# Convolutional Neural Networks and Vision Transformers for Patch-based Mammography Classification: the Role of Preprocessing, Spatial Context and Pretraining

Aneta Gábrišová Lokšová, Michal Pavlov and Ivan Cimrák
University of Žilina
Žilina, Slovakia
Aneta.Loksova, Ivan.Cimrak@fri.uniza.sk

***Abstract*—Convolutional neural networks have long defined the standard for image analysis in radiology, while vision transformers offer an alternative inductive bias based on self-attention over image patches. We compare the two families on patch-based benign-versus-malignant classification of lesions from the CBIS-DDSM, using the official patient-disjoint split. Six controlled experiments vary the patch-extraction strategy, the size of the surrounding spatial context, the lesion types included, the depth of fine-tuning, and the presence of ImageNet pretraining. Each configuration is trained five times and compared using paired and independent $t$-tests. Hyperparameters were selected based on the results obtained on the test split. In pretrained settings, the Data-efficient Image Transformer reaches higher accuracy than ResNet-50 in all five experiments, although its margin over EfficientNet-B7 is smaller and limited to accuracy. ResNet-50 degraded significantly when calcifications were added to both the training and test data and when all layers were unfrozen, whereas no significant change was detected for the Data-efficient Image Transformer. A direct test of the difference between the two architectures' responses was, however, at most borderline, so this asymmetry is suggestive rather than established. Widening the crop from $224 \times 224$ to $448 \times 448$ pixels before downsampling improves both architectures, and the Data-efficient Image Transformer most of all. Without pretraining, no significant difference in accuracy was detected between the two models, and both perform poorly. Absolute performance remains well below clinical usability, and we discuss the design choices that limit it.**

## I. Introduction

Historically, breast cancer has been the primary oncological challenge for the female population. Today, it remains the most frequently diagnosed malignancy and the leading cause of cancer-associated mortality among women [1]. To address this concern, clinical practice relies on mammography screening, a method of early diagnosis. It uses low-dose X-ray technology to perform regular examinations of breast tissue, allowing for the detection of non-palpable lesions before clinical symptoms appear. However, the high volume of screening data requires advanced diagnostic tools to help radiologists maintain high sensitivity and specificity.

During the past decade, deep learning has become the dominant approach to automated mammography analysis, driven both by architectural innovation and by the growth of high-performance computing. Convolutional Neural Networks (CNNs) [2], [3] have long served as the architectural default. Their efficacy in medical imaging is largely attributed to their inherent inductive biases, which are effective in extracting low-level features such as edges and textures in radiographic images. The adaptation of the Transformer architecture to the visual domain, the Vision Transformer (ViT) [4], replaces these biases with self-attention over a sequence of image patches and has been reported to match or exceed CNNs on natural-image benchmarks given sufficient pretraining data. Whether this translates into an advantage on small, low-contrast, single-channel mammographic data is an empirical question, and one in which the published evidence is mixed [5], [6].

We address this question in a deliberately narrow, controlled study in which the classification task is fixed and the surrounding design choices are varied one at a time. Lesion patches are extracted around the annotated regions of CBIS-DDSM, so the models act as second readers on findings a radiologist has already located. We compare ResNet-50 and DeiT-base, with EfficientNet-B0 and B7 as additional convolutional baselines, starting from the protocol of Spijker [7]. Five further experiments then change how the patch is extracted (the whole region resized, a 224 × 224 crop at native resolution, or a 448 × 448 crop downsampled), which lesion types are pooled, how many layers are fine-tuned, and whether ImageNet pretraining is used. Every configuration is trained five times and compared with t-tests, which the original comparison lacked, so that differences between models can be separated from run-to-run noise. The study yields four results:

- **Result 1: Architecture ranking.** With ImageNet pretraining, DeiT reaches higher accuracy than ResNet-50 in all five pretrained experiments, but its advantage over EfficientNet-B7 is limited to accuracy.
- **Result 2: Sensitivity to design.** ResNet-50 degrades significantly when calcifications are added and when all layers are unfrozen, while no significant change is

detected for DeiT; a direct test supports this asymmetry only weakly, so we report it as a hypothesis.

- **Result 3: Spatial context.** Enlarging the crop to 448 × 448 pixels before downsampling improves both architectures and gives DeiT its best result, an effect larger than the gap between the two architectures in Experiment 1.
- **Result 4: Pretraining.** Without ImageNet weights, both models perform poorly, no significant difference in accuracy is detected between them, and they fail in opposite directions.

## II. Related Work

In the literature review [5], the authors compared CNNs and ViTs for image classification in natural and medical imaging. Several studies in this review report improved robustness and generalization of transformer-based models, while others highlight the efficiency and stability of CNNs.

The authors of the systematic review [6] lean towards a similar opinion as the authors of the previous review. Both architectures have been widely applied to diverse imaging tasks, including classification, segmentation, and registration, with radiography being the most common modality studied. The review of 36 studies reveals that transformer-based models, particularly ViTs, often demonstrate competitive or superior performance and robustness relative to CNNs, especially when pre-training is employed, although results can vary depending on dataset and task specifics. The authors also note that studies directly comparing these two architectures are limited and further research is necessary.

Matsoukas et al. [8] conducted the comparison most closely related to ours, evaluating ResNet-50 against DeiT-S on three medical datasets including CBIS-DDSM. Their conclusion was three-part: CNNs are stronger when trained from scratch, off-the-shelf ViTs are on par with CNNs once both are pretrained on ImageNet, and ViTs move ahead when pretrained with self-supervision.

Although the reviews mentioned above are generalized to medical images, the experimental study of Cantone et al. [9] focused on mammography. It evaluated 33 models, 19 convolutional and 14 transformer-based, on the large private OMI-DB dataset, and found that ViTs can achieve performance comparable to established CNNs such as ResNet for whole-image classification. Our study differs in three respects: it operates on lesion patches rather than whole images, it uses the public CBIS-DDSM so that results are reproducible, and it varies preprocessing systematically rather than holding it fixed while varying the architecture.

Several works report transformer-based mammography classifiers with strong headline numbers [10], [11]. These are difficult to place against our results because dataset construction, class balance, and split protocols differ substantially and because the reported accuracies on CBIS-DDSM in the literature span a very wide range. As a reference point on the same dataset, the end-to-end convolutional approach of Shen et al. [12] achieves a per-image AUC of 0.88.

The direct predecessor of this study is the work of Spijker [7], who compared ResNet-50 and DeiT-base on CBIS-DDSM mass regions of interest (ROIs) under ImageNet pretraining with a frozen backbone, reporting accuracy of 0.669 and 0.753 respectively. Our Experiment 1 reproduces that protocol. We obtain the same ordering but lower absolute values for both models and extend the comparison beyond a single preprocessing pipeline and a single CNN.

## III. Preparation for experiments

### A. Mammography Data and Dataset Creation

We use CBIS-DDSM [13], a standardized and curated version of the Digital Database for Screening Mammography (DDSM). The CBIS-DDSM subset was selected and processed by trained mammographers, with images decompressed and converted to DICOM and provides updated ROI annotations, bounding boxes, and curated pathologic diagnoses. Importantly, for this study, CBIS-DDSM ships with an official train/test division that is disjoint at the patient level; we use it unchanged and never resample it.

Rather than using whole-image mammograms, which contain substantial non-informative background and pectoral muscle, we adopt a patch-based approach. Using the provided ROI coordinates, patches are centered on annotated lesions. This mimics a workflow in which the model acts as a second reader for abnormalities that a radiologist has already located, and means that our numbers are not comparable to whole-image classifiers such as [9], [12].

The original images are 16-bit grayscale. For compatibility with the standard 8-bit image pipeline, 16-bit pixel values were divided by 256 and cast to `uint8`. This step is not required by the networks, which accept floating-point input, and it discards grayscale precision; we did not evaluate its effect. Since the models expect three input channels, each grayscale patch was duplicated into three identical channels. The patches were then normalized channel-wise using the mean and standard deviation computed from the converted training data. No data augmentation of any kind was used. In both crop strategies, windows extending beyond the image border were zero-padded.

Three patch-extraction strategies were compared.

- **Whole ROI, resized.** The complete annotated ROI is cropped and resized to $224 \times 224$ pixels. Following [7], ROIs whose aspect ratio (width/height) fell outside the interval $[0.4, 1.5]$ were discarded to bound the geometric distortion introduced by resizing. This filter is the reason why the size of the dataset in Table I differs slightly between experiments.
- **Centered** $224 \times 224$ **crop.** The ROI center is located, and a fixed $224 \times 224$ window is cropped at native resolution, without resizing. This preserves resolution, but truncates lesions larger than the window.
- **Centered** $448 \times 448$ **crop, downsampled.** A $448 \times 448$ window is cropped around the ROI center and then resized to $224 \times 224$. This preserves the square aspect

ratio and includes more peri-lesional tissue at the cost of spatial resolution.

The official training split was divided 80:20 into training and validation subsets at the patch level; this division is not patient-disjoint. The validation subset was used for early stopping and checkpoint selection. The official test split was never used for training or checkpoint selection, but was used to choose among hyperparameter configurations. The results should therefore be read as exploratory rather than as unbiased estimates of generalization performance.

### B. Architectures

Two families were compared. CNNs process images through convolutional layers applied to local receptive fields, followed by pooling and fully connected layers, and benefit from locality and translation-equivariance biases. We evaluate ResNet-50 [2], which uses shortcut connections to enable optimization of deep networks, and EfficientNet-B0 and B7 [3], which apply compound scaling of depth, width, and input resolution. ViTs [4] instead treat an image as a sequence of patches and use self-attention to weight their mutual relevance, giving a global receptive field from the first layer. We use the Data-efficient Image Transformer [14], specifically `deit_base_patch16_224`, which is a ViT-Base trained on ImageNet-1k with the DeiT data-efficient training recipe; the distilled variant was not used.

The four models are not matched in capacity: EfficientNet-B0 has approximately 5.3 M parameters, ResNet-50 25.6 M, EfficientNet-B7 66 M and DeiT-base 86 M. All were run at $224 \times 224$ input, which is below the resolution at which EfficientNet-B7 was designed to operate and therefore handicaps it.

### C. Training Protocol

All backbones except for Experiment 6 were initialized with ImageNet weights. The final classification layer was replaced by two linear layers separated by a ReLU activation, following [7]. Training used Adam optimizer [15] with cross-entropy loss, batch size 32, and a maximum of 50 epochs, with early stopping on validation loss and patience of 5 epochs. The checkpoint with the lowest validation loss was retained for testing.

Two hyperparameters were tuned by grid search: the width of the intermediate linear layer $\in \{128, 256, 512\}$ and the learning rate $\in \{5\times10^{-3}, 5\times10^{-4}, 5\times10^{-5}, 5\times10^{-6}\}$. The selected values are reported alongside every result because they differ substantially between models and experiments and are necessary for reproduction. In particular, the learning rate selected for DeiT varies by more than three orders of magnitude between experiments, which is itself an indication that a single shared learning rate would not have produced a fair comparison.

For each model and experiment, every grid configuration was trained five times, and the configuration with the highest mean test accuracy was retained; the validation subset determined only early stopping and the retained checkpoint. Because selection and reporting use the same data, the reported values are optimistically biased. An unbiased estimate would require an additional held-out set or nested evaluation, which we did not perform.

### D. Evaluation Metrics and Statistical Analysis

The positive class throughout is malignant. We report accuracy, recall (sensitivity), precision, specificity, $F_1$ and balanced accuracy. Specificity, $F_1$ and balanced accuracy were computed from the averaged confusion matrices of the five runs; because they are derived from averaged counts rather than averaged per-run scores, they should be read as summary indicators rather than exact means. The test sets are benign-majority, so a constant classifier predicting benign attains 0.607 accuracy in E1, E5 and E6, 0.601 in E2 and E4, and 0.593 in E3. Accuracy, which was used for selection and testing, should be read against these baselines; balanced accuracy, which accounts for class imbalance, is reported descriptively as a complementary summary. Confidence intervals are t-based intervals for the mean of five runs. Within an experiment, run i of one model was paired with run i of the other and they share the same seed and also data order. Cross-experiment comparisons use Welch's independent t-test. No correction for multiple comparisons was applied; across the roughly 60 tests reported, about three false positives are expected at $\alpha = 0.05$, so isolated results near the threshold should be interpreted with caution. All intervals and tests capture variability due to initialization and training order only. They do not account for sampling of patients or lesions, and since several patches can come from the same patient, including CC (craniocaudal) and MLO (mediolateral oblique) views of the same lesion, the effective test-set size is smaller than the patch count.

### E. Experimental Design

Table I summarizes the six experiments and the resulting dataset sizes. Experiments 2, 3, and 4 vary preprocessing and dataset composition with the training regime held fixed; Experiments 5 and 6 vary the training regime with the Experiment 1 data being held fixed. Each experiment differs from its reference condition in one experimental factor, with the caveats that hyperparameters were re-selected for every experiment, that E2–E4 contain slightly different patches from E1 because the aspect-ratio filter applies only to whole-ROI resizing, and that E3 also changes the test population.

## IV. ResNet-50 versus DeiT under six controlled conditions

### A. Experiment 1: architecture comparison

Experiment 1 assesses how well the four models separate malignant from benign masses when the complete ROI is resized to the input size of the network. The results are given in Table II.

DeiT achieved the highest mean accuracy (0.7034) and ResNet-50 the lowest (0.6475). Paired $t$-tests show that ResNet-50 is significantly worse than both EfficientNet-B7 and DeiT on all three primary metrics, and significantly worse

TABLE I. EXPERIMENTAL DESIGN AND DATASET COMPOSITION. B/M DENOTES BENIGN/MALIGNANT PATCH COUNTS. ALL EXPERIMENTS USE IMAGENET-PRETRAINED WEIGHTS EXCEPT E6, AND FREEZE ALL LAYERS EXCEPT THE REPLACED HEAD EXCEPT IN E5 AND E6.

| | Lesion types | Patch extraction | Layers trained | Pretrained | Train (B/M) | Test (B/M) | Reference |
|---|---|---|---|---|---|---|---|
| E1 | masses | whole ROI $\rightarrow$ resize 224 | head only | ImageNet | 621 / 591 | 215 / 139 | — |
| E2 | masses | centred 224 crop | head only | ImageNet | 635 / 602 | 218 / 145 | E1 |
| E3 | masses + calcifications | centred 224 crop | head only | ImageNet | 1362 / 1111 | 385 / 264 | E2 |
| E4 | masses | centred 448 crop $\rightarrow$ resize 224 | head only | ImageNet | 635 / 602 | 218 / 145 | E2 |
| E5 | masses | whole ROI $\rightarrow$ resize 224 | all layers | ImageNet | 621 / 591 | 215 / 139 | E1 |
| E6 | masses | whole ROI $\rightarrow$ resize 224 | all layers | random init | 621 / 591 | 215 / 139 | E5 |

than EfficientNet-B0 on recall. The comparison between the two strongest models is more restrained: DeiT significantly exceeds EfficientNet-B7 in accuracy, but the differences in recall and precision are not significant. Between the two EfficientNets, only precision differs significantly. In other words, this experiment supports the claim that DeiT outperforms *ResNet-50*, but supports only a narrow claim of superiority over the strongest CNN tested and offers no support at all for a general statement about CNNs as a family.

The ordering reproduces that reported by Spijker [7], but our absolute values are lower for both architectures (0.6475 vs. 0.669 for ResNet-50; 0.7034 vs. 0.753 for DeiT). The most likely explanations are differences in the ROI extraction and in the resulting class balance: [7] reports 784 benign and 912 malignant mass ROIs before splitting, while our extraction yields a benign-majority set (836 benign, 730 malignant across both splits). We were unable to reconcile the two extractions exactly and treat this as a partial rather than an exact replication.

For Experiments 2-6 we retained ResNet-50 rather than the stronger EfficientNet-B7. The reason is comparability with [7] and [8], both of which use ResNet-50, and its continued prevalence in mammography work. This choice must be kept in mind when reading the remaining results: they constitute a controlled ResNet-50-versus-DeiT comparison, not a CNN-versus-ViT verdict, and EfficientNet-B7 was within four accuracy points of DeiT in the one experiment where both were run.

### *B. Effect of the patch-extraction strategy*

Experiment 2 replaces the resize-the-whole-ROI strategy with a fixed $224 \times 224$ crop taken in native resolution around the ROI center; Experiment 4 crops $448 \times 448$ and downsamples. The results appear in Table III, and the cross-experiment tests are summarized in Table IV.

The two architectures showed different patterns in Experiment 2. For ResNet-50, no significant change in accuracy or recall was detected relative to Experiment 1, while precision improved significantly, consistent with resizing-induced geometric distortion having harmed the convolutional model. For DeiT the effect is the opposite: accuracy decreases significantly, with no significant change in recall or precision, suggesting that for the transformer the information lost by truncating large lesions outweighs the distortion avoided.

Experiment 4 is the clearest positive result in the study. Enlarging the crop to $448 \times 448$ before downsampling significantly improved ResNet-50 in recall, and improved DeiT significantly in all three metrics, to its best overall values (accuracy 0.7377, recall 0.7021). Because Experiment 4 differs from Experiment 2 only in the width of the window and the subsequent downsampling, the natural reading is that the peri-lesional context carries diagnostic signal and that the resolution sacrificed to obtain it is worth the trade at this scale. That both architectures benefit indicates this is a property of the task rather than of self-attention, although the effect is larger for DeiT.

### *C. Effect of lesion heterogeneity*

Experiment 3 adds calcifications to the masses for both training and test data, roughly doubling the dataset. ResNet-50 lost significantly in accuracy and recall compared with Experiment 2, while no significant change was detected for DeiT. However, this does not show that the two architectures react differently: when we tested the difference between their changes directly, it was not statistically significant. In addition, the Experiment 3 test set also contains calcifications, so the degradation cannot be attributed to calcifications in the training data rather than to a harder test set. Evaluating the Experiment 3 models on masses only would separate these effects and is left to future work.

### *D. Effect of fine-tuning depth*

Experiment 5 unfreezes all layers, keeping the Experiment 1 data. ResNet-50 significantly decreased accuracy and precision, while DeiT showed no significant change on any metric. A direct test of the difference between the two architectures' changes was only borderline significant, so this does not establish that they respond differently. Full fine-tuning of a 25.6 M-parameter network on roughly 1 200 patches without augmentation is a plausible overfitting regime, and this is our interpretation. However, it is confounded with the learning rate grid: the values available to the search were tuned for head-only training, and full fine-tuning normally requires a smaller learning rate and often a layer-wise schedule. This experiment should therefore be read as showing that naive full fine-tuning did not help under our search space, not that full fine-tuning cannot help.

### *E. Effect of pretraining*

Experiment 6 trains both architectures from random initialization. Because Experiment 1 differs from it in two respects (pretraining and the number of trained layers), we compare it with Experiment 5, which also trains all layers but starts

TABLE II. EXPERIMENT 1: MASSES, WHOLE ROIS RESIZED TO $224 \times 224$, HEAD-ONLY TRAINING. MEAN OVER FIVE RUNS WITH 95 % CONFIDENCE INTERVALS. SPEC., $F_1$ AND BAL. ACC. ARE DERIVED FROM THE AVERAGED CONFUSION MATRICES. LR = LEARNING RATE, $d$ = WIDTH OF THE INTERMEDIATE LINEAR LAYER.

| Model | Params | $d$ | LR | Accuracy | Recall | Precision | Spec. | $F_1$ / bal. acc. |
|---|---|---|---|---|---|---|---|---|
| ResNet-50 | 25.6 M | 256 | $5\times10^{-6}$ | 0.6475 (0.643–0.652) | 0.5252 (0.502–0.548) | 0.5542 (0.546–0.563) | 0.727 | 0.539 / 0.626 |
| EfficientNet-B0 | 5.3 M | 512 | $5\times10^{-6}$ | 0.6576 (0.634–0.681) | 0.5813 (0.562–0.600) | 0.5630 (0.534–0.592) | 0.707 | 0.572 / 0.644 |
| EfficientNet-B7 | 66 M | 512 | $5\times10^{-6}$ | 0.6842 (0.671–0.698) | 0.5612 (0.519–0.603) | 0.6058 (0.587–0.625) | 0.764 | 0.583 / 0.662 |
| DeiT-base | 86 M | 128 | $5\times10^{-5}$ | **0.7034** (0.688–0.719) | **0.6072** (0.576–0.638) | **0.6294** (0.597–0.662) | **0.766** | **0.618 / 0.686** |

from ImageNet weights. Against this reference, DeiT lost significantly in accuracy and precision. For ResNet-50 the picture is different: the accuracy did not change significantly, but the recall dropped sharply, while the precision increased. This does not mean that ResNet-50 is unaffected by the lack of pretraining. Its Experiment 5 reference was already degraded, and without pretraining the model mainly predicted the benign class, which keeps accuracy close to the majority-class baseline while missing most malignant lesions. Therefore, both models lose useful discrimination without pretraining, confirming the importance of transfer learning at this dataset size, in agreement with [8]. Compared with Experiment 1, the best pretrained configuration, the losses are larger for both models, but that comparison mixes two changes.

The comparison between the two models with random initialization requires care. The difference in accuracy (0.6169 for ResNet-50 vs. 0.6056 for DeiT) is *not* statistically significant; DeiT has significantly higher recall and ResNet-50 significantly higher precision. Furthermore, the ResNet-50 operating point is close to degenerate: the mean recall is 0.2317 with a 95 % confidence interval of $[0.070, 0.394]$, meaning that in some runs the model classified almost every patch as benign. Its apparently high derived specificity (0.866) reflects this collapse rather than useful discrimination, and its derived $F_1$ of 0.334 is the worst of any configuration in the study. We therefore do not claim, as would be conventional, that the CNN is superior without pretraining; the correct statement is that no significant difference in accuracy was detected, that they fail in different directions, and that neither is usable.

### *F. Summary across experiments*

Fig. 1 shows accuracy and recall for both architectures across all six experiments with their confidence intervals. Two patterns are visible. DeiT is above ResNet-50 in accuracy in every pretrained setting, but the gap is between two and eight accuracy points; the smaller of these gaps may be within the uncertainty caused by the limited test-set size, which our intervals do not account for. And both models collapse in E6, with ResNet-50's recall interval becoming so wide as to be uninformative.

Two distinct notions of robustness should be separated here. Across design changes, no significant change was detected for DeiT in E3 and E5 while ResNet-50 degraded; because the direct comparison of these changes was not (or only marginally) significant, this does not establish that DeiT is more stable. Across random seeds within a fixed design, the ordering is often reversed – ResNet-50 exhibited the narrower accuracy spread in E1, E2, E3 and E5, with DeiT narrower only in E4 and the two comparable in E6. Claims that transformers are simply "more robust" conflate these, and our data at most suggest the first and do not support the second.

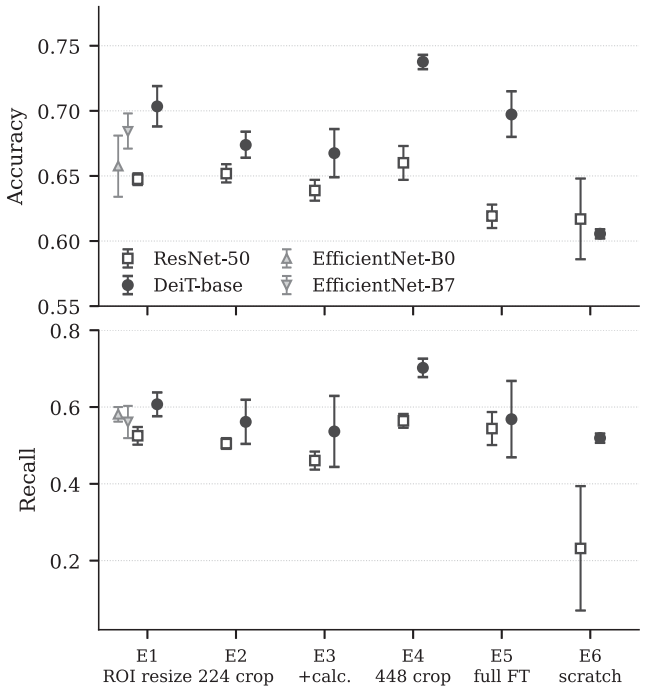



Fig. 1. Accuracy (top) and recall (bottom) across the six experiments, mean of five runs with 95 % confidence intervals for the mean. EfficientNet-B0 and B7 were evaluated in E1 only. E1 = whole ROI resized, E2 = centred 224 crop, E3 = E2 plus calcifications, E4 = centred 448 crop downsampled, E5 = full fine-tuning, E6 = training from scratch. Note that E3 uses a larger and different test population, and that these intervals reflect run-to-run variability only, not finite-test-set uncertainty.

## V. DISCUSSION

*Architecture ranking under pretraining (Result 1).* Designed as a comparison of architecture, this study offers a qualified answer. In every pretrained configuration, DeiT reached a higher accuracy than ResNet-50, and in Experiment 4 it did so by its largest recall margin (13.8 recall points), well beyond the run-to-run variability. However, against the strongest CNN tested, the advantage was limited to accuracy. This pattern – parity or a modest transformer advantage with ImageNet

TABLE III. EXPERIMENTS 2–6: RESNET-50 VERSUS DEIT-BASE. MEAN OVER FIVE RUNS WITH 95 % CONFIDENCE INTERVALS. SPEC., $F_1$ AND BAL. ACC. ARE DERIVED FROM THE AVERAGED CONFUSION MATRICES. BOLD MARKS A WITHIN-EXPERIMENT DIFFERENCE SIGNIFICANT AT $\alpha = 0.05$ BY PAIRED $t$-TEST.

| | Model | $d$ | LR | Accuracy | Recall | Precision | Spec. / $F_1$ / bal. acc. |
|---|---|---|---|---|---|---|---|
| E2 | ResNet-50 | 128 | $5\times10^{-6}$ | 0.6518 (0.645–0.659) | 0.5048 (0.491–0.519) | 0.5727 (0.563–0.583) | 0.750 / 0.537 / 0.627 |
| | DeiT-base | 128 | $5\times10^{-4}$ | **0.6738** (0.664–0.684) | **0.5614** (0.504–0.619) | 0.6019 (0.574–0.630) | 0.749 / 0.581 / 0.655 |
| E3 | ResNet-50 | 512 | $5\times10^{-6}$ | 0.6388 (0.631–0.647) | 0.4606 (0.437–0.484) | 0.5694 (0.557–0.581) | 0.761 / 0.509 / 0.611 |
| | DeiT-base | 128 | $5\times10^{-4}$ | **0.6675** (0.649–0.686) | 0.5364 (0.444–0.629) | 0.6151 (0.567–0.663) | 0.757 / 0.573 / 0.647 |
| E4 | ResNet-50 | 128 | $5\times10^{-5}$ | 0.6601 (0.647–0.673) | 0.5641 (0.546–0.582) | 0.5772 (0.557–0.597) | 0.724 / 0.571 / 0.644 |
| | DeiT-base | 128 | $5\times10^{-5}$ | **0.7377** (0.732–0.743) | **0.7021** (0.678–0.726) | **0.6625** (0.650–0.675) | 0.761 / 0.682 / 0.732 |
| E5 | ResNet-50 | 128 | $5\times10^{-5}$ | 0.6192 (0.610–0.628) | 0.5439 (0.501–0.587) | 0.5148 (0.504–0.526) | 0.668 / 0.529 / 0.606 |
| | DeiT-base | 256 | $5\times10^{-3}$ | **0.6972** (0.680–0.715) | 0.5683 (0.469–0.668) | **0.6394** (0.585–0.694) | 0.781 / 0.602 / 0.674 |
| E6 | ResNet-50 | 512 | $5\times10^{-5}$ | 0.6169 (0.586–0.648) | 0.2317 (0.070–0.394) | **0.6008** (0.523–0.678) | 0.866 / 0.334 / 0.549 |
| | DeiT-base | 128 | $5\times10^{-6}$ | 0.6056 (0.602–0.609) | **0.5194** (0.507–0.531) | 0.4979 (0.493–0.503) | 0.661 / 0.508 / 0.590 |

TABLE IV. CROSS-EXPERIMENT COMPARISONS AGAINST THE REFERENCE CONDITION OF TABLE I, BY INDEPENDENT $t$-TEST AT $\alpha = 0.05$. $\uparrow$ SIGNIFICANT IMPROVEMENT, $\downarrow$ SIGNIFICANT DEGRADATION, $=$ NO SIGNIFICANT DIFFERENCE.

| | Model | Acc. | Recall | Prec. |
|---|---|---|---|---|
| E2 vs. E1 | ResNet-50 | $=$ | $=$ | $\uparrow$ |
| | DeiT-base | $\downarrow$ | $=$ | $=$ |
| E3 vs. E2 | ResNet-50 | $\downarrow$ | $\downarrow$ | $=$ |
| | DeiT-base | $=$ | $=$ | $=$ |
| E4 vs. E2 | ResNet-50 | $=$ | $\uparrow$ | $=$ |
| | DeiT-base | $\uparrow$ | $\uparrow$ | $\uparrow$ |
| E5 vs. E1 | ResNet-50 | $\downarrow$ | $=$ | $\downarrow$ |
| | DeiT-base | $=$ | $=$ | $=$ |
| E6 vs. E5 | ResNet-50 | $=$ | $\downarrow$ | $\uparrow$ |
| | DeiT-base | $\downarrow$ | $=$ | $\downarrow$ |

pretraining, no advantage without it – matches the conclusions of Matsoukas et al. [8] and of the systematic review in [6], and we regard this result as consistent with prior findings rather than as new. This advantage is not fixed, however: it was largest when the patch included more context (Result 3) and disappeared without pretraining (Result 4), so the ranking depends on the conditions under which it is measured.

*Sensitivity to experimental design, a tentative observation (Result 2).* Of the four perturbations we applied to the reference condition, ResNet-50 degraded significantly under two (added calcifications to both training and test data, full fine-tuning) while no significant change was detected for DeiT under either. If the practical question is which architecture is easier to deploy without extensive per-task tuning, such an asymmetry, if confirmed, would matter more than a few accuracy points at a single operating point. We are careful not to over-interpret it: with five replications the non-significant results for DeiT are underpowered, and a larger replication budget could yet reveal smaller degradations. The pattern does not extend to all changes. Enlarging the crop helped both models (Result 3), and removing pretraining harmed both (Result 4), with DeiT losing significantly in accuracy. If DeiT is indeed less sensitive to design, this stability may depend on the pretrained representation rather than on the architecture alone, which our experiments cannot separate.

*Spatial context and patch extraction (Result 3).* The preprocessing results deserve an emphasis independently of the architecture question. Moving from Experiment 2 to Experiment 4 changed nothing but the size of the crop window and produced a greater improvement for DeiT (6.4 accuracy points) than switching from ResNet-50 to DeiT produced in Experiment 1 (5.6 points). In a small-data patch-classification setting, decisions about how the patch is cut appear to be at least as consequential as the choice between these two architectures.

We deliberately do not attribute DeiT's behavior to self-attention capturing global dependencies, although that is the standard explanation and would be convenient here. Nothing in our experiments measures where either model looks. The Experiment 4 result is at least consistent with context mattering, but it is equally consistent with a wider field of view helping any model, since ResNet-50 also improved. Establishing the mechanism requires an interpretability analysis – attention rollout for DeiT alongside Grad-CAM [16] for the CNN – which we did not perform.

*The role of pretraining (Result 4).* Removing pretraining harmed both models, and the transformer's accuracy advantage over ResNet-50 (Result 1) disappeared. Matsoukas et al. [8] found CNNs stronger than transformers when trained from scratch; we detected no significant difference in accuracy, but ResNet-50's near-collapse to the benign class makes accuracy a poor basis for this comparison, and neither model was usable. Together with Results 1 and 2, this suggests that at this dataset size, pretraining matters more than the choice between the two architectures: it is the only change that substantially degraded both.

## VI. LIMITATIONS OF THE STUDY

Several limitations bound the conclusions above, and we state them explicitly rather than leaving them to be inferred.

*Absolute performance.* No configuration exceeded 0.74 accuracy or 0.71 recall. Several configurations exceed the majority-class baseline by only one to four accuracy points, a margin that may not be meaningful given the limited test-set size; only E4 and, more weakly, E1 for DeiT-base separate clearly from it on balanced accuracy. A second reader that misses roughly three in ten malignant lesions is not clinically

usable, and the gap to published results on the same dataset – for example, the 0.88 per-image AUC of [12] – is large. Three of our design choices plausibly account for much of it: the absence of any data augmentation on a training set of roughly 1 200 patches; a hyperparameter grid restricted to two dimensions, with no weight decay, scheduler, or layer-wise learning rates; and a frozen backbone in four of six experiments. These are the first things we would change.

*Statistical power and test set size.* Five replications give the $t$-tests low power, and the test sets contain 354, 363 and 649 patches. Non-significant results should be read as inconclusive. Two improvements would substantially strengthen the analysis without new architectures: reporting AUC, which is threshold-free and standard in this literature, and comparing paired predictions with McNemar's test [17] rather than comparing summary statistics across independent runs.

*Selection on the test split.* All six experiments were evaluated on the same test split, and hyperparameter configurations were chosen by mean accuracy across the five replications. Design decisions (notably the retention of ResNet-50 after Experiment 1) were also made after observing the test results. The reported values should therefore be regarded as optimistically biased by an amount we did not quantify, and the cross-experiment comparisons as exploratory.

*Architecture coverage.* One transformer family and two CNN families were tested, all with $224 \times 224$ input. EfficientNet-B7 was run far below its design resolution. No modern convolutional architecture such as ConvNeXt [18] was included, which is a material omission given the argument in that work that much of the reported ViT advantage is attributable to training recipes rather than to attention. The capacity mismatch between DeiT-base (86 M) and ResNet-50 (25.6 M) is also not addressed; a compute-matched pair such as DeiT-small would be the appropriate control.

*Task framing and data.* Patches are centered on annotated ROIs, so the models are evaluated as second readers on pre-localized findings, and the numbers are not comparable to whole-image or screening-level systems. Data are scanned film mammograms from a single collection, and the generalization to full-field digital mammography is not tested. Experiment 3 changes the test population as well as the training data, so it compares related rather than identical tasks.

*Validation protocol.* The train/validation division was made at the patch level, so CC and MLO views of the same lesion, or different lesions of the same patient, can fall into both subsets. Therefore, validation loss is likely to underestimate the generalization error. This affects early stopping and checkpoint selection and may favor configurations that memorize patient-specific appearance. A patient-grouped validation split would eliminate this; the test split itself is patient-disjoint.

## VII. Conclusion

We compared ResNet-50 and DeiT-base, with EfficientNet-B0 and B7 as additional baselines, on patch-based benign-versus-malignant classification of CBIS-DDSM lesions between six controlled conditions, reporting five replications per configuration with confidence intervals and $t$-tests.

Within the limits of an exploratory analysis, the study yields four results. First, with ImageNet pretraining, DeiT reached higher accuracy than ResNet-50 in all five pretrained experiments, but its advantage over the strongest CNN tested, EfficientNet-B7, was confined to accuracy. Second, ResNet-50 degraded significantly when calcifications were added and when all layers were unfrozen, while no significant degradation was detected for DeiT; a direct test did not establish that the two architectures respond differently, so this remains a hypothesis. Third, spatial context was The largest improvement we identified: enlarging the crop to 448 × 448 pixels before downsampling improved both architectures, gave the best result in the study, and changed DeiT's accuracy more than switching from ResNet-50 to DeiT did in Experiment 1. Fourth, without pretraining, both models performed poorly, no significant difference in accuracy was detected between them, and they failed in opposite directions. Taken together, the results suggest that in this small-data setting, pretraining and the way the patch is cut matter at least as much as the choice between the two architectures.

We also emphasize what the data do not support. The results do not establish that attention-based modeling is preferable in mammography as a general matter, they do not identify a mechanism for the pattern observed for DeiT, and they do not approach clinically usable sensitivity. Future work should add data augmentation, a patient-grouped validation split and an independent held-out test set, a modern convolutional baseline and a capacity-matched transformer; report AUC with paired significance testing over more replications; and pair the comparison with an interpretability analysis capable of testing the global-context hypothesis directly.

## Acknowledgment

This publication was supported by the European Union (EU) Research Program DIGITAL EUROPE through the TEF HEALTH-Testing and Experimentation Facility for Health, Artificial Intelligence (AI) and Robotics under Contract 101100700.

## References

[1] J. Kim, A. Harper, V. McCormack, H. Sung, N. Houssami, E. Morgan, M. Mutebi, G. Garvey, I. Soerjomataram, and M. M. Fidler-Benaoudia, "Global patterns and trends in breast cancer incidence and mortality across 185 countries," *Nature Medicine*, vol. 31, no. 4, pp. 1154–1162, 2025.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[3] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 97, 2019, pp. 6105–6114.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[5] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences*, vol. 13, no. 9, p. 5521, 2023.

[6] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, R. Aoyama, N. Teraya, A. Bolatkan, N. Shinkai *et al.*, "Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review," *Journal of Medical Systems*, vol. 48, no. 1, p. 84, 2024.

[7] M. Spijker, "Comparison of convolutional neural network and vision transformer for breast cancer classification in mammography," MSc thesis, Tilburg University, School of Humanities and Digital Sciences, Tilburg, The Netherlands, 2022, available: https://arno.uvt.nl/show.cgi?fid=161704.

[8] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it time to replace CNNs with transformers for medical images?" *arXiv preprint arXiv:2108.09038*, 2021.

[9] M. Cantone, C. Marrocco, F. Tortorella, and A. Bria, "Convolutional networks and transformers for mammography classification: an experimental study," *Sensors*, vol. 23, no. 3, p. 1229, 2023.

[10] G. Ayana, K. Dese, Y. Dereje, Y. Kebede, H. Barki, D. Amdissa, N. Husen, F. Mulugeta, B. Habtamu, and S.-w. Choe, "Vision-transformer-based transfer learning for mammogram classification," *Diagnostics*, vol. 13, no. 2, p. 178, 2023.

[11] S. Paavankumar, R. Karthik, G. Idayachandiran, P. P. D. Sri, and T. Illakiya, "Classification of benign and malignant breast lesions in mammograms using dense-unified multiscale attention network and data-efficient image transformers," *The European Physical Journal Special Topics*, vol. 234, no. 15, pp. 4337–4355, 2025.

[12] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific Reports*, vol. 9, no. 1, p. 12495, 2019.

[13] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, no. 1, p. 170177, 2017.

[14] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 10 347–10 357.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[16] 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," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[17] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[18] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 976–11 986.