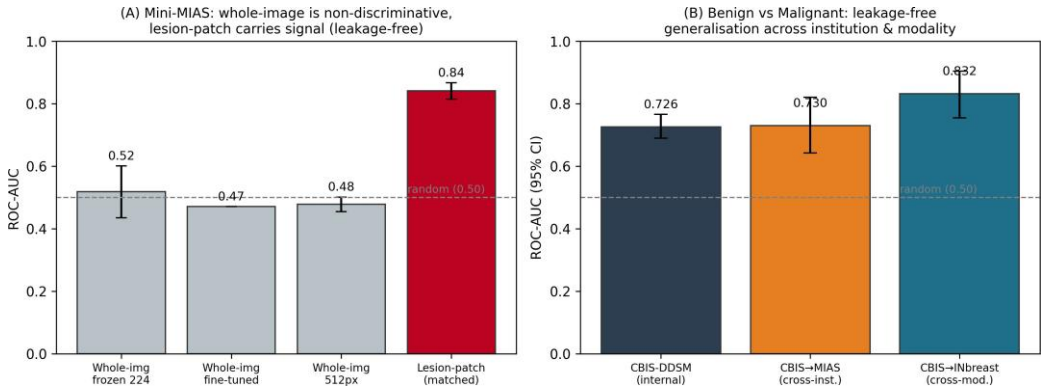


Figure 4. (A) On Mini-MIAS, whole-image classification is non-discriminative under a leakage-free protocol, whereas lesion patches carry signal. (B) Benign-vs-malignant: leakage-free generalization across institution (MIAS) and modality (INbreast), with 95% confidence intervals.

5.3 The CBIS-DDSM Official Split Is Not Patient-disjoint

Applying the patient-independence audit requested during review to our own protocol identified a violation, described below.

We are not the first to observe that this overlap exists. Zafari et al. [12], in the course of harmonizing several mammography datasets, noted patient overlap between the pre-defined CBIS-DDSM training and test sets and discarded the split for that reason. Our contribution here is not the observation, but what follows from it: the mechanism that produces the overlap, its size, its measured effect on a patch-based transfer-learning pipeline and a repaired split that keeps the official test set usable rather than abandoning it.

The submitted manuscript described the CBIS-DDSM official train/test partition as patient- disjoint. For the way the present study and much of the literature, uses it, that description does not hold. The mass and calcification sub-sets were partitioned independently of one another. Each is internally clean: we found zero patients shared between train and test within the mass sub-set and zero within the calcification sub-set. However, 31 patients have both a mass lesion and a calcification lesion and for those patients, the two draws disagreed, placing one lesion type in training and the other in test. Pooling mass and calcification cases into a single benign-*versus*-malignant task, as we and many others do, reintroduces the patient-level dependence that the official split appears to guarantee.

The scale is modest, but not negligible. Sixty of the 704 official test lesions (8.5%) belong to these 31 patients. In 19 of the 31 cases, the shared patient contributes the same breast and the same view to both sides, so the model may encounter the same breast tissue, with the same parenchymal pattern and acquisition characteristics, in training and at test time. The lesions themselves are distinct, as are the image files: we verified that no cropped ROI, no full mammogram and no pixel array is duplicated across the split. The dependence lies at the patient and tissue levels and is not a duplicated-sample error.

We quantify the consequence by retraining the identical model on a repaired split. The repair keeps the official test set untouched, so results remain comparable with published CBIS-DDSM numbers, instead removing every affected patient from training: 83 training lesions are dropped, leaving 2,780 and the residual train/test

patient overlap is zero. Table 10 reports both arms over ten seeds.

The measured effect is negligible. Internal AUC moves by +0.001 (from 0.745 to 0.746) and a paired DeLong test on the shared 704 test lesions does not distinguish the two models ($p = 0.68$). The external results are likewise unchanged and the finding that external performance matches internal performance holds in both arms. We state the null result plainly, since a leak of this kind might otherwise be assumed to have inflated the reported figures; on this dataset, with 8.5% of test lesions affected and the shared material being one breast and not one lesion, it has not.

Three observations nonetheless follow. First, the contamination is a property of a widely used public benchmark and not of our pipeline and its magnitude in another study, for example one pooling differently, using a smaller test set or fitting a higher-capacity model, is not bounded by the value measured here. Second, any study pooling mass and calcification cases from CBIS- DDSM should apply this repair or split by patient directly, since the official split alone does not provide the guarantee its documentation implies. Third, the contamination entered through a dataset convention, not a modelling error and is detectable only by generating and auditing split manifests. This supports treating split manifests and not only source code, as a standard released artefact. That it has been noted independently by Zafari et al. [12] while harmonizing the same benchmark reinforces rather than weakens the point: a property of a widely used dataset went unremarked in the classification literature that relies on it and was surfaced twice, in parallel, only by groups that inspected the split directly.

5.4 CBIS-DDSM Benign *vs.* Malignant: Ablation

For the larger, biopsy-proven CBIS-DDSM, we ablate the pipeline one component at a time on the official test split (Table 4). The only meaningful gains come from fine-tuning (+0.026) and a stronger backbone (+0.027); increasing the input resolution yields no improvement (-0.011), while test-time augmentation recovers part of the loss (+0.013). A leakage-safe hyper-parameter analysis (Sub-section 5.7) confirms that these defaults are near-optimal and the per-dataset confusion matrices are shown in Figure 5. The submitted version continued here with a claim that performance saturates at 0.72–0.73 and that this indicates a task/data ceiling. That sentence has been removed: as the caveats below and the factorial sweep of Sub-section 5.5 show, the single-run ablation could not support it.

Two caveats on how this table should be read. First, each row is a single training run, whereas the headline results of Sub-section 5.8 average ten. Single runs are noisy at this scale: the EfficientNetB0 @224 entry here (0.724) sits at the bottom of the ten-seed range subsequently measured for the same configuration (0.745, range 0.723–0.753), so the two are consistent, but no ordering among rows separated by less than about 0.02 should be trusted. Second, backbone and input resolution change together between the last three rows, so this table cannot attribute the difference to either. Both problems are addressed in Sub-section 5.5, which repeats the comparison with a proper factorial design over multiple seeds and it is that table, not this one, on which we base any claim about a performance ceiling.

Table 4. CBIS-DDSM component ablation (official test split). Single training run per row; see the caveats in the text and the multi-seed factorial comparison in Table 13.

Backbone	Training	Input/TTA Test	AUC
MobileNetV2	Frozen + SVM	224	0.671
MobileNetV2	Fine-tuned	224	0.697
EfficientNetB0	Fine-tuned	224	0.724
EfficientNetB3	Fine-tuned	300	0.713
EfficientNetB3	Fine-tuned	300 + TTA	**0.726**

5.5 Does a Heavier Backbone or a Larger Input Help?

The ablation above varies backbone and input resolution together, leaving the two confounded and all three of its backbones belong to one efficiency-oriented family. Neither is a sound basis for claiming a task/data ceiling, so the experiment was repeated with a design that separates them.

Eleven configurations were trained under the identical leakage-free protocol — same patient-wise splits, same two-phase schedule, same augmentation, same early stopping—spanning classical CNNs (ResNet-

50, DenseNet-121), a modernized CNN (ConvNeXt-T), transformers (ViT-B/16, Swin-T) and the EfficientNet family, from 2.2-86 M parameters. To keep the comparison internally consistent across families, this study runs in a single PyTorch implementation instead of mixing frameworks; EfficientNet-B0 @224 is included in it as a cross-framework anchor against the TensorFlow result of Table 7. EfficientNet-B0, EfficientNet-B3 and ResNet-50 are each run at 224 and 300 px, which gives the clean resolution ablation with the backbone held fixed. Table 13 reports the outcome and Table 11 the paired DeLong tests against the EfficientNet-B0 reference on the shared test cases.

Two results follow, the first of which corrects the submitted manuscript. With the backbone held fixed, resolution does improve performance: EfficientNet-B0 gains +0.036 AUC from 224 to 300 px ($p = 0.003$, paired DeLong) and EfficientNet-B3 gains +0.022. The original ablation reported that resolution brought no benefit (-0.011), but it varied backbone and resolution together; once the two factors are separated, the sign of the effect reverses. The factorial design requested during review is what made this visible.

Second, the performance ceiling is present, but not flat and the claim has been weakened to match. Across 11 configurations spanning 2.2–86 M parameters, CBIS-DDSM test AUC ranges from 0.707 to 0.777. The strongest configuration is ViT-B/16 @224, the advantage of which over the lightweight EfficientNet-B0 baseline is statistically significant (+0.060, $p = 0.0002$); ConvNeXt-T and the higher-resolution EfficientNets are also significantly ahead, while ResNet-50, DenseNet-121, Swin-T and MobileNetV2 are not distinguishable from it (Table 11). Capacity thus retains some influence, in that an 86 M-parameter transformer extracts more than a 4 M-parameter CNN. What the sweep establishes is the scale of the remaining headroom: the full set of architectures, classical and modern, convolutional and attention-based, spans 0.07 AUC and none approaches the 0.94–0.99 figures that leakage-free evaluation dissolves. The difference between a well-evaluated and a leaky protocol is an order of magnitude larger than the difference between the best and the worst architectures.

The ordering also fails to transfer. Swin-T is the strongest model on Mini-MIAS (0.775) while ranking mid-table internally and EfficientNet-B3 is among the weakest externally despite being competitive internally. Selecting an architecture on internal performance alone would not have identified the configuration that generalizes best, which constitutes a further argument for routine external validation.

The anchor proved necessary. An initial PyTorch implementation reproduced it approximately 0.12 AUC too low, with external transfer falling to near chance. The cause was batch normalization, not the architecture: the TensorFlow pipeline holds every BN layer in inference mode, whereas PyTorch by default continues to update BN running statistics during training, which on batches of 16 to 32 grayscale patches is sufficient to destabilize a 10^{-5} fine-tuning stage. In the absence of the anchor, the resulting table would have measured BN conventions while being presented as a comparison of backbones. We record this, because it illustrates the same point the rest of the paper makes: on small datasets, an implementation detail can outweigh the model choice under study.

5.6 Whole Mammogram *versus* Lesion Patch on the Primary Dataset

The submitted manuscript concluded that lesion patches beat whole-image analysis, but demonstrated it only on Mini-MIAS, the whole-image task of which collapses to chance once leakage is removed. That is a weak basis for a general claim, so we added the missing baseline on CBIS-DDSM itself.
Each full mammogram is labelled malignant if any lesion it contains is malignant, giving 2,857 images (2,458 train / 399 test) under the same patient-wise protocol. We train at 224 px and again at 512 px, because a lesion occupying a few hundred pixels in a 3000×5000 mammogram may simply be destroyed by aggressive down-sampling; if the patch advantage were only a resolution artefact, the 512 px model should close much of the gap. Table 14 reports the comparison against the patch pipeline trained and tested on the same patients.

The patch framing performs better, but the resolution effect is substantial. At 224 px the whole- image model reaches AUC 0.638±0.024, more than 0.10 below the patch pipeline; at 512 px it recovers to 0.705±0.014, closing approximately two-thirds of that gap while remaining about 0.04 short. Part of what the submitted manuscript attributed to framing is thus attributable to resolution: reducing a 3000×5000 mammogram to 224 px removes much of the lesion, which is a more accurate account than the claim that whole images carry no signal. The residual gap at 512 px represents the framing effect properly. It is consistent in direction with the Mini-MIAS result while being considerably smaller, as would be expected

on a dataset an order of magnitude larger.

Two differences make this a comparison of framings and not a strictly controlled experiment. The whole-image task carries an image-level label (malignant if any lesion is malignant) over 399 test images, against 704 lesion-level labels for the patch task and a whole-image model must localize as well as classify. We draw a limited conclusion: at the data scale available here, the patch framing extracts more signal. Nothing in these results implies that whole-image mammography classification is un-learnable.

5.7 Hyper-parameter Sensitivity (Leakage-safe)

Tuning is itself prone to optimistic bias, particularly on small datasets, so our main configurations use the literature-standard defaults, with only the epoch count set by early stopping. To verify that these defaults are not materially sub-optimal, we ran a search that never used the official test set for selection. For the frozen-feature SVM (MobileNetV2 features, as in Table 4), a grid search over C and γ was performed with patient-wise five-fold cross-validation (the test set excluded throughout); the best configuration ($C = 1$, $\gamma = 10^{-4}$) attained a CV-AUC of 0.721—0.008 above the default—and a test AUC of 0.679 (Table 5). For fine-tuning, an unfreeze-depth ×learning-rate search selected a 3×10^{-5} configuration, the test AUC of which was 0.733 being only +0.009 over the default (Table 6), with the test set used for reference only. Tuning produced no meaningful test-set gains for either classifier and the ∼0.72–0.73 ceiling is unchanged, confirming that the reported results do not depend on a lucky choice of defaults.

5.8 Cross-dataset External Validation

The transferred model is **EfficientNetB0 @224**, fine-tuned on the CBIS-DDSM official training split and applied to both external sets without any retraining. We state this explicitly, because the submitted version described it as "the best configuration of Table 4", which was an error: that row is EfficientNetB3 @300 with TTA, a different model that happened to reach an internal AUC agreeing with EfficientNetB0's to three decimal places. EfficientNetB0 is also the more appropriate choice for a transfer experiment, since test-time augmentation would confound domain transfer with an inference-time ensembling effect.

Table 5. Leakage-safe SVM grid (patient-wise CV-AUC on the training split; top rows).

C	γ	CV-AUC
1	10^{-4}	**0.721**
1	scale	0.712
1	10^{-3}	0.708
10	10^{-4}	0.708
0.1	10^{-4}	0.706

Table 6. Leakage-safe fine-tuning sweep (validation-selected; test for reference only).

Unfrozen	Phase-2 LR	Val. AUC	Test AUC
30	10^{-5}	0.772	0.710
30	3×10^{-5}	0.779	0.722
50	10^{-5}	0.779	0.723
50	3×10^{-5}	**0.781**	**0.733**

Table 7 reports the results over ten seeds and Table 11 the significance tests. Three points stand out.

5.9 Calibration and Repairing the Operating Point

Table 9 and Figure 6 separate the two things an AUC hides. Discrimination is preserved on transfer, but the expected calibration error roughly doubles on Mini-MIAS and the reliability curves depart from the diagonal in *opposite* directions on the two external sets. On Mini-MIAS, the curve lies below the diagonal: the model is *over*-confident, assigning malignancy probabilities higher than the observed malignant fraction, which is what drives the specificity loss at a fixed 0.5 threshold. On INbreast, it lies above the diagonal—the model is *under*-confident—which is the signature of a cohort the prevalence of which (67% malignant) is far higher than the source's (39%) rather than of a model that has degraded. That the two external sets fail in opposite directions is itself informative and it is why a single blanket recalibration rule

is not appropriate. A two-sample Kolmogorov–Smirnov test against the source validation scores confirms that the score *distribution* shifts in both cases; usefully for deployment, this statistic needs no target labels, so a site can measure its own drift before deciding whether to trust an inherited threshold.

Table 7. Leakage-free external validation. A single EfficientNetB0 @224 trained on the CBIS-DDSM; official training split is applied without retraining to both external sets. AUC is given as the mean±standard deviation over ten seeds, alongside the observed range; the 95% CI is a cluster bootstrap (2,000 resamples of patients for CBIS-DDSM, of source images for the external sets) on the seed-averaged predictions. Sensitivity and specificity are at the 0.5 operating point.

Test set	n	Clusters	AUC (10 seeds)	Range	95% CI	Sens.	Spec.
CBIS-DDSM (internal test)	704	349	0.745±0.008	0.723–0.753	0.703–0.794	0.70	0.65
CBIS→MIAS	119	111	0.734±0.011	0.713–0.750	0.640–0.826	0.78	0.51
CBIS→INbreast	132	107	0.877±0.009	0.858–0.891	0.804–0.946	0.89	0.74

Table 8. Confusion matrices computed directly from the saved per-sample predictions (seed-averaged, threshold 0.5), replacing the matrices reconstructed from summary statistics in the submitted version. FN is the clinically costly error.

Test set	TP	FN	TN	FP	Sens.	Spec.
CBIS-DDSM (internal test)	194	82	277	151	0.70	0.65
CBIS→MIAS	40	11	35	33	0.78	0.51
CBIS→INbreast	79	10	32	11	0.89	0.74

Table 9. Discrimination transfers, but calibration does not. All quantities are computed on the seed-averaged predicted probabilities, so the AUC column is the seed-ensembled value and differs marginally from the seed-mean of Table 7. AUC is essentially unchanged from internal to external data, whereas the Brier score and expected calibration error degrade and the score distribution shifts away from the source (two-sample Kolmogorov–Smirnov statistics against the source validation scores, computable without any target labels).

Test set	AUC (ens.)	Brier	ECE	KS *vs.* source
CBIS-DDSM (internal test)	0.749	0.206	0.104	0.084
CBIS→MIAS	0.737	0.218	0.153	0.210
CBIS→INbreast	0.882	0.131	0.085	0.293

First, discrimination transfers. The external AUC on Mini-MIAS is statistically indistinguishable from the internal CBIS-DDSM AUC, the DeLong test on independent samples giving $p = 0.82$, so equality cannot be rejected. The submitted manuscript instead argued that the external confidence intervals lay above 0.5 and so demonstrated "statistically significant generalization", which conflated beating chance with matching internal performance. That statement has been withdrawn. The INbreast AUC is significantly *higher* than internal ($p < 0.001$), a property of its labelling and not of the model (Sub-section 5.10).

Table 10. Effect of repairing the CBIS-DDSM official split. Pooling mass and calcification cases makes the official partition non-patient-disjoint: 31 patients appear on both sides, affecting 60 of the 704 test lesions. The repair removes those patients from training (83 lesions) and leaves the official test set untouched, so the two columns are evaluated on identical cases. AUC is the mean over ten seeds.

Test set	Official split	Patient-disjoint	Δ
CBIS-DDSM (internal test)	0.745 ± 0.008	0.746 ± 0.004	+0.001
CBIS→MIAS	0.734 ± 0.011	0.731 ± 0.009	-0.003
CBIS→INbreast	0.877 ± 0.009	0.871 ± 0.008	-0.006

Second, the result is stable across training runs: the standard deviation over ten seeds is ≤ 0.011 on every

test set, with the full range spanning less than 0.03 AUC. We report ten seeds in place of three, because three runs estimate this spread poorly.

Third—and this is the finding with the clearest deployment consequence — discrimination transfers, but calibration does not. At a fixed 0.5 threshold, the ranking is intact while specificity on Mini-MIAS degrades, because the whole score distribution drifts upward relative to the source (Sub-section 5.9). A model shipped with its source threshold would therefore over-flag at a new site while its AUC looked healthy.

Table 11. DeLong tests. Internal and external AUCs are computed on different cases, so their ROC curves are independent and the two DeLong variances are combined additively; comparisons on shared cases use the paired test. The internal-*versus*-MIAS comparison is the one that supports the generalization claim: the difference is not significant; i.e., external performance is statistically indistinguishable from internal performance, which is a stronger statement than a confidence interval merely excluding chance.

Comparison	Type	AUC A	AUC B	z	p
CBIS internal *vs.* MIAS	unpaired	0.749	0.737	0.23	0.818
CBIS internal *vs.* INbreast	unpaired	0.749	0.882	-3.67	<0.001
MIAS *vs.* INbreast	unpaired	0.737	0.882	-2.60	0.009
Official *vs.* patient-disjoint (CBIS internal)	paired	0.749	0.750	-0.41	0.681
Official *vs.* patient-disjoint (MIAS)	paired	0.737	0.733	0.74	0.462
Official *vs.* patient-disjoint (INbreast)	paired	0.882	0.878	0.64	0.522

Table 12 then asks what a deployer should actually do about it, comparing four strategies against an oracle that tunes Youden's J on the labelled target set and is therefore unattainable in practice.

Table 12. Choosing the operating point at a new institution. The oracle tunes Youden's J on the labelled target set and is therefore unattainable in deployment; it is shown as an upper bound. Carrying the source threshold across unchanged is the naive baseline. Quantile transfer and its prevalence-matched variant use only un-labelled target scores (plus, for the latter, an estimate of target prevalence). Best attainable J per dataset in bold.

Threshold strategy	Thr.	Sens.	Spec.	Youden J	Needs labels?
CBIS-DDSM (internal test)					
fixed 0.5 (as submitted)	0.500	0.703	0.647	0.350	no
source Youden J, carried across	0.504	0.692	0.652	0.344	no
quantile transfer	0.504	0.670	0.680	0.350	no
prevalence-matched transfer	0.533	0.659	0.699	**0.358**	no
oracle target Youden J	0.563	0.623	0.762	0.385	yes
CBIS→MIAS					
fixed 0.5 (as submitted)	0.500	0.784	0.515	0.299	no
source Youden J, carried across	0.504	0.765	0.529	0.294	no
quantile transfer	0.504	0.686	0.706	**0.392**	no
prevalence-matched transfer	0.546	0.686	0.706	**0.392**	no
oracle target Youden J	0.546	0.686	0.706	0.392	yes
CBIS→INbreast					
fixed 0.5 (as submitted)	0.500	0.888	0.744	**0.632**	no
source Youden J, carried across	0.504	0.888	0.744	**0.632**	no
quantile transfer	0.504	0.640	0.907	0.547	no
prevalence-matched transfer	0.608	0.775	0.837	0.612	no
oracle target Youden J	0.532	0.888	0.791	0.678	yes

Carrying the source-validated Youden threshold across unchanged, which is what the submitted manuscript

implicitly recommended, performs no better on Mini-MIAS than taking no action at all: Youden's J is 0.294 against 0.299 for a fixed 0.5 threshold. Quantile transfer (Equation 1) raises J to 0.392, equal to the value attained by the oracle (0.392) and the prevalence-matched variant yields the same operating point. Specificity on Mini-MIAS rises from 0.51 to 0.71 for a modest loss of sensitivity, so the specificity collapse that motivated the analysis is largely eliminated without annotating any case at the target site. Mini-MIAS is the like-for-like external cohort and this is the principal result of the section.

One negative result is reported alongside it. On INbreast, the plain 0.5 threshold is already close to the oracle and quantile transfer degrades it. The reason is visible in Table 9: the INbreast score distribution is shifted principally, because its prevalence differs (67% malignant against 39% in the source) rather than because the model is miscalibrated on that cohort. Rank matching forces the target to flag the same fraction of cases as the source, which is the wrong correction under prevalence shift; the prevalence-matched variant recovers most, though not all, of the loss. The resulting practical rule is narrow. The operating point should be transferred when the score distribution has drifted and prevalence is comparable and the prevalence-matched form should be preferred wherever a prevalence estimate is available. We make no claim that the procedure is safe to apply without checking these conditions.

Table 13. Architecture and resolution under one identical leakage-free protocol, run in a single PyTorch implementation so the families are directly comparable. EfficientNet-B0/B3 and ResNet-50 appear at 224 and 300 px, so resolution is varied with the backbone held fixed. AUC is the mean over three seeds on the CBIS-DDSM official test split; external AUCs are for the same models applied without retraining.

Backbone	Params (M)	Input	CBIS test AUC	MIAS	INbreast
ViT-B/16	85.8	224	0.777 ± 0.012	0.736	0.838
ConvNeXt-T	27.8	224	0.749 ± 0.005	0.719	0.839
EfficientNet-B3	10.7	300	0.747 ± 0.008	0.697	0.727
Swin-T	27.5	224	0.742 ± 0.006	0.775	0.837
EfficientNet-B0	4.0	300	0.742 ± 0.009	0.684	0.811
ResNet-50	23.5	224	0.738 ± 0.000	0.720	0.836
ResNet-50	23.5	300	0.731 ± 0.007	0.706	0.836
DenseNet-121	7.0	224	0.725 ± 0.005	0.726	0.841
EfficientNet-B3	10.7	224	0.725 ± 0.007	0.689	0.760
MobileNetV2	2.2	224	0.722 ± 0.020	0.688	0.806
EfficientNet-B0	4.0	224	0.707 ± 0.001	0.738	0.818

Table 14. Whole-mammogram *versus* lesion-patch classification on CBIS-DDSM, the primary dataset, under the same patient-wise leakage-free protocol. Two input resolutions are used, so that the gap cannot be dismissed as an artefact of down-sampling a full mammogram until the lesion disappears. Mean over three seeds.

Framing	Input	n (train/test)	CBIS test AUC
Whole mammogram	224 px	2458 / 399	0.638 ± 0.024
Whole mammogram	512 px	2458 / 399	0.705 ± 0.014
Lesion patch	224 px	2863 / 704	0.745

Table 15. Sensitivity to imperfect lesion localization. Each crop is regenerated from the source image with a random translation, expressed as a fraction of the patch side or with the box scaled; the model is not retrained. Because the crops are regenerated rather than taken from the official pre-cropped ROIs, the top row is the correct baseline for reading the degradation. Mean over three random jitter draws and a three-seed model ensemble.

Translation	Scale	CBIS test	MIAS	INbreast
0%	1.00	0.756	0.734	0.887
5%	1.00	0.755	0.757	0.875
10%	1.00	0.735	0.758	0.878
15%	1.00	0.732	0.715	0.864
20%	1.00	0.722	0.719	0.845
0%	0.85	0.750	0.751	0.875
0%	1.15	0.744	0.760	0.861

Table 16. The leakage effect, measured with a single variable changed. Both arms use the same 115 Mini-MIAS films, the same benign-*versus*-malignant labels, the same frozen MobileNetV2 features, the same RBF-SVM and the same 36 rotations; they differ only in whether the split is taken before or after augmentation. Mean ± standard deviation over 5-fold cross-validation. Majority-class accuracy for this task is 0.556.

Protocol	Split unit	Accuracy	ROC-AUC
Augment → split	augmented sample	0.976 ± 0.006	0.997
Split → augment train only	source film	0.513 ± 0.051	0.500 ± 0.072

Table 17. Quantitative Grad-CAM localization against the radiologist annotation, with two null models scored on the identical patches. Crops are deliberately off-centred and the mask is cropped through the same transform. Pointing is the fraction of patches the peak activation of which falls inside the annotation energy; the fraction of activation mass inside it; conc. the energy divided by the lesion's area fraction, which equals 1 for a uniform map and is therefore comparable across regimes. In the native crop convention the lesion fills too much of the patch for the test to discriminate and the centre prior wins; the wide-field crops, where the lesion covers about a tenth of the patch, represent the informative comparison.

Setting	Saliency	Pointing	Energy	Conc.	IoU@top20	Lesion area
CBIS-DDSM, native crop	Grad-CAM	0.375	0.276	1.27	0.226	0.218
	centre prior	0.876	0.448	2.05	-	
	random	0.084	0.218	1.00	-	
Mini-MIAS, native crop	Grad-CAM	0.639	0.549	1.20	0.214	0.459
	centre prior	0.899	0.686	1.50	-	
	random	0.252	0.459	1.00	-	
CBIS-DDSM, wide field	Grad-CAM	0.272	0.107	1.72	0.123	0.062
	centre prior	0.029	0.105	1.69	-	
	random	0.024	0.062	1.00	-	
Mini-MIAS, wide field	Grad-CAM	0.513	0.295	1.89	0.294	0.156
	centre prior	0.202	0.253	1.62	-	
	random	0.042	0.156	1.00	-	

5.10 Why the INbreast Result Must Be Read Cautiously?

The INbreast AUC is the highest of the three, but this warrants caution: its BI-RADS 1–2 cases are unambiguously normal/benign, making the separation easier than in the biopsy-confirmed CBIS-DDSM. The fair, like-for-like external comparison is therefore CBIS→MIAS; the higher INbreast figure should not be read as one modality being inherently more accurate.

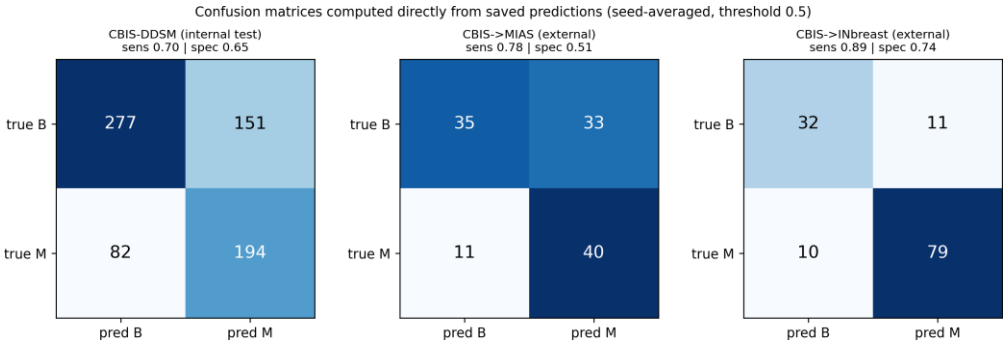


Figure 5. Confusion matrices at the 0.5 operating point, computed directly from the saved per-sample predictions. In the submitted version, these counts were reconstructed from the reported sensitivity and specificity; they are now generated from the prediction files released with the paper. Note the specificity drop on MIAS, caused by un-calibrated cross-domain thresholds rather than by any loss of ranking ability.

Reviewers rightly pressed us to quantify and not merely acknowledge this, so we report the exact BI-RADS composition. Of the 132 INbreast mass patches, 89 fall in the malignant class (BI-RADS 4a–6: 7, 2, 12, 56 and 12 patches for 4a, 4b, 4c, 5 and 6, respectively) and 43 in the benign class. That benign class is *not* homogeneous: 30 patches are BI-RADS 2, genuinely benign, but 13—**30.2% of the benign class**—are

BI-RADS 3, “probably benign”. BI-RADS 3 denotes a lesion with a ≲2% malignancy risk that warrants short-interval follow-up rather than biopsy; it is an assessment of *management*, not a histological finding and a minority of such lesions do prove malignant.

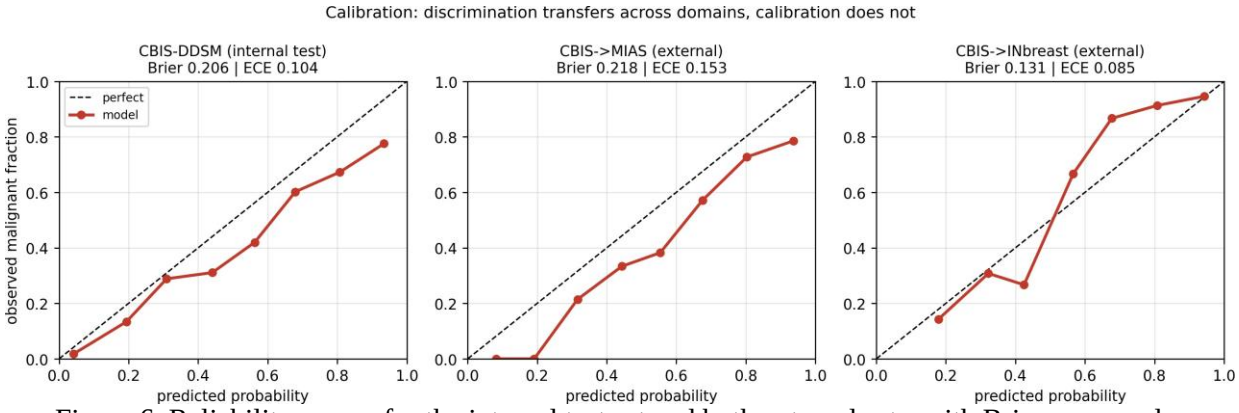


Figure 6. Reliability curves for the internal test set and both external sets, with Brier score and expected calibration error. Points below the diagonal indicate over-confidence. Discrimination is largely preserved across the domain change (Table 7) while calibration is not, which is why a threshold tuned at the source institution misbehaves at the target.

This biases the comparison in two ways, both of which inflate INbreast. First, assignment to BI-RADS 2 *versus* 4–6 is driven by the same morphological features the classifier reads; namely margin, shape and speculation, so the label is partly a function of image appearance rather than an independent ground truth. The task is consequently closer to predicting a radiologist’s assessment than to predicting histology, whereas CBIS-DDSM labels derive from biopsy and are independent of appearance. Second, the genuinely ambiguous cases that a biopsy-confirmed cohort contains, in particular benign lesions with malignant appearance, are largely absent, since such lesions would have been assessed BI-RADS 4 and hence assigned to the malignant class. The 13 BI-RADS 3 patches are the only substantial source of label noise acting in the opposite direction. We therefore treat the INbreast figure as evidence of robustness to a change of acquisition technology rather than as a measure of diagnostic accuracy and base all generalization claims on CBIS→MIAS.

A more principled treatment would not force an ordinal scale into two classes at all. BI-RADS is ordered and the distance from 2 to 3 is not the distance from 3 to 4; ordinal regression or soft labels that encode the published malignancy probability of each category (roughly < 2% for BI-RADS 3 against > 95% for BI-RADS 5), would let a model exploit that structure and would make the uncertainty explicit rather than hiding it in a thresholding decision. Vision–language foundation models pre-trained on report text are well placed to do this, since the reports they are trained on contain the assessment category and the hedging language that accompanies it. We leave this to future work, but note that it is the natural way to use BI-RADS-labelled cohorts as external validation without importing their labelling convention as if it were ground truth.

5.11 How Much Localization Error Can the Classifier Absorb?

Every result so far assumes the lesion’s centre or bounding box is known exactly, because it is read from the dataset’s annotations. A deployed pipeline would receive that box from a detector, with error. To estimate what this costs, we regenerate each crop from the source image with a random translation—5, 10, 15 or 20% of the patch side, in a random direction—and separately with the box scaled by 0.85 or 1.15 to mimic a detector that crops too tightly or too loosely and re-evaluate the trained model without any retraining. Table 15 reports the results on all three datasets.

Because these crops are regenerated rather than taken from the official pre-cropped ROIs, the zero-shift row rather than the headline AUC is the correct baseline for reading the degradation; we would otherwise be attributing the difference between two crop conventions to localization noise.

Degradation is gradual rather than abrupt. A displacement of 15% of the patch side, which corresponds to a realistic detector error, costs 0.024 AUC on CBIS-DDSM, 0.019 on Mini-MIAS and 0.022 on INbreast, each comparable to the seed-to-seed spread of the model itself. Displacements up to 5% have negligible

effect and at 20%, the loss remains within approximately 0.03. Scaling the box by ±15% has less effect than translating it. For an integrator, this indicates that the classifier does not require pixel-accurate boxes and that a detector with typical localization error would not be the binding constraint. The experiment measures only a displaced box on a true lesion; a full detection-plus-classification evaluation would additionally have to account for regions proposed on normal tissue and would yield a lower end-to-end figure.

5.12 Interpretability

We compute Grad-CAM [26] attention maps from the fine-tuned EfficientNet on lesion patches (Figure 7). The submitted version used these panels to *confirm* that predictions are driven by the lesion; that was more than the evidence supported, so we have both softened the claim and measured it.

Measuring it requires care. If each patch is cropped centred on its lesion, then "the peak activation lies inside the annotation" and is satisfied by any model with a centre bias and the metric says nothing. We therefore off-centre every crop and crop the annotation mask through the identical transform, so the lesion sits at a known but non-central position. Against that, we score three quantities—the pointing-game hit rate, the fraction of Grad-CAM energy inside the annotation and the IoU of the top-20% activation with it—and, on the very same patches, two null models: a fixed Gaussian centre prior and a smoothed random map. We also report a chance-normalized concentration, the energy inside the mask divided by the mask's area fraction, which equals 1 for a uniform map regardless of lesion size and so is comparable across regimes.

This precaution proves insufficient on its own. In the crop convention on which the model is trained, the lesion occupies between a fifth and a half of the patch, so it cannot be displaced far enough to leave the central region. The centre prior consequently outscores Grad-CAM, by 0.88 against 0.37 on CBIS-DDSM and 0.90 against 0.64 on Mini-MIAS (Table 17, "native" rows). In this regime, the pointing game provides evidence about the cropping convention rather than about the model, an observation that applies to a substantial body of published saliency evaluation on pre-cropped lesion patches. We therefore repeated the analysis on wide-field crops of three times the lesion box, in which the lesion covers approximately a tenth of the patch and can be placed genuinely off-centre. The ordering then reverses: the Grad-CAM peak falls inside the annotation for 0.27 of CBIS-DDSM patches, against 0.03 for the centre prior and 0.02 for a random map and for 0.51 of Mini-MIAS patches against 0.20 and 0.04. These correspond to approximately nine and two-and-a-half times the centre prior respectively and an order of magnitude above chance.

We draw a bounded conclusion. The maps are consistent with lesion-driven prediction and localize the annotated abnormality considerably better than chance, but on their own, they do not establish that the model reasons from lesion morphology and saliency scores computed on tightly cropped lesion patches should not be read as though they did.

5.13 Comparison with Prior Work

Direct numerical comparison is confounded by differing tasks, splits and (often unreported) leakage control (Table 18). A clear trend nonetheless emerges: the largest headline figures come from the settings most susceptible to optimistic bias. Both report very high single-dataset accuracy without an explicit augment-before-split control—99.8% by Adedigba et al. [20] on a small mammogram set using a discriminatively fine-tuned DenseNet and 98.87% by Salama and Aly [40] using a segmentation-plus-CNN pipeline on DDSM. Methods closer to our own—feature-fusion or deep-feature-plus-classifier approaches—report similarly high single-dataset accuracy, but are likewise never externally validated: Jafari and Karami [41] reach 94.5% on MIAS with an ensemble of CNNs, Ragab et al. [6] obtain 87.2% (AUC 0.94) on DDSM/CBIS-DDSM using AlexNet features with an SVM and Yu et al. [14] fuse deep features with an SVM for an AUC of 0.93 on DDSM. By contrast, our leakage-free CBIS-DDSM AUC (0.745) is close to the only directly comparable patch-based single-dataset CBIS result (0.722) [18] and is unique among these works in providing patient-wise splitting, two-dataset external validation, clustered bootstrap confidence intervals and released per-sample predictions. We therefore present our numbers not as state-of-the-art accuracy, but as a reliable benchmark that most prior studies—lacking leakage control and external validation—cannot offer.

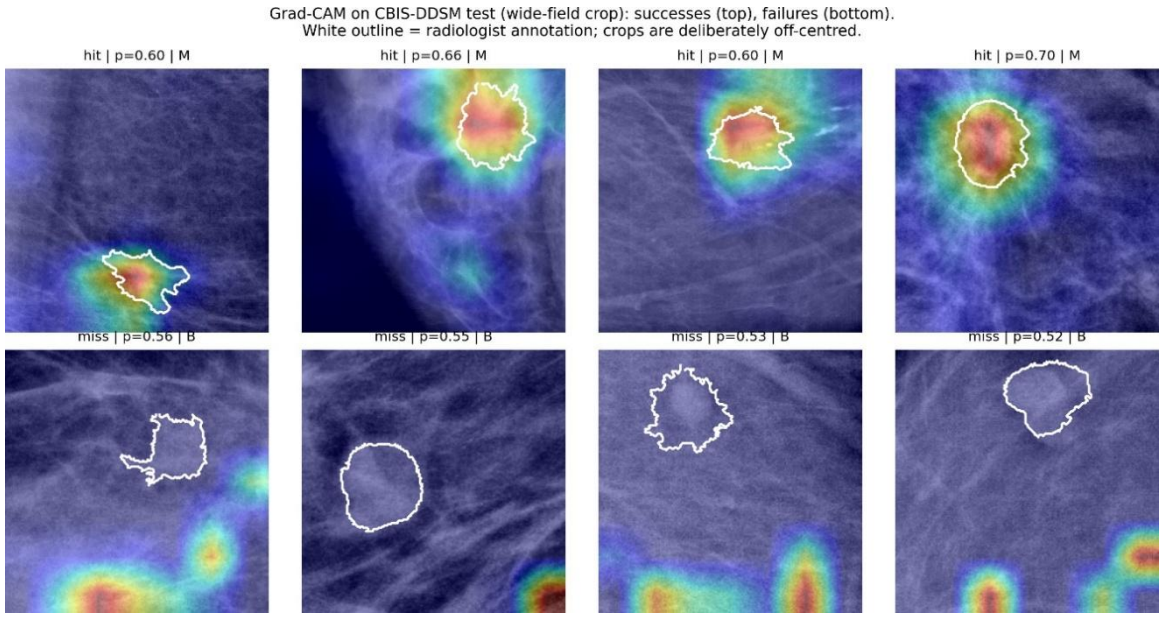


Figure 7. Grad-CAM on deliberately off-centred wide-field CBIS-DDSM crops. Top row: successes. Bottom row: failures, selected automatically as the cases with the least activation mass inside the annotation. The radiologist annotation is outlined in white on every panel, so misses are visible rather than asserted. Every failure shown is a benign lesion scored just above the 0.5 threshold on which the attention has drifted to surrounding fibro-glandular tissue—the false-positive mode discussed in Sub-section 5.15.

Table 18. Contextual comparison with prior work on mammographic lesion classification. Many studies report higher headline numbers, but on whole images or a single dataset and without explicit leakage control or external validation—the dimensions on which this work focuses. The last column states the key limitation relative to our protocol.

Study	Dataset(s)	Method	Reported	Limitation *vs.* this work
Setiawan et al. [7]	MIAS	Law's texture + ANN	93.9% (N/A); 83.3% (B/M)	Whole-image; no leakage control or external validation reported
Agnes et al. [8]	mini-MIAS	Multi-scale CNN (MACNN)	96% sens; AUC 0.99	Whole-image; augment-before-split not controlled; single dataset
Adedigba et al. [20]	Small mammogram set	DenseNet + discriminative fine-tuning	99.8% acc	Near-perfect on small data; standard split, no leakage control or external test
Ragab et al. [6]	DDSM / CBIS-DDSM	AlexNet features + SVM	87.2% acc; AUC 0.94	Image-level; single corpus; no cross-dataset validation
Jafari & Karami [41]	MIAS / DDSM / RSNA	Multi-CNN feature fusion + classifiers	94.5% (MIAS)	Image-level; no patient-wise split; no external validation
Salama & Aly [40]	DDSM / MIAS / CBIS	U-Net segmentation + CNN	98.87% (DDSM)	Segmentation-dependent; very high acc. without reported leakage control
Yu et al. [14]	DDSM	Deep features + dRVFL	AUC 0.93; 81.7% acc	Single dataset; image-level 5-fold; no cross-dataset external validation
Quintana et al. [18]	CBIS-DDSM	Multi-scale (patch)	AUC ≈ 0.722	Patch-based, but single dataset; no external validation
This work	**CBIS (+MIAS, INbreast)**	**EfficientNet patch + SVM/fine-tune**	**AUC 0.745; ext. 0.734 / 0.877**	**Leakage-controlled, patient-wise; 2dataset external validation with clustered 95% CIs and DeLong tests**

5.14 Synthesis

Taken together, the experiments support a single claim: on small mammography datasets, reported performance is governed by evaluation methodology at least as much as by model capacity. The backbone-plus-classifier that appears near-perfect on whole-image Mini-MIAS performs no better than chance once augment-before-split leakage is removed and neither fine-tuning nor higher resolution recovers the apparent skill. The signal that does survive is local: it is recoverable at the lesion-patch scale and withstands a size- and density-matched sampling control. The resulting level is stable across ten seeds,

across a hyper-parameter search, across a repair of the benchmark's own split and across a change of institution. It is not entirely insensitive to capacity, since over 11 configurations spanning 2.2–86 M parameters the test AUC ranges from 0.707 to 0.777, with the largest transformer significantly ahead of the smallest CNN. That range of 0.07 AUC is nevertheless small beside the 0.2 or more separating any of these models from the accuracies that leaky evaluation produces on the same data. The evaluation protocol is thus the larger term. The evidence comes from one lesion-level task on three public cohorts and should not be read as a general law**.**

5.15 Clinical Relevance: What the Errors Would Cost

The intended role of this system is a *second reader* that flags suspicious lesions for radiologist review, never an autonomous diagnostic device. What matters clinically is therefore not the headline AUC, but the asymmetric cost of the two error types and what would have to be true before such a tool could be deployed.

False negatives are the dominant clinical risk. A missed malignancy is the error that changes outcomes: it delays treatment until the next screening round or until the patient presents symptomatically, by which time stage and mortality risk have risen. The per-dataset confusion matrices (Table 8) quantify this: at a 0.5 threshold the model misses 82 of 276 malignant lesions on the internal test set (30%) and 11 of 51 on Mini-MIAS (22%). No operating point available to us makes the system safe to run without a radiologist and a leakage-free figure is what makes this visible; a leaked accuracy of 0.94 would have implied the opposite. In deployment the threshold should therefore be set well below 0.5 to increase sensitivity at the cost of recall workload and the system should be positioned so that its negatives do not suppress a human read.

False positives set the workload cost. Each false positive propagates into recall imaging, patient anxiety and, at the end of the chain, avoidable biopsy of benign tissue. The cross-domain calibration failure matters most here: at a fixed 0.5 threshold, specificity on Mini-MIAS collapses even though the ranking is intact, so a model shipped with its source threshold would flood a new site with false alarms and be switched off long before its discrimination was ever assessed. The transfer procedure of Sub-section 4.6 exists for exactly this reason.

Deployment challenges. Four remain and none is solved here. *(i) Localization.* The pipeline classifies a lesion patch and presumes something has already found the lesion. In a real workflow a detector supplies that box with error; Sub-section 5.11 measures how quickly performance decays as the box degrades and it is the first thing an integrator should re-measure with its own detector. *(ii) Threshold governance.* Because calibration does not transfer, every site needs a defined recalibration step, plus monitoring for drift as scanners and protocols change. The burden here is regulatory and procedural, not only a technical one. *(iii) Case mix.* Our evaluation is lesion-level and enriched with abnormalities; a screening population is overwhelmingly normal, so the positive predictive value in service will be far below anything reported here. *(iv) Scope of evidence.* Both external cohorts are small and retrospective, drawn from curated public archives rather than consecutive clinical series and neither support a claim about a specific device or population.

On the other hand, the computational barrier is genuinely low: the whole pipeline trains and runs on a single 6 GB consumer GPU, so on premise deployment needs no specialized infrastructure and retraining as local annotations accrue is cheap. The binding constraints are clinical and procedural, not computational.

6. REPRODUCIBILITY

Every number in this paper is produced by a script that is released with the manuscript, together with the per-sample predictions those scripts emit. Publishing the raw prediction files matters as much as publishing the code: it is what allows a reader to re-compute our confusion matrices, confidence intervals and significance tests without a GPU and it is what we ourselves lacked when the first version of this manuscript reconstructed its confusion matrices from summary statistics.

The package is available at https://github.com/hitesh712/leakage-free-mammography, archived at DOI:10.5281/zenodo.22122946. Two commands reproduce the statistical results of this paper on a laptop, with no GPU and without downloading the imaging datasets: rev08_analysis.py regenerates the

headline table, the DeLong tests, the calibration figures and the threshold-transfer comparison from the released predictions and rev11_validate_stats.py re-runs the validation of our DeLong implementation reported below.

Released artefacts. *(i)* The full experiment suite: data preparation and patch extraction, the split audit, the ten-seed training run, the patient-disjoint repair, the architecture and resolution ablation, the whole-image baseline, the crop-perturbation study, the Grad-CAM localization analysis and the statistical analysis. *(ii)* The exact split manifests, listing every lesion with its patient identifier, its official partition and its partition after the patient-disjoint repair. *(iii)* Per-sample predicted probabilities for every dataset, every seed and every architecture. *(iv)* The environment records emitted by each run.

Validation of the statistical code. Because the significance claims rest on our own implementation of DeLong's method, we validated it rather than assume it and release the validation script and its output. On the real prediction files, its AUC matches scikit-learn to ten decimal places and its standard error agrees with a 20,000-resample bootstrap standard error to within 0.1%, 1.0% and 0.4% on the internal, Mini-MIAS and INbreast sets, respectively—two independent estimators of the same quantity. Comparing a score vector with itself returns $z = 0$ exactly. Under simulation with two scorers of genuinely equal AUC, the paired test rejects at 0.053 and the unpaired test at 0.046 against a nominal 0.05, so both are correctly calibrated; the unpaired variant matters most here, since it is the less standard of the two and carries the internal-*versus*-external comparison. Power against a large true difference is 0.998 and heavily tied scores are handled correctly.

Settings. Table 19 lists the complete configuration. The ten seeds are 2021, 7, 123, 42, 2024, 5, 777, 31337, 99 and 1234; each seed sets NumPy and framework random state and re-draws the patient-wise validation carve-out. Two software stacks are used and both are reported, since the extended architecture comparison of Sub-section 5.5 runs in PyTorch while the main pipeline runs in TensorFlow.

7. LIMITATIONS

Several caveats apply to these findings. First, the practical ceiling on biopsy-confirmed benign-*versus*-malignant data is malignant sensitivity (Table 8), which would likely require larger cohorts, multi-view fusion or detection-and-aggregation pipelines to improve. Second, INbreast labels are BI-RADS-derived rather than strictly biopsy-confirmed, so its stronger external result reflects greater robustness rather than higher accuracy—the like-for-like comparison remains CBIS→MIAS (Sub-section 5.10). Third, discrimination transfers, but calibration does not, so per- dataset recalibration is required for deployment. Fourth, the approach is intentionally patch-level, relying on per-patch ROI annotations; it does not exploit whole-image triage, multi-view or temporal information and uses a lightweight backbone trainable on a single consumer GPU.

Table 19. Complete training configuration and software environment.

Item	Value
Input	224×224 lesion patch, grayscale replicated to 3 channels, raw [0,255]
Pre-processing	percentile windowing (16-bit only) → resize → CLAHE (clip 2.0, 8 × 8)
Head	global average pooling → dropout 0.4 → dense(1, sigmoid)
Phase 1	head only; Adam, lr10^{-3}; 5 epochs; batch 16
Phase 2	top 40 backbone layers unfrozen; Adam, lr10^{-5}; ≤ 20 epochs
Early stopping	monitor validation AUC, patience 5, restore best weights
Class imbalance	balanced class weights $n/(2n_c)$
Augmentation	rotation ∈ {0, 90, 180, 270} and horizontal flip; training folds only
Validation	GroupShuffleSplit, 15%, grouped by patient, re-drawn per seed
Batch Normalization	kept in inference mode throughout
Seeds	2021, 7, 123, 42, 2024, 5, 777, 31337, 99, 1234
Bootstrap	2,000 resamples, clusters = patients (CBIS) or source images (external)
Main pipeline	Python 3.10.11, TensorFlow 2.10.0, tensorflow-directml-plugin 0.4.0
Architecture study	PyTorch 2.5.1+cu124, timm 1.0.27, mixed precision
Shared libraries	NumPy 1.23.5, pandas 2.3.3, scikit-learn 1.7.2, SciPy 1.15.3, OpenCV 4.13.0, pydicom 3.0.2
Hardware	single NVIDIA RTX 4050 Laptop GPU, 6 GB

Fifth, and most important, is how the external numbers should be read, the external cohorts are small ($n \approx$ 120–130) and our uncertainty estimates understate the true uncertainty. A bootstrap resamples the cases we happen to have collected; it can only express how the estimate would move under re-weighting of *those* cases. It cannot express between-cohort variability—how the estimate would move on a genuinely new sample of patients from a different site, scanner fleet and referral pattern—which for cohorts of this size is likely the larger component. Our intervals are therefore a lower bound on total uncertainty, compounded by both external sets being curated public archives rather than consecutive clinical series. Larger, prospectively assembled cohorts, such as VinDr-Mammo [17] and ideally several of them, are needed before the transfer result can be treated as an estimate of field performance rather than as evidence that transfer occurs at all.

Sixth, the pipeline consumes a lesion patch and therefore presumes that localization has already happened. Sub-section 5.11 measures sensitivity to imperfect boxes, but a full detection-plus-classification evaluation, in which the detector also generates false regions on normal tissue, is outside our scope and would give a lower and more realistic end-to-end figure.

Finally, our claims about whole-image classification are bounded by what we tested: two datasets, EfficientNet-family backbones and input resolutions up to 512 px. We do not claim that whole- image mammography classification is unlearnable in general—with far larger cohorts, multi-view input or detection-and-aggregation architectures demonstrably learnable—only that it did not learn under the data scale and configurations examined here.

8. Conclusion and Future Work

On small mammography datasets, it is methodology rather than architecture that determines reported performance. We identified a widely reproducible leakage pitfall that inflates Mini-MIAS whole-image classification from chance to ~98% and demonstrated leakage-free generalization of a CBIS-DDSM-trained benign-*versus*-malignant classifier (AUC 0.745 internal; 0.734 on MIAS and 0.877 on INbreast, with clustered confidence intervals and ten-seed stability). A DeLong test cannot separate the internal and cross-institution results, which is the specific sense in which the model generalizes.

Two findings sharpen the methodological message. Auditing our own protocol showed that the CBIS-DDSM *official* split is not patient-disjoint once mass and calcification cases are pooled, so leakage entered through a dataset convention rather than a modelling error—an illustration of why split manifests, not just code, should be published. And extending the comparison to six architecture families, including ViT-B/16 and Swin-T, left the performance band unchanged, which converts the ceiling from a conjecture into a tested claim. Finally, because discrimination transfers while calibration does not, we give a label-free procedure that recovers the operating point at a new site.

The result is not a claim of state-of-the-art accuracy, but a reproducible, honestly measured baseline. In future work, we will validate on larger external cohorts, such as VinDr-Mammo [17], investigate domain-generalization and vision–language foundation models—including ordinal or soft-label treatments of BI-RADS—and evaluate the pipeline end-to-end behind a real detector rather than behind ground-truth boxes.

References

[1] A. Bhushan, A. Gonsalves and J. U. Menon, "Current State of Breast Cancer Diagnosis, Treatment and Theranostics," Pharmaceutics, vol. 13, no. 5, p. 723, 2021.

[2] American Cancer Society, "Limitations of Mammograms," [Online], Available: https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html, Jun. 2023.

[3] L. Luo et al., "Deep Learning in Breast Cancer Imaging: A Decade of Progress and Future Directions," IEEE Reviews in Biomedical Engineering, 2024, DOI: 10.1109/RBME.2024.3357567.

[4] H. E. Kim et al., "Transfer Learning for Medical Image Classification: A Literature Review," BMC Medical Imaging, vol. 22, p. 69, DOI: 10.1186/s12880-022-00793-7, 2022.

[5] D. Lévy and A. Jain, "Breast Mass Classification from Mammograms Using Deep Convolutional Neural Networks," arXiv Preprint, 1612.00542, [Online], Available: http://arxiv.org/abs/1612.00542, 2016.

[6] D. A. Ragab, M. Sharkas, S. Marshall and J. Ren, "Breast Cancer Detection Using Deep Convolutional Neural Networks and Support Vector Machines," PeerJ, vol. 7, p. e6201, DOI: 10.7717/peerj.6201, 2019.

[7] A. S. Setiawan et al., "Mammogram Classification Using Law's Texture Energy Measure and Neural

Networks," Procedia Computer Science, vol. 59, pp. 92–97, DOI: 10.1016/j.procs.2015.07.341, 2015.
[8] S. A. Agnes et al., "Classification of Mammogram Images Using Multiscale All Convolutional Neural Network (MA-CNN)," J. of Medical Systems, vol. 44, p. 30, DOI: 10.1007/s10916-019-1494-z, 2020.
[9] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in Machine-learning-based Science," Patterns, vol. 4, no. 9, p. 100804, DOI: 10.1016/j.patter.2023.100804, 2023.
[10] A. Lemay et al., "Improving the Repeatability of Deep Learning Models with Monte Carlo Dropout," NPJ Digital Medicine, vol. 5, p. 174, DOI: 10.1038/s41746-022-00709-3, 2022.
[11] Y. Yang et al., "MammoDG: Generalizable Deep Learning Breaks the Limits of Cross-domain Multi-center Breast Cancer Screening," arXiv preprint, arXiv: 2308.01057, 2023.
[12] Y. Zafari et al., "MammoClean: Toward Reproducible and Bias-aware AI in Mammography through Dataset Harmonization," arXiv preprint, arXiv Preprint, arXiv: 2511.02400, Nov. 2025.
[13] S. Benbakreti et al., "Using ResNet-18 in a Deep-learning Framework and Assessing the Effects of Adaptive Learning rates in the Identification of Malignant Masses in Mammograms," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 10, no. 1, pp. 93–107, DOI: 10.5455/jjcit.71-1699818406, 2024.
[14] X. Yu, Z. Ren, D. S. Guttery and Y.-D. Zhang, "DF-dRVFL: A Novel Deep Feature Based Classifier for Breast Mass Classification," Multimedia Tools and Applications, vol. 83, no. 5, pp. 14393-14422, 2023.
[15] R. S. Lee et al., "A Curated Mammography Data Set for Use in Computer-aided Detection and Diagnosis Research," Scientific Data, vol. 4, p. 170177, DOI: 10.1038/sdata.2017.177, 2017.
[16] I. C. Moreira et al., "INbreast: Toward a Full-field Digital Mammographic Database," Academic Radiology, vol. 19, no. 2, pp. 236– 248, DOI: 10.1016/j.acra.2011.09.014, 2012.
[17] H. T. Nguyen et al., "VinDr-Mammo: A Large-scale Benchmark Dataset for Computer-aided Diagnosis in Full-field Digital Mammography," Scientific Data, vol. 10, p. 277, DOI: 10.1038/s41597-023- 02100-7, 2023.
[18] G. I. Quintana et al., "Exploiting Patch Sizes and Resolutions for Multi-scale Deep Learning in Mammogram Image Classification," Bioengineering, vol. 10, no. 5, p. 534, DOI: 10.3390/bioengineering10050534, 2023.
[19] P. Oza, P. Sharma, S. Patel, F. Adedoyin and A. Bruno, "Image Augmentation Techniques for Mammogram Analysis," Journal of Imaging, vol. 8, no. 5, p. 141, DOI: 10.3390/jimaging8050141, 2022.
[20] A. P. Adedigba, S. A. Adeshina and A. M. Aibinu, "Performance Evaluation of Deep Learning Models on Mammogram Classification Using Small Dataset," Bioengineering, vol. 9, no. 4, p. 161, 2022.
[21] A. G. M. Al Mansour, F. Alshomrani, A. Alfahaid and A. T. M. Almutairi, "MammoViT: A Custom Vision Transformer Architecture for Accurate BI-RADS Classification in Mammogram Analysis," Diagnostics, vol. 15, no. 3, p. 285, DOI: 10.3390/diagnostics15030285, 2025.
[22] S. Elumalai et al., "Breast Cancer Classification Based on Micro-calcifications Using Dual Branch Vision Transformer Fusion," Scientific Reports, vol. 15, DOI: 10.1038/s41598- 025-34377-6, 2025.
[23] S. Sharma, Y. Singh and T. Choudhury, "Advanced Deep Learning Architectures for Enhanced Mammography Classification: A Comparative Study of CNNs and ViT," Discover Artificial Intelligence, vol. 5, Article no. 187, DOI: 10.1007/s44163-025-00426-2, 2025.
[24] S. Ghosh et al., "Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography," Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI 2024), vol. LNCS 15012, DOI: 10.1007/978-3-031-72390-2_59, Morocco, 2024.
[25] F. M. Talaat et al., "Grad-CAM Enabled Breast Cancer Classification with a 3D Inception-ResNet V2: Empowering Radiologists with Explainable Insights," Cancers, vol. 16, no. 21, p. 3668, 2024.
[26] 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," Proc. of the IEEE Int. Conf. on Computer Vision (ICCV), pp. 618–626, Venice, Italy, 2017.
[27] 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, p. 110492, 2025.
[28] H. C. Lu, E. W. Loh and S. C. Huang, "The Classification of Mammogram Using Convolutional Neural Network with Specific Image Pre-processing for Breast Cancer Detection," Proc. of the 2nd IEEE Int. Conf. on Artificial Intelligence and Big Data (ICAIBD), pp. 9–12, Chengdu, China, 2019.
[29] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Proc. of the 36th Int. Conf. on Machine Learning (ICML), pp. 6105–6114, PMLR 97, L. B., California, USA, 2019.
[30] S. R. Salian and S. D. Sawarkar, "Melanoma Skin Lesion Classification Using Improved Efficient-NetB3," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 8, no. 1, pp. 45–56, DOI: 10.5455/jjcit.71-1636005929, 2022.
[31] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
[32] G. Huang et al., "Densely Connected Convolutional Networks," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, 2017.

[33] Z. Liu et al., "A ConvNet for the 2020s," Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986, 2022.

[34] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv Preprint, arXiv: 2010.11929, 2021.

[35] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proc. of the 2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 9992–10002, 2021.

[36] E. R. DeLong et al., "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.

[37] X. Sun and W. Xu, "Fast Implementation of DeLong's Algorithm for Comparing the Areas under Correlated Receiver Operating Characteristic Curves," IEEE Signal Process. Lett., vol. 21, no. 11, pp. 1389–1393, 2014.

[38] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks," Proceedings of the 34th International Conference on Machine Learning (ICML'17) – vol. 70, pp. 1321 – 1330, 2017.

[39] W. J. Youden, "Index for Rating Diagnostic Tests," Cancer, vol. 3, no. 1, pp. 32–35, 1950.

[40] W. M. Salama and M. H. Aly, "Deep Learning in Mammography Images Segmentation and Classification: Automated CNN Approach," Alexandria Engineering Journal, vol. 60, no. 5, pp. 4701–4709, 2021.

[41] Z. Jafari and E. Karami, "Breast Cancer Detection in Mammography Images: A CNN-based Approach with Feature Selection," Information, vol. 14, no. 7, DOI: 10.3390/info14070410, 2023.

ملخص البحث:

غالباً ما تكون معدّلات الدّقّة المرتفعة المبلَّغ عنها في مجموعات بيانات تصوير الثّدي بالأشعة السّينية مبالغاً فيها بسبب تسرُّب البيانات الذي ينجم غالباً عن عمليات تعزيز البيانات المطبّقة قبل تقسيم البيانات إلى مجموعاتٍ للتّدريب والاختبار. فعلى سبيل المثال، يؤدّي مجرّد تغيير ترتيب التّقسيم في مجموعة بيانات (Mini-MIAS) إلى انخفاض دقة نموذج يعتمد على استخراج السّمات العميقة وآلة المتّجهات الدّاعمة من 0.976 إلى 0.513، وهي نسبة تعادل الصّدفة العشوائية، في حين تظلّ الإشارة التّنبؤية صالحة عند مستوى رقعة الآفة.

تأتي مساهمتنا على مستوى البروتوكول المنهجي وليس على مستوى البنية المعمارية للنّموذج، وتتمثّل في إجراء تدقيقٍ قابل للتّكرار لكشف تسرُّب البيانات، وتحديد ومعالجة مشكلة تداخل المرضى، حيث تتبّعنا السّبب إلى استخلاص أقسام الكُتل الورمية والتكلّسات بشكلٍ مستتلّ، مما شمل 31 مريضاً؛ أي 8.5% من آفاق مجموعة الاختبار في مجموعة البيانات (CBIS-DDSM)، وإجراء مسح شامل للبنى المعمارية لاختبار سقف الأداء بدلاً من افتراضه، وابتكار إجراء لا يعتمد على التّصنيفات لنقل عتبة اتّخاذ القرار إلى مؤسسة طبية جديدة.

وقد حقّق نموذج خفيف الوزن يعتمد على معالج الرُّقع تمّ تدريبه على مجموعة بيانات (CBIS-DDSM) وفق بروتوكول صارم نتائج جيدة في الفصل بين المرضى وقيماً جيدة من حيث مؤشرات الأداء اتسمت بالاستقرار لعددٍ من محموعات البيانات التي استخدمت لتقييم أداء النموذج.

مما سبق نستنتج أنّ التّحكُّم في تسرُّب البيانات والتحقق الخارجي من صحة النّموذج هما أمران أكثر من مجرّد النّظر إلى أرقام الدّقّة المعلنة عند التّعامل مع مجموعات بيانات آفات الثّدي.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).