Classification of Suspicious Microcalcification Clusters: Applying Vision Transformers and Mammography-Specific Pretraining

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

Abstract*—Background/Motivation: Clusters of microcalcifications are among the most difficult findings in mammography because the descriptive rules to mark a cluster as suspicious are not definitive and the diagnosis is confirmed only by biopsy. Convolutional neural networks remain the default choice for this task, while newer architectures such as Vision Transformers and models pretrained by self-supervised learning have not yet been evaluated on the biopsy-confirmed binary classification of these clusters. Methods: We fine-tuned two Vision Transformers, a hybrid convolutional-transformer architecture, and a recent mammography foundation model, and compared them with the ResNet101 architecture from our previous work. In the first experiment, patches of 674×674 pixels were resized to 224×224 pixels; in the second, they were cropped directly at 224×224 pixels. Dataset: Patches with biopsy-confirmed benign and malignant microcalcifications were extracted from the CBIS-DDSM database, and an external malignant-case sensitivity evaluation from CSAW-CC was created to evaluate the models on images acquired with a different technology. Results: None of the newer models exceeded the accuracy of 0.705 previously reported for ResNet101 on CBIS-DDSM; the highest accuracy among them was 0.65. Since the ResNet101 result was not reproduced under the present protocol, this comparison is indicative rather than controlled. VersaMammo improved substantially when patches were cropped at native resolution instead of being resized; changes for the other models were small and, given a single training run and small validation sets, should not be over-interpreted. Native-resolution patches also led to markedly more balanced sensitivity and specificity, suggesting that data preparation has a substantial influence on model behavior. On the external malignant-case sensitivity evaluation, they recognized between 52 and 62 percent of the malignant clusters. Conclusion: The newer architectures did not outperform the previously reported convolutional network, and patch preparation had a substantial effect on model behavior, suggesting that data-related factors may limit performance more than the choice of architecture.

I. Introduction

A. Background

Breast cancer is one of the most common cancers among women and mammography remains the main technique used for its early detection. Because mammography is based on X-ray (ionizing) radiation, screening programs are usually limited to a defined age range, most often between 50 and 70 years, when the incidence of breast cancer is highest [1].

Mammograms can show several types of findings related to breast cancer. Asymmetries are usually recognized by comparing an image with the patient's previous examinations, whereas masses appear as round, dense structures that are often palpable. A more difficult case is presented by calcifications, and in particular microcalcifications, which are not palpable and do not form a dense visible structure on the image. Microcalcifications are usually found in clusters and it is often difficult for radiologists to decide whether a given cluster is benign or malignant. Some descriptive rules exist to mark a cluster as suspicious, for example, a group of more than five microcalcifications, or an irregular shape, but these rules do not give a strict answer, so the final diagnosis is normally confirmed only by biopsy [2].

B. Related Work

Due to this diagnostic difficulty, deep learning (DL) and convolutional neural networks (CNNs), in particular, have become a common approach for classifying microcalcification clusters. The authors of [3] proposed a computer-aided diagnosis system for breast cancer that combined CNN-based feature extraction with a support vector machine (CNN–SVM) and with a k-nearest-neighbor classifier (CNN–KNN), reaching an accuracy of 98.79% for the CNN–SVM model and 96.49% for the CNN–KNN model. In [4], the authors compared several CNN architectures - ResNet50, ResNet101, and five variants of EfficientNet (B0–B4) - with a dataset of 3,674 microcalcification images taken from mammograms with biopsy-confirmed diagnosis (BI-RADS 1–2 for benign, BI-RADS 5–6 for malignant). All variants of EfficientNet performed significantly better than the ResNet models, and the best model, EfficientNet-B3, reached an accuracy of 86.9%, while ResNet-50 and ResNet-101 reached only 64.6% and 62.8%. The authors in [5] built an automated DL pipeline based on a Faster R-CNN detector with a feature-pyramid backbone that locates and classifies microcalcifications. Using an internal dataset of 4,810 mammograms with 6,663 microcalcification lesions confirmed by biopsy, the pipeline reached an overall accuracy of about 72% and a sensitivity of approximately 78% for malignant lesions on the test set. In our previous work [6] we addressed a related task: classifying high-resolution mammogram patches (674×674 pixels) from the CBIS-DDSM and OMI-DB databases as containing or not containing a

suspicious microcalcification cluster, as a first step towards scanning whole mammograms with a sliding window; the best model, ResNet101, reached a validation accuracy of 98.2% (F1-score 0.955). The task addressed in the present article, the classification of a cluster as benign or malignant, was studied in [7], where the same convolutional architecture reached an accuracy of 70.5% when trained and validated on CBIS-DDSM alone, with results varying by up to ±1.5 percentage points between training runs, which shows that deciding whether a cluster is malignant is considerably harder than detecting it.

In addition to CNNs, newer architectures, such as Vision Transformers (ViTs) and self-supervised learning (SSL) are becoming available and may offer further improvement or at least a different angle. ViTs have already been used in mammography, but mostly for detecting lesions on mammograms or for classifying all types of breast cancer in general, rather than focusing on a classification of biopsy-confirmed benign/malignant microcalcification clusters. The authors of [8] proposed a dual-branch ViT model that fuses the craniocaudal (CC) and mediolateral oblique (MLO) views to classify microcalcifications into four morphology categories (amorphous, pleomorphic, fine linear/branching, and punctate) on the CBIS-DDSM dataset, reaching an accuracy of 96.8%, but to our knowledge there is no study focusing on binary classification of suspicious clusters using ViTs.

Moreover, no published study has yet applied SSL to the binary, biopsy-confirmed classification of microcalcification clusters. This is not surprising, since both SSL and ViTs typically require considerably more training data than CNNs. However, general-purpose SSL-pretrained models for mammography have recently appeared, for example, VersaMammo [9], a foundation model pretrained on more than 700,000 mammograms.

C. Objective

In this work, we investigate whether these newer architectures can match CNNs on this task. We fine-tune several pretrained ViT models, DeiT, Swin Transformer, and MammoViT, together with VersaMammo, for the binary classification of benign and malignant microcalcification clusters, using biopsy results with patch-based approach as ground truth. Our contribution is to evaluate whether ViTs and SSL are comparable to conventional CNN-based approaches for this specific task.

II. Methodology

A. Mammography Data and Dataset Creation

For the purpose of our study, we decided to use two well-known and publicly available databases: CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography) [10] and CSAW-CC (Cohort of Screen-Aged Women – Case Control) [11]. Both databases mainly contain two types of breast lesions: masses and calcifications. CBIS-DDSM is commonly used for DL tasks, and it was also used in our previous studies discussed in the related work, which allows a comparison with them. However, this database is relatively old and its images were digitized from screen-film mammography rather than acquired directly in digital form. Because microcalcifications are small and have low contrast, this is a relevant limitation for our task. For this reason, we also use the more recent CSAW-CC database, which was collected in Sweden and is one of the largest modern databases of high-quality digital mammograms. It serves as an external malignant-case sensitivity evaluation so that the models trained on digitized screen-film images can also be evaluated on directly acquired digital images. Both databases use a similar annotation format: lesions are marked with binary masks, where the area of the lesion is shown in white against a black background, which gives the exact location of the microcalcification cluster or mass.

From both databases, we selected only the findings with microcalcifications, which were then divided into two classes, benign and malignant. In both cases, class labels are based on histopathological results, so every finding used in this study was confirmed by biopsy. In CBIS-DDSM, findings annotated as benign without callback were grouped together with the benign class. The annotated findings in CSAW-CC are biopsy-proven cancers, so the external malignant-case sensitivity evaluation contains only malignant clusters and, therefore, the reported results describe sensitivity rather than overall accuracy.

Two training datasets were created for the individual experiments. For the first, patches of 674×674 pixels were cropped from the mammograms at the positions given by the binary masks and then resized to 224×224 pixels, so that they can be used as input for the vision transformers. This dataset contains data from CBIS-DDSM only and allows a comparison with our previous study. For the second, the patches were cropped directly at 224×224 pixels, which preserves the original pixel resolution of the findings instead of downsampling them. If the cluster of microcalcifications extends the patch, multiple patches from one sample are created. This dataset is again based on CBIS-DDSM only. In addition, an external malignant-case sensitivity evaluation was created from CSAW-CC in the same way as the second dataset. This set is never used for training and serves only to evaluate how the models generalize to images from a different acquisition technology.

The remaining steps were identical for all datasets. The training data were augmented by rotations of 90, 180, and 270 degrees and by flipping the images. The CBIS-DDSM data were divided into a training set and a validation set following the official distribution provided with the database, which is defined at the patient level, so images of one patient never appear in both sets. The numbers of patches in each subset are given in Table I.

B. Models

This study compares four recent models - two pure Vision Transformers, one hybrid CNN-transformer architecture, and one mammography foundation model pretrained with SSL - against a conventional **convolutional neural network**. The result of ResNet101, which we decided to use as a reference

TABLE I. NUMBER OF PATCHES IN SUBSET

Dataset	Subset	Amount
CBIS-DDSM 674×674	Train (augmented)	3192
	Validation	225
CBIS-DDSM 224×224	Train (augmented)	28680
	Validation	1044
CSAW-CC 224×224	External evaluation	739

point, is taken from our previous work [7] and was not retrained for this study. It was obtained on CBIS-DDSM microcalcification findings using the same official patient-level train/test split, but with 674×674 patches used at their native resolution without resizing, ImageNet-pretrained weights fine-tuned with the Adam optimizer (learning rate 1e-6) and class weighting, and without data augmentation. The reported value (70.5%) corresponds to the best of 40 training epochs on a CBIS-DDSM validation set of 241 patches (74 malignant, 167 benign), which differs from the validation sets used here. Because it was not produced under the protocol described here, and in particular was never evaluated on the native-resolution 224×224 patches of Experiment 2, it should be read as a historical reference point rather than a controlled baseline. A matched re-evaluation is left for future work.

CNNs remain a strong choice for image-based tasks because their convolutional filters process an image locally and hierarchically: the first layers detect simple characteristics such as edges and textures, and deeper layers combine them into increasingly complex patterns and structures. This local and hierarchical way of processing an image lets CNNs learn efficiently even from a relatively small number of training images, which is one of the main reasons they have remained the default architecture for medical image analysis for many years. However, newer architectures have outperformed CNNs on many tasks in recent years, so we consider it important to test them as well, even though this does not mean they are automatically the better choice for every task.

Vision Transformers are built on the original Transformer architecture and its attention mechanism, first developed for natural language processing. The attention mechanism allows a model to weigh how relevant every part of the input is to every other part, instead of only looking at a small, fixed neighborhood like a CNN does. The ViT adapts this idea to images by splitting an image into fixed-size patches, treating each patch as a token and processing the whole sequence of patches with the same self-attention mechanism [12]. Because of this, a ViT can, in principle, relate any two regions of an image directly, giving it a more global view of the image than a CNN. We fine-tune several pretrained ViT models, each using a different way of processing images, to see whether these newer architectures can compete with conventional CNNs.

DeiT (Data-efficient Image Transformer) was developed to train ViTs competitively using only the standard ImageNet-1k dataset, instead of the hundreds of millions of images required by the original ViT. This data efficiency comes mainly from an improved training based on strong data augmentation and regularization; knowledge distillation is then added on top of this as a further, separate improvement. For our task, where a large amount of data is not available, both properties are relevant. DeiT introduces a dedicated distillation token, along with the usual class token, through which the student ViT is trained to reproduce the class predicted by a pretrained CNN ”teacher” model. The authors show that this makes the transformer’s decisions noticeably more similar to those of the convolutional teacher than to those of another transformer trained without distillation, which they attribute to the inductive bias toward local image structures that CNNs have but ViTs do not naturally possess [13].

Swin Transformer (Shifted Window Transformer) addresses a different problem: in a standard ViT, the cost of self-attention grows quadratically with the size of the input image because every patch attends to every other patch. The Swin Transformer therefore computes self-attention only inside local, non-overlapping windows, which makes the cost grow only linearly with the image size. To avoid isolating these windows from each other, their position is shifted between consecutive layers, so neighboring windows become connected, and the receptive field of the model grows with depth. Swin does not compute global attention at any point; instead, the global view is built up gradually. In addition, patches are merged in deeper layers, producing multi-level feature maps similar to those of a CNN and making the model usable as a general-purpose backbone for classification and dense prediction tasks [14].

MammoViT is a hybrid architecture proposed for mammographic image analysis, originally proposed for the four-class BI-RADS classification, which combines a CNN feature extractor with a small Vision Transformer [15]. In the original work, mammograms are first resized to 224×224 pixels and passed through a ResNet50 pretrained on ImageNet, and the resulting 2048-dimensional feature vectors are reshaped into a tensor that is then divided into patches for the transformer. The transformer part therefore does not process the mammogram directly, but only the feature maps produced by the CNN. It consists of a small number of blocks with standard multi-head self-attention, with the number of blocks, attention heads, and the patch size selected by hyperparameter search. We include it in the comparison because it is one of the few transformer architectures designed specifically for mammograms.

Finally, we also included **VersaMammo**, a recent foundation model for mammography [9]. It was trained in two stages on a collection of 706,239 mammograms from 21 sources. In the first stage, a ViT teacher was trained by self-supervised learning on unlabeled mammograms, using a combination of masked image modeling and a contrastive objective. In the second stage, this knowledge was transferred to an EfficientNet-b5 student network, which is the released VersaMammo model, through supervised learning on clinical labels such as BI-RADS category and breast composition, combined with knowledge distillation from the teacher. It is worth pointing out that the released model is therefore a CNN

and not a Vision Transformer, and that it was not trained by self-supervised learning itself; it only inherits the self-supervised representations of its teacher. We include it here not as a ViT, but as a representative of self-supervised pretraining on large amounts of mammographic data, which gives it a favorable starting point for fine-tuning for our task.

III. Results

A. Experimental Setup

Standard classification metrics were used to evaluate the trained models. Because this is a medical task, the two types of error do not have the same consequences, so overall accuracy alone is not sufficient. A false negative means that a malignant cluster is classified as benign, which delays diagnosis and treatment, while a false positive means that a benign cluster is classified as malignant, which leads to unnecessary biopsy and additional stress for the patient, but not to direct harm. Therefore, the models were evaluated using accuracy, sensitivity, specificity, and F1-score. In all experiments, the malignant class is treated as the positive class, so sensitivity expresses how many malignant clusters were correctly recognized and specificity how many benign clusters were correctly recognized.

All models were implemented in PyTorch and trained in Visual Studio Code with the Google Colab extension, on T4, L4, or A100 graphics cards. For ViTs, we used the Swin Transformer Base checkpoint published by the authors[1] and the distilled DeiT Small checkpoint from the timm library[2]. The models differ considerably in size, which should be kept in mind when comparing them: DeiT Small has approximately 22 million parameters, the ResNet101 reference point approximately 44 million, the EfficientNet-b5 backbone of VersaMammo approximately 30 million, and Swin Transformer Base approximately 88 million. All models were trained with the AdamW optimizer except the reference point, which was trained with the Adam optimizer. The patches were converted to three channels, and for DeiT and Swin they were normalized with the mean and standard deviation of the ImageNet dataset, which corresponds to the preprocessing used during their pretraining.

The same two-phase fine-tuning procedure was used for each model. This procedure follows the common practice for adapting pretrained networks to small medical datasets, rather than the original training of DeiT and Swin Transformer, which were designed for training these models from scratch on ImageNet. In the first phase, the backbone of the network was frozen, so its weights were not updated, and only the classification head was trained. The purpose of this step was to bring the linear classifier to a reasonable starting point before the pretrained weights themselves start to change. The first phase typically took 5 to 7 epochs with a higher learning rate of approximately 1e-3, and at the end we kept the head parameters that reached the best accuracy.

[1]https://github.com/microsoft/Swin-Transformer

[2]https://github.com/huggingface/pytorch-image-models

In the second phase, the backbone was unfrozen and the whole model was trained with lower learning rates. Instead of one common value, we used a layer-wise learning rate decay: the layers closer to the input image received a lower learning rate than the layers closer to the head, because the early layers contain general features that we did not want to change much. For the transformer models, the learning rate gradually increased from the first transformer blocks towards the deeper parts of the network, and the classification head was trained with the highest value. The second phase ran for approximately 40 epochs with early stopping, where the patience was set to 10 epochs, so the training was stopped if the accuracy did not improve during the last ten epochs. For the transformer models, we additionally used a learning rate schedule with a linear warm-up during the first 5 epochs, followed by a smooth decrease until the end of the training, which prevented instability at the beginning of training.

B. Experiments

Table II summarizes the results of the first experiment, where patches of 674×674 pixels were resized to 224×224 pixels. None of the four models reached the accuracy of the ResNet101 reference. The best result was obtained by DeiT-S with an accuracy of 0.63, followed by Swin-B with 0.62, MammoViT with 0.60 and VersaMammo with 0.53.

TABLE II. Experiment 1 – resized 674×674 patches

Model	Acc.	Sens.	Spec.	F1
ResNet101† [7]	0.705	0.67	0.71	0.58
DeiT-S	0.63	0.21	0.91	0.31
Swin-B	0.62	0.46	0.72	0.49
MammoViT	0.60	0.66	0.57	0.57
VersaMammo	0.53	0.88	0.29	0.60

†Historical reference; not retrained under the present protocol. Obtained on native-resolution 674×674 patches (no resizing).

More informative than overall accuracy is the balance between sensitivity and specificity, which was very poor for two of the models. DeiT-S reached a specificity of 0.91 but a sensitivity of only 0.21, so it classified most of the clusters as benign and recognized only about one malignant cluster out of five. VersaMammo behaved in the opposite way, with a sensitivity of 0.88 and a specificity of 0.29, which means that it marked most of the findings as malignant. Neither of these operating points is useful in practice, even though the overall accuracy does not look extremely low. Only MammoViT produced a more balanced result, with a sensitivity of 0.66 and a specificity of 0.57, but its accuracy stayed below the ResNet101 reference point as well.

Table III shows the results of the second experiment, where the patches were cropped directly at 224×224 pixels. The accuracy of DeiT-S changed from 0.63 to 0.65, Swin-B from 0.62 to 0.65, VersaMammo from 0.53 to 0.63, and MammoViT from 0.60 to 0.57. Only VersaMammo showed a substantial change; the differences for the other models were small. The ResNet101 result from our previous work [7] is included in Table III as a reference point.

TABLE III. EXPERIMENT 2 – 224×224 PATCHES

Model	Acc.	Sens.	Spec.	F1
ResNet101† [7]	0.705	0.67	0.71	0.58
DeiT-S	0.65	0.73	0.56	0.67
Swin-B	0.65	0.57	0.72	0.61
MammoViT	0.57	0.85	0.30	0.66
VersaMammo	0.63	0.73	0.55	0.66

†Historical reference; not retrained under the present protocol.

The operating points also became considerably more balanced. DeiT-S, Swin-B and VersaMammo reached sensitivities between 0.57 and 0.73 with specificities between 0.55 and 0.72, instead of the extreme values observed in the first experiment. The F1-score of DeiT-S increased from 0.31 to 0.67 and that of Swin-B from 0.49 to 0.61. MammoViT remained strongly shifted towards the malignant class, with a sensitivity of 0.85 and a specificity of 0.30. Even after this improvement, however, the accuracy of the best models remained below the 0.705 of the ResNet101 reference point, and none of the tested architectures outperformed the reference convolutional model.

Finally, the models trained in the second experiment were evaluated on the external malignant-case sensitivity evaluation created from the CSAW-CC database. Because this set contains only biopsy-proven malignant clusters, the accuracy is equal to the sensitivity and the remaining metrics cannot be computed. DeiT-S reached a sensitivity of 0.62, Swin-B 0.60, VersaMammo 0.59, and MammoViT 0.52. For DeiT-S, VersaMammo, and MammoViT these values are lower than the sensitivities measured on the CBIS-DDSM validation set (0.73, 0.73, and 0.85 respectively), while Swin-B was the only model whose sensitivity did not decrease on the external malignant-case evaluation, changing from 0.57 to 0.60. The largest decrease was observed for MammoViT, which lost more than 30 percentage points and missed nearly half of the malignant clusters. It should be noted that a model that classified all findings as malignant would reach an accuracy of 1.00 on this test set, so these values must be interpreted together with the specificity measured on the validation set.

IV. DISCUSSION

The main result of this study is that none of the tested architectures outperformed the conventional convolutional reference point. Compared with the accuracy of 0.705 previously reported for ResNet101 [7], the best of the newer models, DeiT-S and Swin-B, reached 0.65. Given that the reference was obtained under a different protocol and on a different validation set, this indicates that the newer models did not show an advantage, rather than that they are worse. The two ViTs, the hybrid MammoViT architecture and the VersaMammo foundation model, therefore, did not show an improvement over the reference on this task, even though all of them have been reported to perform very well on other image classification problems. This does not mean that these architectures are generally weaker than CNNs but suggests that their advantages do not automatically transfer to the binary classification of microcalcification clusters, which is a very specific task.

We believe that the main reason is the amount of training data available. ViTs do not have the inductive bias towards local image structures that convolutional filters provide, so they usually need considerably more training images to learn comparable representations. Our dataset contains a limited number of lesions, expanded to several thousand to tens of thousands of patches by cropping and augmentation per experiment, which is several orders of magnitude less than the datasets on which these models are normally trained. This is consistent with our previous observation that all convolutional architectures we tested performed within about 1.5 percentage points of each other [7], and with the data being the main limiting factor, although our experiments do not test this directly.

The comparison between the two experiments suggests that the way the patches are prepared, together with the resulting dataset characteristics, may have a substantial effect on model performance. When the patches were cropped directly at 224×224 pixels instead of being resized from 674×674 pixels, VersaMammo improved clearly, from 0.53 to 0.63, while the accuracy of the other models changed by only 2–3 percentage points, although their operating points became more balanced. The effect can be expected to be stronger for CBIS-DDSM than for modern digital databases, because its images were digitized from screen-film mammograms and already contain less detail before any resizing is applied.

The limited benefit of VersaMammo deserves comment, given that it was pretrained on more than 700,000 mammograms. Its pretraining operated on whole mammograms: the breast region was cropped and resized to between 224×224 and 518×518 pixels, and the supervised stage used image-level labels such as BI-RADS category and breast composition [9]. At this scale, an individual microcalcification occupies at most a few pixels, so the learned representations are oriented towards the global appearance of the breast rather than the morphology of individual calcifications, whereas our patches show a small region at a much higher magnification. This mismatch in scale may explain why VersaMammo gained the most when the native resolution was preserved in Experiment 2. CBIS-DDSM is among the datasets used to develop VersaMammo, so its weak performance cannot be explained simply by unfamiliarity with digitized screen-film images; and because the authors of [9] used its own patient-level data split, we cannot exclude that some of our validation images were seen during its pretraining, which, if anything, would favor VersaMammo. Its extreme operating point in Experiment 1, with a sensitivity of 0.88 and a specificity of 0.29, further suggests that part of its weak result reflects the placement of the decision threshold rather than missing features. Finally, the authors of VersaMammo report that replacing its convolutional student network with a ViT reduced performance, which they attribute to the importance of modeling local texture; this is consistent with our observation that the transformer models did

not exceed the convolutional reference on a task that depends on fine local detail.

An additional observation concerns the balance between the two types of errors. In the first experiment, two models settled at operating points that would be unusable in practice: DeiT-S reached a specificity of 0.91 with a sensitivity of only 0.21, and VersaMammo the opposite combination of 0.88 sensitivity and 0.29 specificity. Their overall accuracy of 0.63 and 0.53 does not reveal how extreme this behavior is, which confirms that accuracy alone is not a sufficient metric for this task. In the second experiment, the operating points became much more balanced for three of the four models, but MammoViT remained strongly shifted towards the malignant class, with a sensitivity of 0.85 and a specificity of 0.30. Since all of our metrics were computed at a fixed decision threshold of 0.5, part of these differences reflects where the threshold falls rather than how well the models separate the two classes, and optimizing the threshold on a held-out portion of the training data, not on the evaluation set, is a straightforward improvement for future work.

The external evaluation from CSAW-CC shows a moderate loss of performance on data acquired with a different technology. Three of the four models detected fewer malignant clusters than on the CBIS-DDSM validation set, and the largest decrease was observed for MammoViT, whose sensitivity decreased from 0.85 on the CBIS-DDSM validation set to 0.52 on the external malignant-case evaluation. Swin-B was the only model whose result did not decrease. These values have to be interpreted with caution, because a model that classified every finding as malignant would reach an accuracy of 1.00 on this test set; they are informative only in combination with the specificity measured on the validation set. Even with this limitation, the results are in line with our earlier finding that models trained on digitized screen-film mammograms do not transfer well to full-field digital mammography [7], and suggest that, at least in our fine-tuning setting, large-scale mammography pretraining did not close this gap.

Our study has several limitations. The ResNet101 result was taken from previous work and not retrained under the present protocol, so the comparison is indicative rather than controlled. Each configuration was trained once. The two experiments differ in resolution, number of patches, and validation set, so the effect of resolution cannot be isolated. The validation set was also used for model selection and early stopping, so the reported values may be optimistic; the same applies to the reference result from [7]. Evaluation was performed at patch level; in Experiment 2, correlated patches from the same lesion may be counted several times and larger lesions weighted more heavily. All metrics were computed at a fixed threshold of 0.5, and threshold-independent metrics (ROC-AUC, PR-AUC) and calibration were not assessed. Training used only digitized screen-film data, and the external set contains only malignant findings, so specificity could not be verified externally. CBIS-DDSM labels include findings with very few microcalcifications, introducing label noise.

Future work should therefore focus primarily on the data rather than on further architectures. Adding databases with biopsy-confirmed benign and malignant microcalcifications at the lesion-level would allow both classes to be represented in more than one acquisition technology, which is the main condition for a reliable generalization evaluation. Repeated training runs with several random seeds and optimization of the decision threshold should also be part of any further comparison of this kind. Finally, ResNet101 should be retrained under the present protocol to allow a controlled comparison.

V. Conclusion

In this study, we evaluated whether Vision Transformers and models pretrained by self-supervised learning can compete with conventional convolutional networks in the binary classification of biopsy-confirmed microcalcification clusters. We fine-tuned DeiT-S, Swin-B, the hybrid MammoViT architecture, and the VersaMammo foundation model, and compared them with the ResNet101 reference point from our previous work on patches from the CBIS-DDSM database.

None of the tested models exceeded the reference point. The best results were obtained by DeiT-S and Swin-B with an accuracy of 0.65, compared to 0.705 for ResNet101. The differences between the four newer models were small and, given that only a single training run was performed for each configuration, they should not be interpreted as a reliable ranking. The more informative result is that none of them showed an advantage over the previously reported convolutional result.

The results also differed considerably between the two patch-preparation settings. When the patches were cropped directly at 224 by 224 pixels instead of being resized from 674 by 674 pixels, VersaMammo improved substantially (0.53 $\rightarrow$ 0.63). The changes for the other three models were small (2–3 percentage points), although the operating points of DeiT-S and Swin-B became markedly more balanced. This is consistent with the hypothesis that downscaling removes fine detail relevant to individual microcalcifications, although in our design the effect of resolution cannot be separated from the larger number of training and validation patches in Experiment 2.

The evaluation also showed that the overall accuracy is not sufficient for this task. Two models reached accuracy values that did not look unreasonable while classifying most findings into a single class, and therefore the balance between sensitivity and specificity has to be reported together with accuracy. On the external data from the CSAW-CC database, three of the four models recognized fewer malignant clusters than on the validation data, indicating that the transfer between digitized screen-film and full-field digital mammography remains an open problem, even for a model pretrained on hundreds of thousands of mammograms.

Overall, our results suggest that, in our setting, data-related factors limit performance more than the choice of architecture. More progress will most likely require databases that provide biopsy-confirmed benign and malignant microcalcifications annotated at the level of individual lesions and acquired with

more than one imaging technology, rather than the application of further architectures to the data that are currently available.

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] T. Hunger, E. A. Nekolla, E. Wanka-Pail, K. S. Winter, and G. Brix, "Extending the age range in mammography screening: A benefit-risk assessment from a radiation protection perspective," in *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Georg Thieme Verlag KG, 2025.

[2] H. Liu, Y. Chen, Y. Zhang, L. Wang, R. Luo, H. Wu, C. Wu, H. Zhang, W. Tan, H. Yin *et al.*, "A deep learning model integrating mammography and clinical factors facilitates the malignancy prediction of bi-rads 4 microcalcifications in breast cancer screening," *European Radiology*, vol. 31, no. 8, pp. 5902–5912, 2021.

[3] H. Alsajer, "Classification of microcalcifications in breast cancer using hybrid deep learning techniques, and image analysis," in *Computer Science On-line Conference*. Springer, 2025, pp. 160–171.

[4] W.-C. Shia, "Deep learning-based classification of benign and malignant breast microcalcifications in mammography," *Scientific Reports*, vol. 15, no. 1, p. 42768, 2025.

[5] Q. Lin, W.-M. Tan, J.-Y. Ge, Y. Huang, Q. Xiao, Y.-Y. Xu, Y.-T. Jin, Z.-M. Shao, Y.-J. Gu, B. Yan *et al.*, "Artificial intelligence-based diagnosis of breast cancer by mammography microcalcification," *Fundamental research*, vol. 5, no. 2, pp. 880–889, 2025.

[6] A. Mračko, A. Gábrišová, D. Narova, I. Cimrák, and L. Vanovčanová, "Detection of suspicious microcalcification clusters in mammography: A patch classification approach," in *2025 38th Conference of Open Innovations Association (FRUCT)*. IEEE, 2025, pp. 231–239.

[7] A. Mracko, I. Cimrák, L. Vanovcanová, and V. Lehotská, "Deep learning in breast calcifications classification: Analysis of cross-database knowledge transferability." in *BIOSTEC (1)*, 2024, pp. 527–535.

[8] S. Elumalai, S. Rajendran, and M. Khalid, "Breast cancer classification based on microcalcifications using dual branch vision transformer fusion," *Scientific Reports*, vol. 16, no. 1, p. 4249, 2025.

[9] F. Huang, J. Zhu, Y. Yu, Y. Xie, Y. Guo, Q. Kong, M. Wu, X. Jiang, S. Yang, J. Ma *et al.*, "A versatile foundation model for ai-enabled mammogram interpretation," *arXiv preprint arXiv:2509.20271*, 2025.

[10] 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.

[11] F. Strand, "Csaw-cc (mammography)–a dataset for ai research to improve screening, diagnostics and prognostics of breast cancer (version 1)," *Karolinska Institutet*, vol. 10, 2022.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention. arxiv 2020," *arXiv preprint arXiv:2012.12877*, vol. 2, no. 3, 2020.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.

[15] A. G. M. Al Mansour, F. Alshomrani, A. Alfahaid, and A. T. M. Almutairi, "Mammovit: A custom vision transformer architecture for accurate birads classification in mammogram analysis," *Diagnostics*, vol. 15, no. 3, p. 285, 2025.