Mammography classification with multi-view deep learning techniques: investigating graph and transformer-based architectures

Medical Image Analysis 99 (2025) 103320

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures

Francesco Manigrasso [a], Rosario Milazzo [a], Alessandro Sebastian Russo [a], Fabrizio Lamberti [a], Fredrik Strand [b,c], Andrea Pagnani [d], Lia Morra [a,*]

[a] Politecnico di Torino, Dipartimento di Automatica e Informatica, Corso Duca degli Abruzzi 24, 10129, Turin, Italy
[b] Department of Oncology-Pathology, Karolinska Institute, Stockholm, Sweden
[c] Department of Breast Radiology, Karolinska University Hospital, Stockholm, Sweden
[d] Politecnico di Torino, Dipartimento di Scienza Applicata e Tecnologia, Corso Duca degli Abruzzi 24, 10129, Turin, Italy

ARTICLE INFO

Keywords:
Mammography
Visual transformers
Computer-aided diagnosis

ABSTRACT

The potential and promise of deep learning systems to provide an independent assessment and relieve radiologists' burden in screening mammography have been recognized in several studies. However, the low cancer prevalence, the need to process high-resolution images, and the need to combine information from multiple views and scales still pose technical challenges. Multi-view architectures that combine information from the four mammographic views to produce an exam-level classification score are a promising approach to the automated processing of screening mammography. However, training such architectures from exam-level labels, without relying on pixel-level supervision, requires very large datasets and may result in suboptimal accuracy. Emerging architectures such as Visual Transformers (ViT) and graph-based architectures can potentially integrate ipsi-lateral and contra-lateral breast views better than traditional convolutional neural networks, thanks to their stronger ability of modeling long-range dependencies. In this paper, we extensively evaluate novel transformer-based and graph-based architectures against state-of-the-art multi-view convolutional neural networks, trained in a weakly-supervised setting on a middle-scale dataset, both in terms of performance and interpretability. Extensive experiments on the CSAW dataset suggest that, while transformer-based architecture outperform other architectures, different inductive biases lead to complementary strengths and weaknesses, as each architecture is sensitive to different signs and mammographic features. Hence, an ensemble of different architectures should be preferred over a winner-takes-all approach to achieve more accurate and robust results. Overall, the findings highlight the potential of a wide range of multi-view architectures for breast cancer classification, even in datasets of relatively modest size, although the detection of small lesions remains challenging without pixel-wise supervision or ad-hoc networks.

1. Introduction

Mammography is the main imaging modality for breast cancer screening, and hence one of the most important tools available to reduce breast cancer mortality (Broeders et al., 2012; Morra et al., 2015). Given the high reading volumes, combined with a well-defined diagnostic task and a fairly standardized acquisition process, screening mammography is an ideal candidate for automated or semi-automated reading. The promise of deep learning systems to provide an independent assessment that could relieve the burden on radiologists has been recognized in several recent studies (Rodríguez-Ruiz et al., 2019; Kyono et al., 2018; Dembrower et al., 2020). However, designing deep learning systems for mammography remains a challenging problem due to a number of issues: with a cancer prevalence lower than 1%, mammography screening is the typical needle-in-the-haystack search problem, requiring very large and enriched datasets to achieve high performance (Wu et al., 2020; Schaffter et al., 2020); the need to process high-resolution images (Wu et al., 2020); and the need to combine information at multiple scales (Shen et al., 2021b; Pinto Pereira et al., 2009), and from multiple views (Van Schie et al., 2011; Samulski and Karssemeijer, 2011; Perek et al., 2018; Famouri et al., 2020; Ren et al., 2021).

One of the most complete approaches for automated processing of screening mammography are so-called *multi-view* architectures that combine information from the four views typically included in a screening exam and produce an exam-level classification score indicating,

* Corresponding author.
E-mail address: lia.morra@polito.it (L. Morra).

https://doi.org/10.1016/j.media.2024.103320
Received 5 December 2023; Received in revised form 20 June 2024; Accepted 19 August 2024
Available online 2 September 2024

e.g., the probability of the exam containing a cancer. Multi-view architectures are able to perform both *ipsi-lateral* and *contra-lateral* analysis: the former searches for lesions on either breast by combining the cranio-caudal (CC) and medio-lateral-oblique (MLO) views to compensate for the effects of high breast density and tissue superposition (Sacchetto et al., 2016; Wei et al., 2011; Van Gils et al., 1998; Ren et al., 2021; Samulski and Karssemeijer, 2011); the latter combines information from both breasts and can, for example, detect asymmetries that would not emerge if individual views were separately considered (Rangayyan et al., 2007).

Another promising aspect of these architectures is that, in principle, they can be trained from exam-level labels, bypassing the need to acquire costly pixel-level supervision. In the first large-scale attempt to train deep neural networks (DNNs) from weak image-level labels, the DREAM challenge, DNNs trained using strongly annotated external data significantly outperformed DNNs trained on image labels alone (Schaffter et al., 2020). With advances in deep learning architectures, breast cancer detection from image-level supervision appears to be substantially improved (Wu et al., 2020; Shen et al., 2021b). Yet, while most state-of-the-art solutions for mammography are based on deep convolutional neural networks (CNNs), and especially residual networks, alternative deep architectures are emerging in the literature. On the one hand, Visual Transformers (ViT) are outperforming CNNs in several medical and non-medical tasks (Dosovitskiy et al., 2020; He et al., 2022; Xu et al., 2022; Matsoukas et al., 2022). Compared to CNNs, transformers are particularly promising for three properties: (i) optimal computational allocation on relevant regions of the image (pixels are not all equal paradigm), (ii) optimal semantic encoding; (iii) relating spatially distant semantic features through the so-called self-attention mechanisms (Dosovitskiy et al., 2020). The ability to model long-range dependencies through cross-view attention allows to naturally integrate information from multiple mammography views (van Tulder et al., 2021).

On the other hand, several architectures have been proposed to explicitly mimic the radiologist's interpretation pattern for both ipsi–lateral and contra–lateral analysis (Ren et al., 2021; Du et al., 2019; Liu et al., 2021b; Zhang et al., 2021; Yang et al., 2021). These architectures integrate information from ipsi-lateral views to solve the tissue superposition problem, and thus search for structures that are both spatially co-located and with similar visual features as potential lesion candidates (Wei et al., 2011; Samulski and Karssemeijer, 2011; Van Gils et al., 1998; Ren et al., 2021; Yang et al., 2021). Contra-lateral analysis, on the other hand, aims at detecting asymmetries and dissimilarities between the two breasts. These architectures are typically composed of a convolutional backbone followed by a module that is more apt for relational reasoning, such as graph convolutional networks (GCNs) (Liu et al., 2021b) or relational networks (Yang et al., 2021; Ren et al., 2021).

Direct comparison of architectures with different inductive biases from existing studies is hindered by their different experimental settings, not only in terms of dataset size and composition, but also in terms of task, learning setting, and performance metrics. For instance, van Tulder et al. (2021) evaluated transformers in a dual-view setting, thus performing only ipsi-lateral analysis, whereas CNN-based architectures are often evaluated on the four mammographic views (Wu et al., 2020). Radiologist-inspired architectures were instead evaluated on lesion detection tasks (Liu, 2010). In this paper, we aim to compare multi-view architectures characterized by different inductive biases, trained under a comparable weakly supervised setting. Specifically, our contributions are three-fold:

- we extend existing architectures based on transformers (van Tulder et al., 2021; Matsoukas et al., 2022) and graph convolutional networks (Liu et al., 2021b) to handle four mammographic views;
- we introduce a new transformer-based architecture with ipsi-lateral and contra-lateral cross-view attention;
- we evaluate different architectures not only performance-wise, but also with respect to how they integrate local and global features. Our results suggest that different architectures are complementary in nature, in the sense that they are preferentially sensitive to specific signs, and that breast cancer detection could benefit from their integration even though transformers outperform convolution-based architectures.

The rest of the paper is organized as follows. Section 2 reviews the main architectures that were proposed for exam-level mammography analysis. The architectures explored in our experiments are analyzed in Section 3. Datasets and experimental settings are described in Sections 4 and 5, respectively. Results are presented and discussed in Sections 6 and 7, respectively. Finally, brief conclusions are drawn in Section 8.

2. Related work

Applications of deep learning in mammography have flourished in the past five to ten years: as a result, reviewing all possible architectures and applications would be beyond the scope of this section (Morra et al., 2019; Ou et al., 2021; Jiménez-Gaona et al., 2020). Here we summarize currently available multi-view architectures for breast cancer detection and triage, divided in three broad categories: CNN-based, radiologist-inspired and transformer-based.

2.1. Models based on convolutional neural networks

CNNs are the *de facto standard* in many application domains for medical and non-medical image interpretation. In mammography, recent CNN models have reached performance close to human radiologists, or even outperformed them, in laboratory conditions (Rodríguez-Ruiz et al., 2019; McKinney et al., 2020). Both single-view (Maqsood et al., 2022; Lotter et al., 2021; Samee et al., 2022) and multi-view (Carneiro et al., 2017; Khan et al., 2019; Kyono et al., 2018; Nawaz et al., 2018) architectures have been proposed to detect and classify malignant lesions. However, in many multi-view architectures the image is aggressively downsized in order to reduce the computational and memory footprint of the network. Downscaling is not desirable as many lesions signs are typically only discernible at higher resolution. Wu et al. (2020) show the possibility of training multi-view models for high resolution images from weak image-level labels. They based their architecture on a ResNet-22 backbone modified to cope with high resolution images, and fuse the features from multiple views by simple concatenation. This architecture was chosen as the baseline in our study, with some adjustments to account for the different settings in which the training datasets were acquired.

2.2. Radiologist-inspired models

Several architectures have been designed to more closely mimic the way radiologists integrate information from contra–lateral and ipsi–lateral views (Yang et al., 2021; Ren et al., 2021; Du et al., 2019; Zhang et al., 2021; Liu et al., 2021b). Broadly speaking, these architectures attempt to explicitly match corresponding regions from the different views, based on their geometrical and visual properties, to emphasize either abnormalities that appear consistently in the CC and MLO views, or asymmetries between the left and right breasts. To avoid the need to register the four views, which may be unfeasible due to the effect of compression on soft tissues (Famouri et al., 2020), researchers have sought to introduce additional modules into CNN-based architectures, such as relational networks (Ren et al., 2021) or graph convolutional networks (GCNs) (Zhang et al., 2021; Liu et al., 2021b; Du et al., 2019). In the latter case, a weighted graph models the relationship between local regions from different views.

Numerous GCNs have been proposed in mammography, for example Du et al. (2019) combine a CNN with two graph attention networks,

the first classifying each node to predict ROIs and the second classifying the entire image. Instead, Zhang et al. (2021) use a CNN to extract features and a GCN to learn the relationships between views, each represented as a node in a graph.

A particularly promising approach is the Anatomy-aware Graph convolutional Network (AGN) (Liu et al., 2021b) which introduces two graph-based modules, each dedicated to modeling either ipsi-lateral or contra-lateral analysis. The original AGN architectures takes as input three views – an examined view (e.g., the right CC view), the contra-lateral view (e.g., the left CC view) and the ipsi-lateral view (e.g., the right MLO view) – and produces as output a combined feature map for the examined view. Liu et al. (2021b) demonstrated that an object detector based on the proposed network outperformed standard architectures for mammographic mass detection. In this work, the architecture was extended to support the simultaneous analysis of four views for the task of weakly supervised exam-level classification.

2.3. Transformers-based models

Transformers have revolutionized language modeling, and Vision Transformers (ViTs) are steadily and increasingly outperforming convolutional networks in computer vision. More recently, researchers have started investigating the relative performance of ViTs and CNNs on medical image analysis (He et al., 2022; Matsoukas et al., 2021; Li et al., 2023; Matsoukas et al., 2022). Due to their weaker inductive biases, ViTs can scale up to much larger datasets than their CNN counterparts, but at the same time require a larger amount of training data to achieve desirable performance, which can be a significant challenge in medical image analysis. To address this, many methods have combined convolutional layers with ViTs to improve performance with limited medical images, or have extensively leveraged transfer learning and self-supervised learning to reduce data requirements (He et al., 2022). However, existing research has not yet shown that ViTs outperform CNNs in all scenarios, particularly in low-resolution and few-shot medical image analysis.

On the specific task of mammographic image interpretation, transformer-based models have been compared to CNNs with mixed results, also due wide differences in experimental setups and architectures (He et al., 2022; van Tulder et al., 2021; Miller et al., 2022; Matsoukas et al., 2021, 2022). Most studies have compared ViTs to CNNs on the task of single-view image classification in a transfer learning setting, e.g., when fine-tuning from ImageNet1K. Architectures such as DeiT (Touvron et al., 2021) and Swin Liu et al. (2021a) are the best candidates for high resolution medical images (Matsoukas et al., 2022; Betancourt Tarifa et al., 2023; Cantone et al., 2023; Li et al., 2023). In particular, the latter is the prime candidate due to its hierarchical nature and associated lower computational requirements. However, the performance gap between ViTs and CNNs in typical mammography datasets, such as CBIS-DDSM and OmniDB, was either small (Matsoukas et al., 2022) or in favor of CNNs (Cantone et al., 2023). In Cantone et al. (2023), the hierarchical Swin-v2 transformer was the only architecture that achieved competitive results (that increased with higher input resolution), indicating an advantage in incorporating a locality bias. In a self-supervised setting, Miller et al. (2022) found that pre-training ViTs using the masked autoencoder framework had poor performance compared to pre-training CNNs using contrastive self-supervised techniques.

Larger benefits are potentially associated with the use of ViTs in the multi-view setting, thanks to their ability to model long-range visual relationships. When training transformers on multiple views it is important to find a good trade-off between performance and computational complexity: for instance, for RGB images a reasonable compromise was obtained by employing a unified backbone to which each view is passed separately, for then concatenating the individual output features to perform the final classification (Chen et al., 2001). In the field of mammography, van Tulder et al. (2021) proposed a Cross-View attention layer to perform inter-view feature mixing based on a CNN backbone: this solution, while reaching promising results, does not exploit ViTs to model the content of each breast. On the other hand, Chen et al. (2022) have proposed a multi-view network based on the DeiT transformer that can process up to four low resolution input views, and showed moderate increase in performance compared to CNNs on a small scale dataset of roughly 1K mammograms. Despite these promising results, several research gaps need to be addressed including how transformer-based architectures perform compared to existing architectures; and what is the influence of different transfer and self-supervised settings on their relative performance.

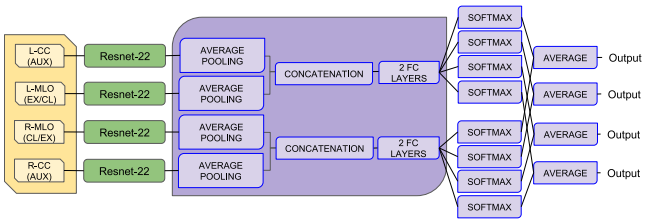


Fig. 1. A schematic representation of the NYU model from Wu et al. (2020). The backbone parameters are shared among images of the same view (CC and MLO), as indicated by the different colors. The loss is calculated from the softmax output, and the predictions of the CC and MLO views are averaged at inference time.

3. Architectures

We formulate the overall problem as a multi-task learning framework with two independent classification tasks: malignant vs. normal, and recalled vs. not recalled. Each breast consists of the standard four views, denoted in the following as $\mathbf{x}_{\mathrm{R-CC}}$, $\mathbf{x}_{\mathrm{L-CC}}$, $\mathbf{x}_{\mathrm{R-MLO}}$ and $\mathbf{x}_{\mathrm{L-MLO}}$. Each breast side (left and right) is associated with binary labels indicating the presence or absence of malignant cancer ($y_{\mathrm{R,m}}$ and $y_{\mathrm{L,m}}$) and whether the image was recalled by the radiologists for further workup ($y_{\mathrm{R,r}}$ and $y_{\mathrm{L,r}}$). It should be noted that the dataset was collected in the context of a European screening program, and the recall status is established based on the consensus of two independent readers. The task is then formulated as predicting the two labels for each side ($\hat{y}_{\mathrm{R,m}}, \hat{y}_{\mathrm{L,m}}, \hat{y}_{\mathrm{R,r}}$ and $\hat{y}_{\mathrm{L,r}}$).

3.1. Baseline

The first architecture consists of the "image-wise" CNN proposed by Wu et al. (2020). The architecture comprises a backbone that maps each view into a fixed–dimension space, and two fully connected layers that convert the representations into the output prediction. Following the original implementation, the weights of the four backbones (L-CC, R-CC, L-MLO, and R-MLO) are shared, and the representations are concatenated to produce four independent predictions for the CC and MLO views. A modified Resnet-22 is used as backbone to compute the embedding space of 256 dimensions for each view, as documented by Wu et al. (2020), to obtain an appropriate threshold between the width and depth of the model with respect to the high resolution of the input image. The final architecture is shown in Fig. 1. Compared to Wu et al. (2020), the classification head was modified to obtain predictions for two classification tasks: malignant vs. non-malignant and recall vs. non-recall.

As in the original architecture, the final prediction ($\hat{y}_{\mathrm{R,m}}, \hat{y}_{\mathrm{L,m}}, \hat{y}_{\mathrm{R,r}}$ and $\hat{y}_{\mathrm{L,r}}$) is determined as the average of the predictions from the two views CC and MLO.

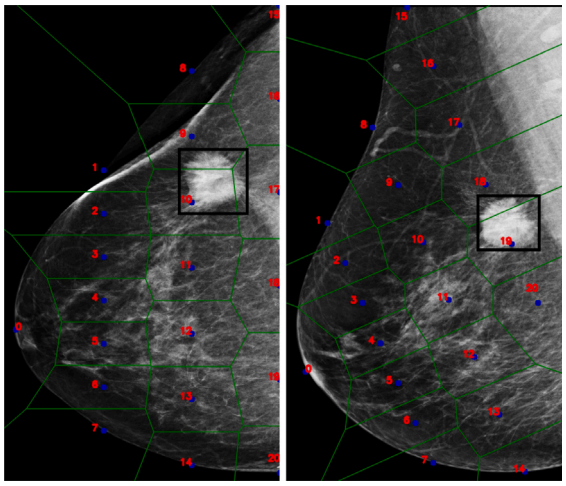


Fig. 2. A representation of pseudo-landmark and the respective tessellation of the same breast in the two projections.

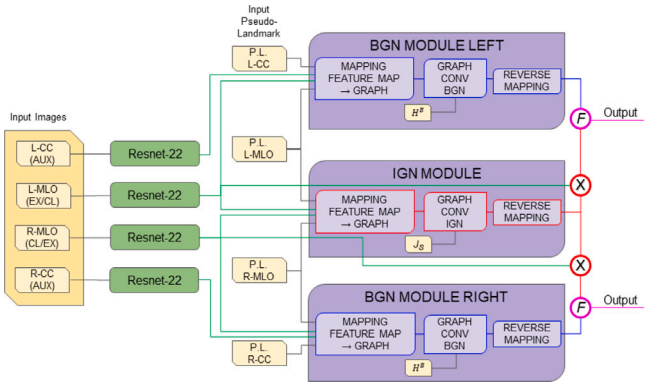


Fig. 3. A representation of the entire AGN4V model. It is worth noting that this architecture requires an additional set of inputs (that is, the pseudo-landmarks and their position), which are used by the IGN and BGN modules to simulate radiologists' analysis.

3.2. AGN4V

The second proposed architecture, named Anatomy-aware Graph Convolutional Network Four Views (AGN4V), is an extension of the AGN introduced in Liu et al. (2021b). The proposed architecture mainly consists of a backbone and two modules based on GCNs, which receive graphs obtained from the feature maps through a specific mapping function. These modules respectively model the geometric relationships of the ipsilateral views (Bipartite Graph convolutional Network or BGN) and the structural similarities between left and right breasts (Inception Graph convolutional Network or IGN). Each node is associated to a (irregular) region of the breast (tessellation), and each region is in turn associated to a point in the image, called in the following pseudo-landmark. An example of pseudo-landmarks with the respective tessellation can be seen in Fig. 2. Both BGN and IGN include a mapping function, which produces a graph encoding the pseudo-landmarks, their features and their geometrical inter-relationships, followed by a Graph Convolutional Network (GCN), and a reverse mapping function which projects the processed graph back to an attention map in the feature space. The mapping function differs between the BGN and IGN modules, which take as input different images and encode different geometric and semantic properties, as further detailed in Section 3.2.2. The main difference with respect to the original architecture is the duplication of the BGN module, with shared weights, in order to obtain the prediction for both sides and both tasks simultaneously. For each side a triplet of images is thus integrated, the examined view (EX) on which the prediction is performed, the contralateral view (CL) and the auxiliary view (AUX), setting the MLO as the examined view and the CC as the auxiliary view. Again, as with the baseline, the loss is calculated using the softmax output, which contains the probabilities of cancer presence and recall for each view examined. The overall architecture can be seen in Fig. 3.

3.2.1. Pseudo-landmark extraction

To extract the pseudo-landmarks, the three guidelines defined in Liu et al. (2021b) were followed:

- Each pseudo-landmark should represent a region with relatively similar positions between the same projections of different breasts;
- Distinct pseudo-landmarks should represent distinct regions of the breast;
- The combination of all pseudo-landmarks should cover the entire breast.

For the extraction of the landmarks from CC views, we opted to start from the only reference point available, namely the nipple. To extract further landmarks, both the longitudinal position of the landmark just calculated and the breast contour were used, so that the final pseudo-landmarks are evenly spaced along both axes. In the MLO views, the position of the pectoral muscle, together with the nipple, was used to extract the landmarks. Starting from the position of the nipple landmark, a line perpendicular to the pectoral muscle is traced to find the intersection between these two, and subsequently the landmarks located on the pectoral muscle are placed at even intervals. Finally, as done for the CC projection using the contour of the breast and, in this case, two lines parallel to the pectoral muscle, the remaining landmarks are placed.

3.2.2. Implementation details

The backbone of AGN4V coincides, in this case, with the ResNet-22 described in Section 3.1. The GCNConv layer was selected to implement the GCNs components for both the IGN and BGN, each consisting of 4 layers. It takes as input a weighted graph in sparse form, i.e., as a pair of arrays, one encoding the graph arcs as pairs of source/target nodes (H^B and J for the BGN and IGN, respectively), and one encoding the corresponding arc weights (W^B and W^I). H^B is the combination of two graphs, the geometric graph, which represents the geometric constraints across views and the semantic graph, which characterizes the semantic similarities between nodes. The two graphs jointly regularize the propagation of ipsilateral information. Instead, J characterizes the relations of nodes across different views: since the bilateral views may not be aligned perfectly due to the inherent geometric distortions, J is reformulated as J_s which links each node to its top nearest neighbors in the contralateral view, in this case s is equal to 2. Further details on the structure of the IGN and BGN modules are defined in Liu et al. (2021b).

3.3. MaMVT

An overview of the proposed Mammography Multi-View Transformer (MaMVT) is shown in Fig. 4. The four input views are separately fed into a shared backbone network. At the halfway-point of the third block of the network a cross-view attention layer (defined in Section 3.3.3) is applied in order to combine the representations of coupled views inside the network, i.e., the corresponding views from the left and right breasts, and the CC and MLO view from each breast. Then, each view feature vector is output from the backbone network and passed through a corresponding classification layer, with shared weights for the CC and MLO views. Based on the findings of Chen et al. (2001), two additional classification layers with shared weights for the left

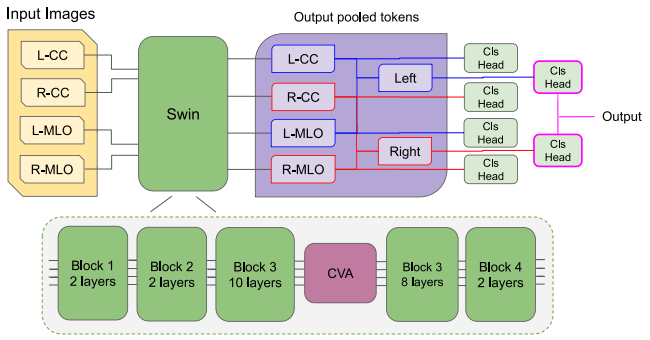


Fig. 4. Representation of the MaMVT architecture used in this work: the four views are passed through a shared Swin backbone, with an additional cross-attention block inserted inside the backbone after the 10th layer of the 3rd block to perform cross-attention between each view. The final output for each view is then passed through a classification layer and used for additional loss computation. The two left and right views are additionally concatenated to obtain a left and right representation as well, which are also passed through a classification layer and are used to perform both loss computation and to obtain the final classification result for the exam.

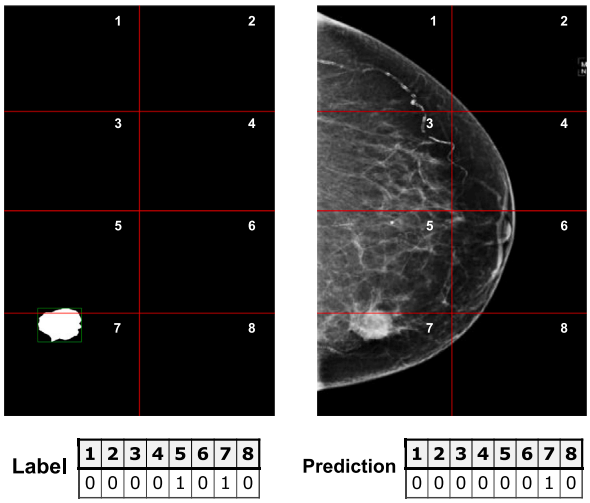


Fig. 5. Simplified example of the patch supervision method: shown on the left is the image mask, split into patches and converted into the label vector below, where each value corresponds to one patch: indices 5 and 7 are set to 1, since their respective patches contain the lesion. Shown on the right is a hypothetical prediction of each image patch following the same structure: in this example, all patches were predicted correctly with the exception of patch 5.

breast and the right breast perform classification on the combinations of L-CC, L-MLO and R-CC, R-MLO, respectively. The latter single view predictions are only used for loss computation as an additional task.

3.3.1. Backbone

One problem with using high-resolution images with regular vision transformers like ViT is the fact that these architectures do not perform any kind of convolutional or pooling operations, thus increasing the computational and memory requirements by processing the image always at its full size. To tackle this issue and based on the findings from Matsoukas et al. (2022), the Swin (Liu et al., 2021a) architecture was selected for the backbone. This architecture aggregates patches at different levels of the network in order to achieve better performance by increasing cross-patch attention, and at the same time decrease both computational complexity and memory footprint by reducing the number of patches. An important difference from ViTs is the lack of a classification token, since classification is instead performed on the combined global pooling of the output patches. Experiments were performed using the Swin-v1 (Liu et al., 2021a) and Swin-v2 (Liu et al., 2022) architectures, generating MaMVT-v1 and MaMVT-v2 respectively.

3.3.2. Patch-level supervision

Taking advantage of the patch-based nature of the transformer architecture, an additional patch-level supervision task was also introduced on each of the four views. Briefly, this additional task entails predicting whether a lesion is present in each patch, taking into account that Swin has a hierarchical structure and that neighboring patches are gradually merged in deeper transformer layers. This weak supervision has minimal computational and memory overhead, only introducing an additional classification head that performs a small number of patch-level classification. The patch-level reference standard is achieved by resizing and splitting the mask into the same number of patches as the output layer of the Swin backbone. The array is then used to generate a one-hot binary vector label of the image that indicates patches that contain (parts of) lesions. The number of patches in the backbone output is ultimately equal to the windows count, that is 12×12 output patches. For the sake of exposition, Fig. 5 shows a simplified example on a reduced number of patches.

3.3.3. Cross-view attention layer

Inspired by previous work by van Tulder et al. (2021), in which a similar cross-view module was placed between the third and fourth layers of a ResNet architecture, the cross-view attention layer is here

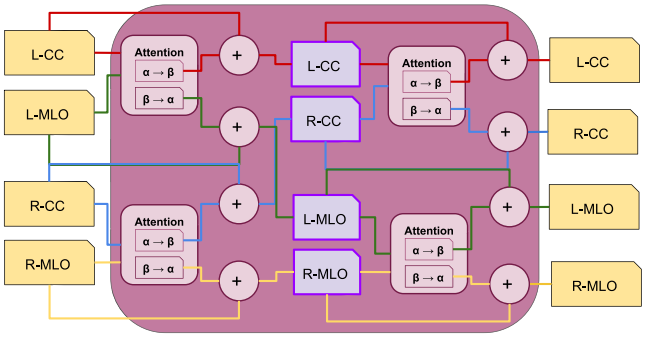


Fig. 6. Side-variant four-view cross-attention module scheme. First, for each side (L-CC and L-MLO, R-CC and R-MLO), the pair-wise attention operations are performed and then added to their respective views. Then the same operation is applied for each type of view (L-CC and L-MLO, R-CC and R-MLO).

placed at a similar depth in the Swin backbone, i.e., at the half-way point of the third block. This layer implements Multi-Head Attention between two views by deriving the query Q matrix of the attention mechanism from a source view α, while the key K matrix and the value V matrix are derived from a target view β. The Multi-Head Attention layer is asymmetrical, and therefore the outcome depends on the views that are selected as source query α and target key–value β, respectively. To compensate for this, multi-head attention is computed in both directions, alternating the target α and source β view: the respective results are then added to the corresponding target view. As an example, given a pair of views such as L-CC and L-MLO, multi-head attention is first computed setting the L-CC view as the target α and the L-MLO view as source β, then setting the L-MLO view as source and the L-CC view as target, and then both output features as summed to the original target (L-CC) features. As shown in Fig. 6, the layer is applied to each combination of views: the combinations belonging to the same side are performed first (L-CC and L-MLO, R-CC and R-MLO), followed by the combinations of the same type of view (L-CC and L-MLO, R-CC and R-MLO), for a total of four times. In this way, both ipsi-lateral and contra-lateral attention is implemented.

Empirically, we observed that standard multi-view architectures are not invariant with respect to the order in which the two sides (left

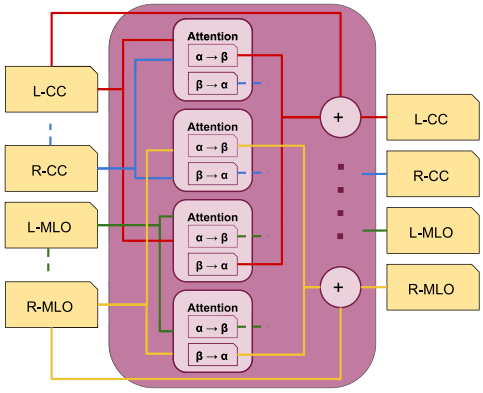


Fig. 7. Side-invariant four-view cross-attention module scheme. All the pair-wise attention operations are performed first, and then added to their respective views. Only the sum operations for the L-CC and R-MLO views are shown for clarity.

and right) are presented, even if both sides share the same backbone weights. Thus, the network may learn to introduce spurious associations between the side and the prediction. This issue is not limited to transformers, but is shared by all architectures: in the case of the baseline, permutation invariance is lost when features are concatenated before being fed to the final fully connected layer. In the case of MaMVT, the properties of the network depends on how cross-view attention is implemented in the backbone. We therefore designed and tested an additional side-invariant version of the Swin-v2 backbone. In this variant, instead of applying a single two-view cross-attention module four times to different CC and MLO pairs, a single four-view cross-attention module is applied on all four views simultaneously: the bi-directional attention operations are performed first on each view pair, and then the respective add operations are performed at the same time for each view, as shown in Fig. 7.

3.4. Pretraining

Exam-level labels, while offering reduce annotation costs, provide a very weak and sparse supervisory signal. Many authors have found benefits by pre-training the network on a more balanced task, such as BI-RADS classification (Wu et al., 2020), which, however, requires additional labels.

In recent years, self-supervised visual representations learning have been introduced as a pre-training mechanism to increase the robustness of the learned feature representation. By learning to associate different transformed versions of the same image, while simultaneously discriminating them from different images, the learned feature representations are invariant with respect to the selected transformations, as demonstrated by Misra and van der Maaten (2019), Maaz et al. (2021), Chen et al. (2021), Jiang et al. (2020).

Our work leverages previous experiments by Miller et al. (2022), who tested four different self-supervised architectures applied to breast cancer classification: SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), SWaV (Varamesh et al., 2020) and Vit-MaE (He et al., 2021). For convolutional and residual networks, BYOL is preferable for patch-level classification tasks, while SWaV is best used as pretraining for full image classification tasks (Miller et al., 2022). In the latter case, the models are pre-trained using tiled patches, and then the features are transferred to the entire mammograms. Briefly, SWaV is a self-contrastive supervised method based on an online learning mechanism: given two augmented versions of the same patch, matching "code" are computed from a given set of prototypes and then swapped to calculate the loss function in order to find matching information in the two augmented versions.

In our experiments, the baseline backbone (modified Resnet22) is pre-trained using SWaV, before being fine-tuned on the full image classification task. For the AGN4V, the backbone is pre-trained with SWaV, fine-tuned with the baseline architecture, and then transferred.

For the transformer-based architecture, we compared two choices of pre-training: ImageNet and a patch-level self-supervised method denoted as PEAC (patch embedding of anatomical consistency), proposed by Zhou et al. (2023). Previous experiments by Miller et al. (2022) found that traditional self-supervised transformer training, based on the concept of the masked autoencoder, brought little benefit to the final classification task. This result can be explained by the different nature of self-supervision in transformers and convolutional networks. Masked autoencoders seek to "fill in the blanks" by predicting the masked patches based on context: however, it is unlikely that the presence of lesions can be predicted based on the surrounding tissue, hence it offers little benefit compared to ImageNet pretraining. On the other hand, PEAC is a contrastive teacher-student framework that enforces both global and local (at the patch level) consistency between two augmented views. It was shown to improve performances over ImageNet pretraining in other medical modalities, such as chest X-rays (Zhou et al., 2023). We selected this framework as it is compatible with the Swin-v1 architecture.

3.4.1. Loss

All three architectures use the NLL loss for both cancer and recall prediction tasks.

For the Baseline, the loss is calculated as the sum of the individual recall and cancer components for left and right, respectively:

$$\mathcal{L} = \mathcal{L}_{LCC} + \mathcal{L}_{RCC} + \mathcal{L}_{LMLO} + \mathcal{L}_{RMLO} \tag{1}$$

Instead, for the AGN4V, the loss is calculated by adding the losses of the two sides for both classes of the examined view, in our case the MLO:

$$\mathcal{L} = \mathcal{L}_{LMLO} + \mathcal{L}_{RMLO} \tag{2}$$

In both cases, Baseline and AGN4V, each $\mathcal{L}_x$ is equal to:

$$\mathcal{L}_x = \mathcal{L}_{cancer_x} + \mathcal{L}_{recall_x} \tag{3}$$

where x = LCC, RCC, LMLO, RMLO. For Transformers, the loss is calculated by combining the losses obtained from the classification head of each separate view, and the two losses obtained from the classification heads of the left and right representations, which are obtained by concatenating the features of the two left and right views, respectively. Finally, the Focal Loss (FL) (Lin et al., 2017) is used for patch-level classification in transformers:

$$\mathcal{L} = \mathcal{L}_{LCC} + \mathcal{L}_{RCC} + \mathcal{L}_{LMLO} + \mathcal{L}_{RMLO} + \mathcal{L}_{Left} + \mathcal{L}_{Right} \tag{4}$$

where each $\mathcal{L}_i$ is equal to:

$$\mathcal{L}_i = \mathcal{L}_{cancer_i} + \mathcal{L}_{recall_i} + \sum_{k}^{n} \mathcal{L}^{k}_{patch_i} \tag{5}$$

where x = LCC, RCC, LMLO, RMLO, k represents the kth patch, and n is equal to the number of patches.

4. Dataset

Experiments were conducted on the Karolinska Cohort of Screen-Aged Women (CSAW) dataset, further enriched with the DDSM dataset as well as synthetically generated lesions to augment the number of cancers available. In this section, more details are given on each of the datasets considered.

4.1. CSAW case-control subset (Karolinska)

The CSAW dataset by Dembrower et al. (2019) is a population-based cohort of women aged 40 to 74 years who were invited for screening between 2008 and 2015 in the Stockholm region, Sweden. From the original dataset, a random sample of 30% of the population was withheld before release and not included in this study. The available CSAW subset includes 8723 women, of which 873 are *cases* diagnosed with cancer during the observation period, and 7850 are *controls*. Overall, the dataset includes 524 women with screen-detected cancer, 217 with interval cancers, and 132 with prior images/other; based on the available annotations, approximately 50% of interval cancers are not visible. To reduce the number of ambiguous cases, all exams considered priors, i.e., examinations in which more than 730 days had elapsed between the date of screening and the date of diagnosis, were excluded; women for which only prior images were available were eliminated. The dataset was split at the individual level into a train/validation/test set with a 70%/10%/20% ratio, stratified by age, case/control, date of diagnosis, and number of exams. Screen-detected (SD) and interval cancers (ICs) were defined based on the time from screening to diagnosis (SD: < 60 days, IC: 60–729 days). The dataset was split after applying exclusion criteria as shown in Fig. 8. To avoid training or testing on mammographically-occult cancers, all cases with lesions that were not visible in the available images, based on the available pixel-level annotations, were eliminated. For each case, only the examination acquired closest to the diagnosis date was retained for both training and testing. In contrast, for women classified as controls, all examinations were included in the training set, whereas the first examination, in chronological order, was included in the validation and test set.

The CSAW dataset includes information about the recall status according to the screening protocol in Sweden, which is double reading with arbitration. In all experiments, the recall label was defined based on the consensus of the two independent radiologists. Because the recall status was not defined in some cases, a positive recall value was set for SD cancers. For cases classified as ICs, as well as for controls with missing recall status, recall was set as negative.

4.2. DDSM dataset

The use of multiple datasets requires harmonization to a common reference standard. The DDSM dataset (Heath et al., 1998) is enriched and acquired in the United States, therefore reflecting a much different screening organization than Europe. It should be noted that we refer here to the original DDSM dataset,[1] rather than the more recent CBIS-DDSM version (Lee et al., 2017), since the latter does not include negative exams (controls) or full mammographic exams with four views.

The DDSM is labeled as as follows: negative exams with a follow-up of at least four years (normal), benign lesions where no further films or biopsies were obtained (benign without a callback), benign lesions found in recalled examinations (benign with callback), and screening examinations where at least one cancerous pathology was found (malignant). For the cancer label, mapping is relatively unambiguous as malignant cancers are assumed to be cases, and all remaining exams are taken as controls. For the recall prediction task, the ground truth is aligned with the CSAW reference standard in the following way: malignant and benign lesions with callbacks were annotated as recalls, and all remaining images as not recalled. Lesion laterality was extracted from the segmentation masks. The DDSM dataset contains 1952 subjects, 1273 controls (including normal images and women with benign lesions), and 679 cases (histologically proven malignant cases); 1341/1952 subjects were labeled as recalled. In this case, the number of exams is equal to the number of subjects. All exams were included in the training set.

[1] Available at https://www.kaggle.com/datasets/skooch/ddsm-mammography

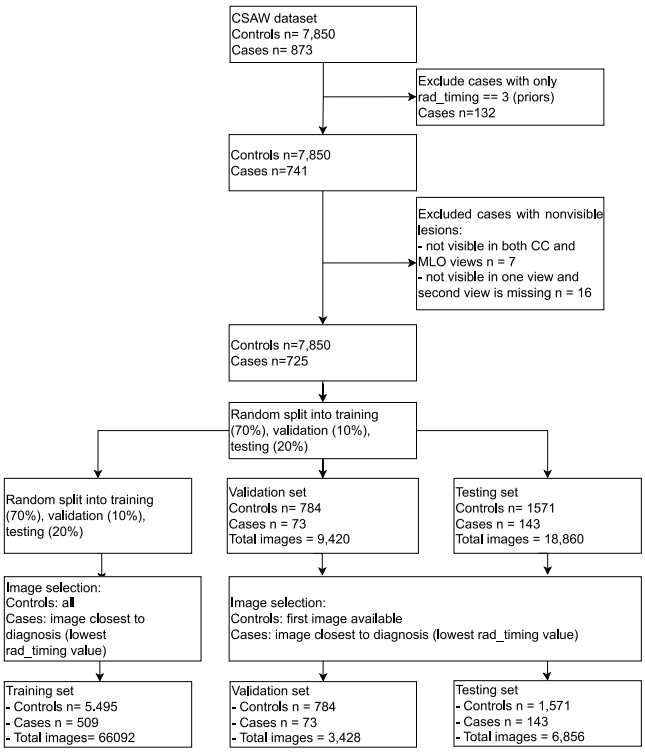


Fig. 8. Flowchart for exam selection and stratification in training, validation and test set, with number of exams and images included at each step.

4.3. Synthetic Karolinska dataset

The number of cases is enriched with a synthetic dataset to improve the generalization properties of the trained models. We opted for inserting lesions through Poisson blending (Pezeshk et al., 2016), instead of more advanced generative models (Garcea et al., 2023; Shen et al., 2021a; Wu et al., 2018) to allow greater control over the generation process and ensure consistency between appearance of the synthetic lesions in the CC and MLO views. The procedure for generating synthetic images is divided into three different steps. First, the lesions with available segmentation in the CSAW training set were cropped from the corresponding images.

Then, data augmentation transformations were applied to the cropped lesions to improve the variability of the examples, including random resizing (same resize factor for both axes in [0.8, 1.2]), color jittering (random contrast and random brightness jittering, range [0.8, 1.2]), and random rotation (in the range [−30°, 30°]). Transformations between the CC and MLO views of the same lesion were performed with identical configuration to maintain consistency between the two views.

Finally, the augmented lesion crops were inserted at compatible points in the CC and MLO views. To select realistic lesion locations, the adjacency matrix (H_g) used to construct the BGN graph in the AGN4V architecture was exploited. The H_g matrix, calculated as detailed in Zhang and Yeung (2012) based on the respective positions of existing lesions, was used to encode the geometric constraints for ipsi – lateral consistency between the CC and MLO pseudo – landmarks (defined in Section 3.2).

To generate a new synthetic case, a normal exam and a pair of pseudo-landmarks for the CC and MLO views were sampled with probability distribution encoded by the H_g matrix. Random noise was added to the pseudo-landmarks to generate the insertion coordinates. A further check was introduced to ensure that the resulting lesion masks did not exceed or was excessively close to the breast boundary. Once valid coordinates were obtained, the lesions were added to the normal breast image using Poisson blending (Pezeshk et al., 2016).

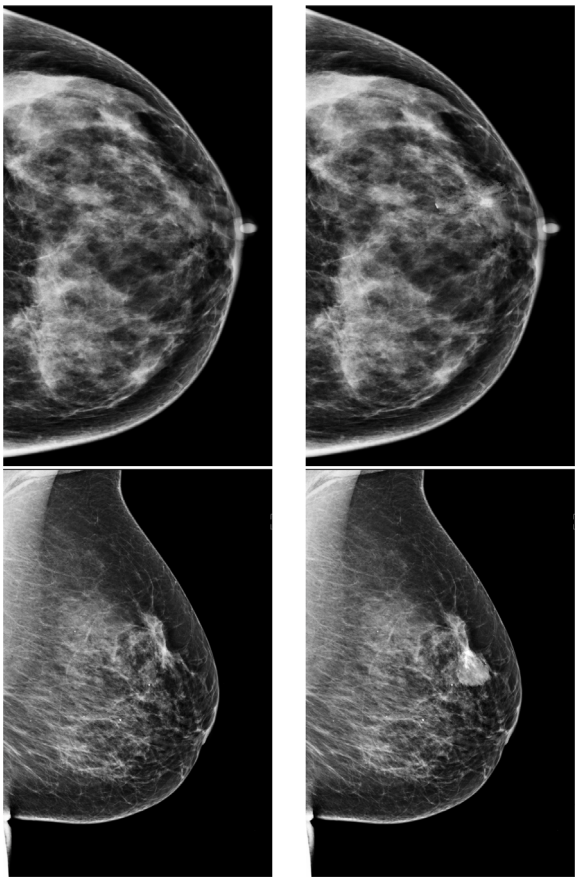


Fig. 9. Two examples of synthetic cases comparing the original healthy control image (left) and the result of the synthetic lesion insertion (right). The network was trained on both the original healthy control and the synthetic lesions.

Synthetic cases were generated starting from the 241 annotated cases in the CSAW dataset in which the lesion was visible and segmented in both the CC and MLO views; a total of 1112 synthetic cases were rendered, about five times the number of annotated cases. Examples of synthetic lesions are demonstrated in Fig. 9.

4.4. Patch extraction for SwaV pretraining

Based on the collected training set, a series of patches was extracted for self-supervised pre-training (Varamesh et al., 2020). Drawing from the experience of Miller et al. (2022), patches of size 256 × 256 were extracted from the original full-resolution 16-bit images. Compared to Miller et al. (2022), our training dataset is larger and contains a much larger number of negative examples. For this reason, more patches were extracted from positive cases than from normal images by reducing the overlap between adjacent patches from 50% to 30%. For the screen-film mammography, positive patches were extracted from the CBIS-DDSM rather than DDSM dataset, in order to exploit the higher quality annotations (Lee et al., 2017)

In total, roughly 950,000 patches were extracted from DDSM and 160,000 from CBIS-DDSM, of which 30,000 were positive patches overlapping with lesions.[2] From the Karolinska dataset, roughly 400,000 positive patches and 13,000,000 negative patches were extracted.

[2] Although the labels were not used during self-supervised pre-training, the dataset was enriched to ensure that SWaV was able to learn to model potential lesions.

5. Experimental settings

5.1. Preprocessing and data augmentation

Preprocessing was performed to find a balance between computational requirements and accuracy, also taking into account the different needs and constraints imposed by each model. The preprocessing is described in detail for the baseline; for the AGN4V and MaMVT, the same preprocessing steps are performed unless otherwise noted.

Baseline. Following the approach in Wu et al. (2020), each image in the CSAW dataset used was cropped from the initial resolution (4096 × 3328 or 3328 × 2560 pixels) and then both CSAW and DDSM, already cropped, were resized to a final resolution of 2677 × 1942 pixels for CC views and 2974 × 1748 pixels for MLO views. To reduce computational requirements, all experiments were performed after further downsampling of the images by a factor of 2, hence a final resolution of 1335 × 971 and 1487 × 874, respectively.

The images were padded until the aspect ratio of the model was reached before resizing. Each image was individually padded with black pixels along the X or Y axis, depending on the original width to height ratio. Padding on the X-axis was added to the left side of the image. For Y- axis, CC views were padded at the top and bottom of the image to center the breast, while MLO views were padded only at the bottom. The left views were flipped along the vertical axis to align all breasts to the right side. Finally, each image was normalized by subtracting its mean and dividing its standard deviation. The same normalization technique was applied as in the original work by Wu et al. (2020), which has the advantage of centering all images on approximately the same input range regardless of the vendor.

AGN4V. The same size for the CC and MLO view was needed in order to simplify the tessellation operation and the extraction of the pseudo-landmarks. Hence, all images were resized to 2974 × 1942 regardless of the view.

MaMVT. Transformers, instead, require square images that can be more easily divided into patches. Hence, a resolution of 1536 × 1536 was used without additional prior padding. Instead, the breasts were stretched to cover the entire available space. However, it should be noted that, unlike padding, this transformation does not preserve pixel spacing.

Data augmentation. For all architectures, the following transformations were applied as data augmentation:

- *Random rescaling:* The resizing factor along each axis is chosen randomly in the following interval [0.8, min(S, 1.2)], where S denotes the maximum factor at which it is possible to stretch the four images without cropping the breasts.
- *Random contrast:*. The contrast factor is randomly selected in the interval [0.8; 1.2]. The following transformation is applied to each pixel of a given image: $x = (1-c) * x_m + c * x$ where c is the contrast factor, and x_m is the mean pixel value of the image.
- *Gaussian noise:* To improve the generalization properties of the model, noise is added to the center of each image, as in Wu et al. (2020).

Rescaling and contrast are applied with the same parameters to all four views.

Finally, an additional form of exam-level data augmentation was introduced by randomly swapping the left and right breast. As introduced in Section 3.3, all the investigated models are not invariant to the order with which the two breasts are presented, usually due to the presence of concatenation layers, and may learn to spuriously associate specific imaging features with the breast laterality. By promoting invariance, this simple form of data augmentation improves generalization across all architectures. Given that this transformation was not included in previous works, experiments were executed with and without random swapping.

5.2. SWaV hyperparameter settings

Contrastive self-supervised learning like SWaV requires aggressive data augmentation to learn robust features. Following previous work by Miller et al. (2022), multi-crop at different scales and sizes was applied. For each 256 × 256 patch, a first crop is applied in the range [1, 0.5] before resizing the patches to 128 × 128; a second image is obtained by cropping in the range [0.8, 0.14] and resizing the crop to 96 × 96. Other transformations applied were contrast jittering (range [0.8, 1.2]) and gamma correction (randomly chosen from [0, 0.25, 0.5, 0.75, 1], where 0 represents the original image).

Using SWaV, the backbone was trained from scratch for 200 epochs with SGD optimizer, batch size 2048, learning rate 2.4 and weight decay of 1e−5. SWaV temperature parameter was set to 0.1, and Sinkhorn regularization to 0.05. We used 50 prototypes and a queue length of 300.

5.3. PEAC hyperparameter setting

The original training hyper-parameters from Zhou et al. (2023) were used, with a cosine learning rate scheduler with a maximum learning rate of 0.1 and no warmup epochs; batch size was reduced to 16 for computational requirements. For the self-supervised pre-training input image and patch size were increased to 1536 × 1536 and 16 × 16 respectively, consistently with the MaMVT architecture; the PEAC implementation available for the Swin-v1 architecture, made available by the original authors, was used for the experiments. The architecture was pre-trained for 100 epochs on the training dataset described in Section 4. The final weights were then transferred to the Swin-v1 backbone of MaMVTv1, and the four-view architecture was trained for another 60 epochs.

5.4. Sampling strategy

Underrepresentation of the positive class was mitigated by sampling an equal number of positive and negative cases at each epoch (Wu et al., 2020). Since DDSM has a higher proportion of positive cases, the exams were separately sampled from DDSM and CSAW, thus artificially balancing the dataset and mitigating possible spurious correlations between the type of mammography (screen-film or digital) and cancer status. AGN4V is the only exception since only the Karolinska dataset was used, due to difficulties in reliably extracting the pseudo-landmarks from screen-film mammograms.

5.5. Hardware

All experiments were carried out on workstations equipped with an Intel® Core™ i9-10980XE CPU @ 3.00 GHz × 36, 192 GB of RAM and two NVIDIA RTX 3090 GPUs with 24 GB VRAM. Two GPUs were used to train the MaMVT and for SWaV pre-training, whereas the baseline and AGN4V were trained on a single GPU. Average inference times were estimated on a subset of 15 exams on the same hardware architecture used for training using a single GPU for the AGN and baseline and two GPUs for MaMVT with a batch size of 1.

5.6. Hyperparameter settings

Baseline The baseline was trained until convergence with a batch size of 32, image size of 1338 × 971 for CC views and 1487 × 874 for MLO, SGD optimizer with 0.9 momentum, weight decay of 1e−3, learning rate of 1e−3, without scheduler, and label smoothing with 0.2 smoothing value. Dropout with rate of 0.5 was applied before the output layer.

AGN4V For the AGN4V most of the hyperparameters are the same as the baseline, except for image size, which is 1487 × 971 for all views. The examined views are L-MLO and R-MLO and the backbone weights are initialized from the Baseline. Dropout is used after every graph convolutional layer except the last, with a dropout rate of 0.5.

MaMVT All architectures were trained for 60 epochs with a batch size of 8, SGD optimizer with 0.9 momentum, 1e−4 learning rate, gradient clipping by norm with value 5.0, weight decay of 1e−3, label smoothing with 0.2 smoothing value, and ImageNet 22k pre-training.[3] For both versions of the backbone Swin architectures, we set an input patch size of 16 with a window size of 12 patches, image input size of 1536 with 1 channel, input embedding size of 128, 4 transformer blocks containing 2, 2, 18, 2 transformer layers with 4, 8, 16, 32 attention heads per attention layer, respectively, no dropout, and drop path with 0.1 rate applied both before and after the attention layer inside each transformer layer for Swin-v2. For Swin-v1, a window size of 12 was used, while for Swin-v2 the window size was increased to 24. For the Focal Loss we set γ to 2 and α to 0.25.

5.7. Performance metrics

All architectures were evaluated mainly in terms of AUC (area under the roc curve) to classify views as cancer/noncancer and recalled/nonrecalled for both breast (or view) and patient (or exam). Predictions were aggregated at the exam-level by taking the maximum between the predictions of the two views. Given the interest in using AI systems as possible rule-out systems that can identify negative cases (Dembrower et al., 2020), we further reported the patient-level false positive rate at a sensitivity of 99% (FPR99). This metric provides a rough estimate of the ability of the system to identify negative exams without missing any cancer. All metrics were calculated for all cancers (including SD and interval cancers), as well as for SD cancers only.

To account for random initialization, all experiments were repeated three times: for each run, the checkpoint with highest cancer AUC on the validation set was selected. Bootstrap with 1000 repetitions, sampled randomly from the three runs, was used to obtain 95% confidence intervals. For XAI analysis, the best performing repetition was sampled for each architecture.

Bootstrap with 1000 repetitions was also used to test statistical significance by comparing each architecture and training setting with the baseline. To reduce the number of tests, we only tested for statistically significant differences in cancer AUC. We further controlled for multiple hypothesis testing using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995); an adjusted p-value < 0.05 indicated statistical significance.

5.8. Explainability metrics

In order to further evaluate the results obtained, a visual explanation technique, Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) was used to produce heatmaps that attribute the classification to specific areas of the image. For a more quantitative analysis of the Grad-CAM attention maps, three evaluation metrics were introduced, denoted in the following as DICE, Intersection Over Breast (IOB) and Intersection Over Lesion (IOL). The metrics are defined as follows:

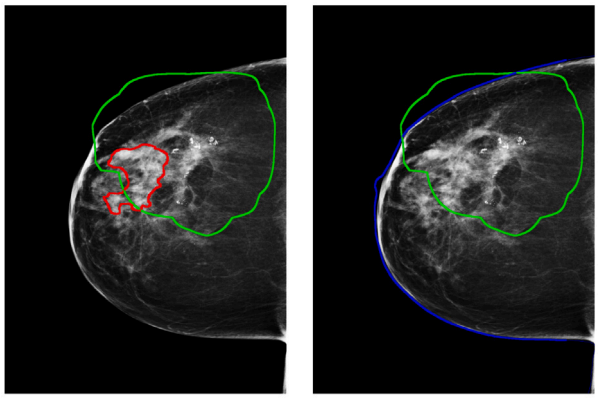
$$DICE_t = \frac{2 * |AP \cap GP_t|}{|AP \cup GP_t|} \tag{6}$$

$$IOB_t = \frac{|GP_t \cap BP|}{|BP|} \tag{7}$$

$$IOL_t = \frac{|GP_t \cap AP|}{|AP|} \tag{8}$$

where AP represents the pixels in segmentation mask, GP_t the pixels highlighted in the Grad-CAM heatmap binarized at the threshold t,

[3] We selected publicly available weights from Hugginface's Pytorch Image Models https://github.com/huggingface/pytorch-image-models



(a) (b)

Fig. 10. The red line (a) represents the lesion annotation pixel (AP), while the green one (a) underlines the area covered by the Grad-CAM heatmap (GP_t). These two quantities were used to calculate the DICE score and the Intersection over Lesion. The blue (b) represents the area covered by the entire breast (BP), and is used to calculate the Intersection over Breast.

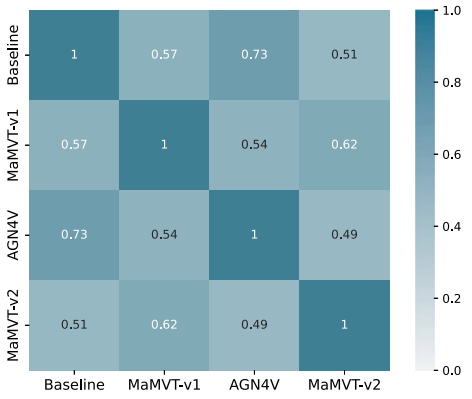


Fig. 11. Correlation between the predictions of three best run of each architecture on the validation set.

and BP the pixels in the breast area. All metrics were calculated for $t \in \{0.2, 0.4, 0.6, 0.8\}$. Intuitively, the DICE score indicates to what extent the network is capable of localizing the lesion, on the one hand, and attribute the cancer prediction to its presence, on the other. IOB instead quantifies whether the explanation is spread over the entire breast and implicitly whether the network relies more on global rather than local features. Finally, IOL calculates to what extent the lesion is part of the explanation: a high IOL combined with a low DICE implies that the explanation focuses on a large portion of the breast that includes, but is not limited to, the lesion. Fig. 10 shows an example of pixel-to-pixel annotation for annotation pixels (AP), Gradcam Pixels (GP) and Breast Pixel (BP), respectively.

The explainability metrics were calculated separately for the training set and the combined validation/test set, and for correctly and incorrectly classified exams. Only cases for which the lesion was segmented are included in this analysis, we include a total of 498/59 views for baseline and MaMVT while 251/29 for AGN4V, respectively, for the training and validation set.

6. Results

In this section, we analyze the results from two distinct perspectives. The first subsection presents the results obtained using the reference performance metrics, broken down by breast and patient, while the second presents a punctual analysis on single views to evaluate the accuracy of the provided predictions.

6.1. Predictive performance

The performance of all architectures introduced in Section 3 for both cancer detection and recall prediction tasks are summarized in Table 1 and Table 2 for the validation and test set, respectively. The Baseline and AGN4V were trained from scratch, as well as using pre-trained weights from SWaV; the MaMVT architectures were trained using pre-trained Imagenet weights, while for the MaMVT-v1 architecture we also compared against PEAC self-supervised pretrained; for MaMVT-v2 experiments in the first section of both tables also use the side-invariant four-view version of the cross-attention module. Performance was calculated at both the breast and patient level; in the text, all metrics are reported on the test set and the patient level, unless otherwise noted.

By exploiting *self-supervised pre-training*, performance improved by approximately 6% for both baseline (Cancer AUC=69.9 vs. 76.3, $p = 0.126$ and Recall AUC=70.1 vs. 76.5) and AGN4V (Cancer AUC=63.7 vs. 69.7, $p = 0.250$ and Recall AUC= 66.2 vs. 73.7). Differences between the AGN4V and the baseline, under this training regime, were not statistically significant (Cancer AUC=63.7 vs. 69.9, $p = 0.979$). We observed a small 1 to 2% boost for MaMVT-v1 (Cancer AUC=79.2 vs. 80.3, $p = 0.04$ and Recall AUC= 78.9 vs. 82.9). Given the small improvement compared with the additional pre-training effort required, for MaMVT-v1 the ImageNet pre-trained version was used in the rest of this paper.

All architectures and metrics benefit by introducing random swapping of the left and right breast during training, but the highest improvement can be seen on the AGN4V (Cancer AUC=69.7 vs. 74.9, $p = 0.125$ and Recall AUC=73.7 vs. 75.4). The benefit is more pronounced at the breast level and on the validation set: indeed, the performance increase was mostly measured on the left side.

Among the three architectures, *the single best performing architecture is MaMVT-v2*, trained without side invariance and with random flipping augmentation (Cancer AUC=80.1, $p = 0.053$, Recall AUC=81.3), followed by self-supervised PEAC pre-trained MaMVT-v1 (Cancer AUC=80.3, $p = 0.053$, Recall AUC=82.9), ImageNet pre-trained MaMVT-v1 (Cancer AUC=79.2, $p = 0.062$, Recall AUC=78.9), Baseline (Cancer AUC=74.9 and Recall AUC=77.1) and AGN4V (Cancer AUC=74.9, $p = 0.105$, Recall AUC=75.4), with all p values computed against the baseline trained with random flipping augmentation and self-supervised pre-training. Baseline and AGN4V show higher signs of overfitting the validation set compared to MaMVT; indeed, performance on the validation set is comparable for the Baseline, AGN4V, MaMVT-v1 and MaMVT-v1 (PEAC) architectures (Cancer AUC (breast)=81.8, 81.0, 81.7 and 80.8, respectively), with MaMVT-v2 outranking other methods (Cancer AUC (breast)=85.0). Despite similar performance, Fig. 11 shows that the correlation between the predictions in the validation set is low, in particular between both versions of MaMVT, showing the influence of the different backbone, and thus the four architectures appear to focus on different patterns.

The complementary nature of the three types of architectures is further demonstrated by the performance of a simple ensemble obtained by averaging the predictions of the four architectures (ImageNet pre-trained MaMVT-v1 was used in the ensemble), randomly sampled among all the available runs. The ensemble outperforms the four individual architectures (Cancer AUC = 82.6, $p = 0$, Recall AUC = 84.5 for all cancers, Cancer AUC = 87.5 and Recall AUC = 84.8 for SD cancers). This improvement is also visible in the bootstrapped ROC curves in Fig. 12.

Further insights are obtained from the *score distributions of the positive and negative exams* in Fig. 13. As expected, all architectures tend to produce overconfident scores. Most errors are due to false negatives, that is, cancer cases classified as negatives with very low probability.

Table 1
Performance metrics, at breast and patient level, calculated on the validation set. Performance metrics reported include the Area under the ROC Curve (AUC) for cancer detection (Cancer, for brevity) and recall prediction (Recall, for brevity). Performances are separately calculated on all cancers (including screen detected and interval cancers), and screen detected cancer only; 95% confidence intervals are calculated based on 1000 bootstrap repetitions from three training runs. Models indicated with * (top rows) are trained from scratch. All other models are either pre-trained on ImageNet (MaMVT-v1, MaMVT-v2) or using self-supervised learning (Baseline, AGN4V, MaMVT-v1 (PEAC)). All models indicated with † (bottom rows in the table) are trained using random swapping of the left and right breast. The remaining models (intermediate rows) are trained using standard data augmentation. MaMVT models indicated with • are trained using the side invariant version of the cross-attention module. **Best** and *second–best* models are indicated in bold and underline characters, respectively.

Model	All cancers					Screen detected cancers				
	Cancer (Breast) ↑	Cancer (Patient)↑	Recall (Breast)↑	Recall (Patient)↑	FPR99 (Patient)↓	Cancer (Breast)↑	Cancer (Patient)↑	Recall (Breast)↑	Recall (Patient)↑	FPR99 (Patient) ↓
Baseline*	77.2 (65.4-86.0)	73.4 (64.2-82.4)	75.6 (64.0-84.7)	73.5 (65.0-82.0)	87.2 (72.3-97.5)	81.3 (70.2-90.0)	76.2 (66.7-86.3)	75.6 (64.1-84.6)	73.6 (65.7-81.9)	79.3 (67.6-89.4)
AGN4V*	70.4 (58.0-80.3)	67.5 (60.9-73.5)	70.1 (56.3-81.3)	67.3 (59.8-73.7)	92.2 (78.0-97.6)	73.7 (60.6-84.0)	68.7 (60.9-75.4)	71.0 (57.2-81.2)	67.2 (60.0-73.2)	89.7 (69.8-97.6)
Baseline	79.1 (67.5-87.5)	79.7 (73.5-85.1)	79.4 (68.8-87.7)	80.2 (72.5-86.4)	85.3 (72.8-96.2)	87.4 (78.7-93.9)	85.5 (77.7-91.1)	79.5 (68.9-88.1)	80.7 (72.3-87.2)	74.3 (45.7-83.8)
AGN4V	76.3 (66.6-84.5)	73.9 (67.4-79.7)	75.4 (64.9-84.0)	73.1 (65.7-79.7)	92.7 (80.6-99.2)	80.7 (69.3-88.9)	78.4 (71.2-84.1)	75.7 (64.6-84.2)	73.2 (65.9-79.6)	83.5 (60.7-92.0)
MaMVT-v1	79.8 (69.0-88.4)	78.6 (70.3-85.2)	81.0 (71.1-89.2)	79.3 (70.9-86.2)	87.2 (64.1-97.4)	85.1 (74.0-93.7)	82.1 (71.7-89.9)	81.2 (71.6-89.3)	79.4 (70.3-86.5)	82.5 (56.6-97.6)
MaMVT-v2•	86.1 (77.9-91.7)	83.1 (77.5-87.3)	85.5 (75.0-93.1)	83.1 (76.9-88.6)	*76.3 (53.7-95.5)*	90.1 (81.9-95.9)	87.4 (81.9-91.9)	85.9 (75.4-93.4)	83.3 (76.7-88.6)	62.8 (44.5-79.9)
Ensemble ‡	*86.3 (78.3-91.9)*	*84.2 (77.9-88.9)*	*86.6 (77.8-93.0)*	*85.6 (79.6-90.5)*	**76.2 (55.8-88.4)**	*92.5 (86.5-96.8)*	*90.5 (85.6-94.4)*	*86.8 (77.5-93.1)*	*85.7 (79.2-90.5)*	*56.6 (33.5-73.1)*
Baseline†	81.8 (69.8-90.0)	83.5 (77.8-88.7)	81.1 (70.5-88.7)	78.8 (71.5-84.8)	86.6 (61.6-99.6)	90.3 (81.8-95.7)	87.8 (81.5-92.5)	81.5 (70.7-89.4)	79.5 (72.4-85.4)	73.3 (49.5-90.1)
AGN4V†	81.0 (72.3-88.0)	78.3 (71.5-84.1)	85.0 (78.5-90.7)	81.1 (75.1-86.5)	94.2 (82.5-99.6)	89.3 (82.3-94.6)	86.2 (79.7-91.0)	85.4 (78.3-90.9)	81.4 (74.7-86.5)	79.8 (51.8-90.3)
MaMVT-v1†	81.7 (72.5-88.8)	79.8 (73.8-85.1)	81.6 (71.7-89.5)	80.2 (72.3-85.9)	80.5 (64.4-88.3)	87.7 (78.5-94.4)	84.6 (77.1-89.9)	82.0 (72.3-90.1)	80.2 (73.2-86.2)	74.9 (47.5-88.4)
MaMVT-v1 (PEAC)†	80.8 (70.8-88.4)	81.3 (74.7-86.2)	83.5 (75.3-90.0)	82.6 (75.9-88.2)	81.4 (68.0-92.9)	86.9 (77.9-93.7)	84.6 (77.7-90.1)	83.8 (74.8-90.5)	82.9 (76.5-88.2)	72.3 (56.4-83.9)
MaMVT-v2†	85.0 (73.6-92.8)	83.7 (76.4-89.7)	84.8 (74.5-92.3)	83.9 (76.1-90.0)	83.7 (59.6-91.9)	88.8 (77.8-95.9)	87.3 (77.3-93.3)	85.2 (75.7-92.7)	84.1 (75.9-90.1)	76.7 (48.3-92.0)
Ensemble†	**88.4 (79.7-94.2)**	**87.2 (81.6-91.7)** ‡	**89.3 (82.7-94.4)**	**87.3 (81.4-91.6)**	79.7 (45.5-93.3)	**94.8 (89.9-97.9)**	**92.5 (87.8-95.8)**	**89.7 (82.8-94.8)**	**87.5 (82.2-92.1)**	**50.3 (27.9-72.4)**

Table 2
Performance metrics, at breast and patient level, calculated on the test set. Performance metrics reported include the Area under the ROC Curve (AUC) for cancer detection (Cancer, for brevity) and recall prediction (Recall, for brevity). Performances are separately calculated on all cancers (including screen detected and interval cancers), and screen detected cancer only; 95% confidence intervals are calculated based on 1000 bootstrap repetitions from three training runs. Models indicated with * (top rows) are trained from scratch. All other models are either pre-trained on ImageNet (MaMVT-v1, MaMVT-v2) or using self-supervised learning (Baseline, AGN4V, MaMVT-v1 (PEAC)). All models indicated with † (bottom rows in the table) are trained using random swapping of the left and right breast. The remaining models (intermediate rows) are trained using standard data augmentation. MaMVT models indicated with • are trained using the side invariant version of the cross-attention module. **Best** and *second–best* models are indicated in bold and underline characters, respectively.

Model	All cancers					Screen detected cancers				
	Cancer (Breast)	Cancer (Patient)	Recall (Breast)	Recall (Patient)	FPR99 (Patient)	Cancer (Breast)	Cancer (Patient)	Recall (Breast)	Recall (Patient)	FPR99 (Patient)
Baseline*	72.8 (62.9-81.3)	69.9 (60.8-76.6)	74.0 (64.3-82.1)	70.1 (63.0-75.7)	93.3 (86.4-99.8)	77.7 (66.9-86.7)	74.2 (64.2-81.7)	74.1 (63.7-82.9)	70.2 (62.7-76.2)	90.5 (80.2-95.3)
AGN4V*	64.4 (55.9-73.1)	63.7 (56.4-71.3)	67.6 (59.0-75.0)	66.2 (59.5-73.1)	95.7 (90.6-99.0)	68.1 (58.4-76.5)	65.9 (57.5-74.8)	68.1 (59.6-75.2)	66.2 (59.5-73.1)	93.4 (86.2-97.2)
Baseline	78.4 (70.8-85.2)	76.3 (70.9-81.0)	78.2 (69.6-85.2)	76.5 (70.9-81.3)	93.5 (81.0-99.6)	82.3 (73.5-89.4)	80.1 (74.7-85.0)	78.5 (70.1-85.5)	77.2 (72.1-81.6)	91.5 (73.3-99.6)
AGN4V	68.9 (60.2-76.9)	69.7 (63.2-75.4)	76.0 (68.0-82.5)	73.7 (68.1-78.8)	96.5 (92.5-98.9)	76.0 (66.5-84.0)	74.3 (68.0-79.8)	76.4 (68.0-83.2)	73.6 (68.4-79.2)	96.0 (89.9-98.8)
MaMVT-v1	80.8 (74.2-86.6)	79.2 (74.0-83.8)	80.2 (72.1-86.9)	77.9 (71.3-83.0)	93.8 (74.1-99.7)	84.9 (77.6-90.7)	82.8 (77.4-87.8)	80.3 (72.3-86.9)	78.2 (72.1-83.1)	91.8 (67.7-99.6)
MaMVT-v2•	*80.4 (72.9-86.9)*	*78.6 (72.3-83.3)*	82.1 (74.7-88.1)	79.1 (73.7-83.9)	90.2 (81.1-98.0)	84.5 (76.0-90.7)	82.8 (76.8-87.9)	82.4 (74.6-88.6)	79.6 (74.5-84.2)	88.4 (77.6-96.3)
Ensemble	*83.6 (76.9-89.3)*	*82.3 (77.6-86.2)* ‡	*85.4 (79.1-90.9)*	*82.8 (78.2-86.9)*	*89.8 (76.4-98.3)*	*87.9 (80.7-93.5)*	*86.5 (81.9-90.9)* ‡	*85.8 (79.3-91.0)*	*83.4 (78.5-87.5)*	*87.0 (63.9-96.2)*
Baseline†	77.8 (70.5-84.3)	74.9 (69.2-79.9)	79.0 (71.2-85.5)	77.1 (71.3-82.5)	92.5 (85.1-99.4)	81.4 (73.1-88.2)	79.8 (73.4-85.4)	79.2 (71.7-85.7)	77.2 (71.2-82.4)	90.4 (77.6-96.3)
AGN4V†	76.7 (70.1-82.9)	74.9 (69.7-79.0)	79.5 (72.5-85.0)	75.4 (70.1-79.7)	95.4 (86.7-98.7)	82.7 (76.2-88.3)	78.4 (73.1-83.0)	80.3 (74.0-85.8)	75.7 (70.6-80.4)	91.1 (76.6-96.5)
MaMVT-v1†	80.2 (72.5-86.2)	79.2 (73.1-84.0)	81.4 (74.0-87.7)	78.9 (73.0-83.7)	91.9 (74.2-99.6)	84.7 (76.9-91.0)	83.1 (77.4-87.9)	81.6 (73.5-88.1)	79.1 (73.1-84.2)	90.5 (72.1-99.7)
MaMVT-v1 (PEAC)†	80.7 (73.5-87.0)	80.3 (75.2-84.4)	82.8 (75.4-88.7)	82.9 (78.0-87.3)	92.4 (76.2-99.0)	86.7 (78.5-92.5)	85.8 (80.4-90.3)	83.1 (75.7-89.3)	83.1 (78.2-87.5)	92.1 (73.9-99.0)
MaMVT-v2†	82.3 (75.9-87.4)	80.1 (75.2-84.1)	83.8 (77.2-89.1)	81.3 (76.2-85.5)	**89.7 (74.9-98.3)**	86.8 (80.6-91.9)	84.6 (79.8-88.7)	84.0 (76.9-89.4)	81.7 (76.6-85.9)	**86.4 (70.3-95.9)**
Ensemble†	**84.4 (78.1-89.6)**	**82.6 (77.7-86.5)** ‡	**86.9 (80.6-91.8)**	**84.5 (79.2-88.6)**	93.3 (73.1-99.2)	**88.9 (82.3-94.0)**	**87.5 (82.6-91.6)** ‡	**87.2 (81.3-92.0)**	**84.8 (80.0-88.9)**	90.3 (64.6-98.3)

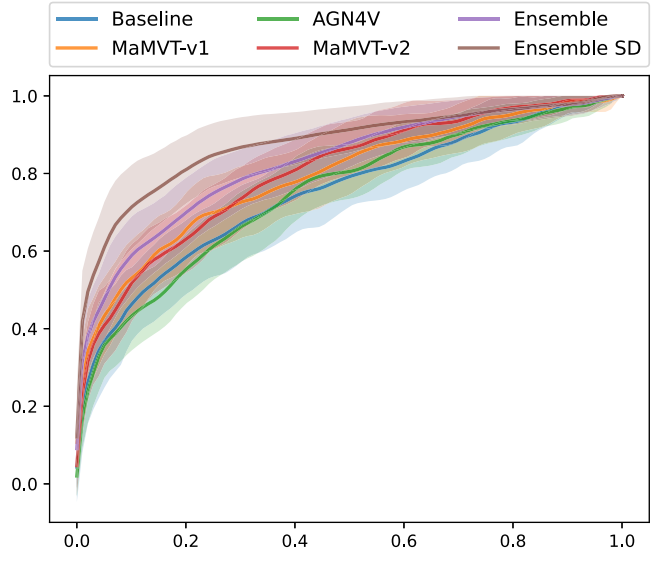


Fig. 12. ROC curve on the test with 95% confidence intervals calculated by bootstrapping. All ROCs are calculated on all cancers, except for the ensemble that are calculated on both all cancers and SD cancers.

Upon visual inspection, lesions that are consistently missed by all architectures can be attributed to relatively smaller and borderline lesions. It is interesting to note that both MaMVT versions tend to detect more cancers, with MaMVT-v2 in particular also assigning higher scores to negative exams, whereas other architectures are more accurate in classifying negative cases. False positives are also dependent on the architecture: for instance, AGN4V is more apt at detecting asymmetries between the two breasts, even in cases that are negative for cancers (see example in Fig. 14).

In Table 3 the best models for each architecture were further evaluated by introducing test-time augmentation (TTA), by averaging the predictions on 100 variants of the same image. TTA improves the performance of all individual models in terms of AUC, but does not improve that of the ensemble, which offers similar performance benefits with a much smaller computational footprint, requiring only three instead of 100 evaluations. In these results, the same data augmentation was applied to each input view; we also tried to apply different data augmentations to each of the four views, as suggested in Kyono et al. (2018), but did not observe any substantial change (results not shown).

The higher performance of the MaMVT architecture comes with a slightly higher computational cost. Table 4 compares the difference architectures in terms of model parameters and throughput, measured as the number of exams (four views) processed per second. MaMVT-v2 has 43% more parameters than the baseline and inference is 34% slower. The slowest architecture at inference time is the AGN4V, due to the need to extract and process the graphs. It should be noted that AGN4V requires several preprocessing steps such as segmenting the images, computing the landmarks and retrieving information from the adjacency matrix to build the feature graphs; such preprocessing steps do not benefit, in the current implementation, from GPU acceleration.

Table 3
Patient-level AUC calculated for the best run of each architecture and for the ensemble with and without test-time augmentation (TTA). TTA increases performance for each individual architecture, but has minimal effect on the ensemble.

Type	Model	AUC Cancer	AUC Recall	FPR99
Pre-Training	Baseline	74.1	75.8	94.0
	AGN4V	74.6	75.1	99.0
	MaMVT-v2	77.3	77.5	99.0
	Ensemble	80.7	81.6	95.0
Pre-Training and test-time augmentation	Baseline	**75.6**	**77.3**	96.2
	AGN4V	**75.1**	**77.0**	97.0
	MaMVT-v2	**78.2**	**77.5**	97.1
	Ensemble	80.9	81.7	97.4

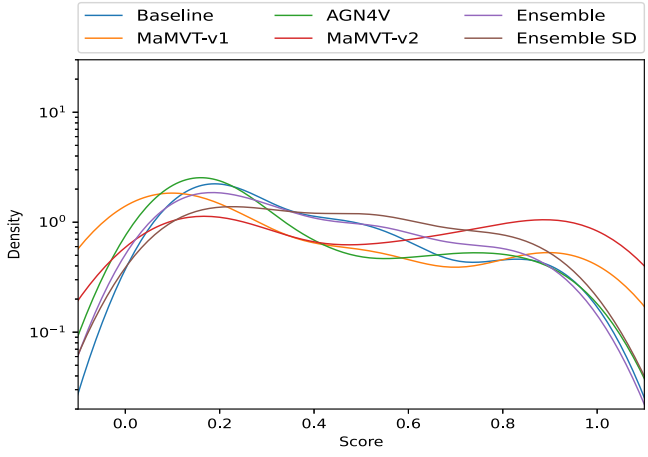


Fig. 13. Score distribution on the cancer cases (a) and negative control (b) exams for each architecture and ensemble (y axis in logarithmic scale). The best performing run is selected for each architecture on the validation set. MaMVT-v2 assigns high score to the most cancer cases, but also generates the highest percentage of highly scored false positives. For the ensemble, distribution of positive cases is reported for all cancers and for screen detected cancers (SD) separately.

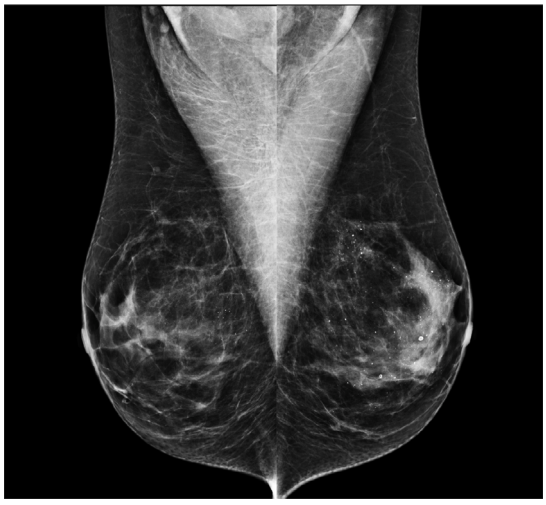


Fig. 14. Example of negative control that was classified as positive by the AGN4V architecture. Note the asymmetry between the left and right breast.

Table 4
Comparison of model parameters (millions) and throughput (exams processed per second) for the four architectures.

Model	Parameters	Inference time (Sample/s)
Baseline	6.13 M	3.45
AGN4V	7.57 M	0.28
MaMVT-v1	8.8 M	2.31
MaMVT-v2	8.7 M	2.27

6.2. Explainability metrics

To quantify the agreement between the attention maps produced by the four models and the radiologists' annotations, we analyzed the annotated images separately in the training set (498 for the baseline and MaMVT and 251 for the AGN4V), and the combined validation and test set, which we rename Validation Test (59 images for baseline and MaMVT and 29 images for the AGN4V). The number of images is lower for the AGN4V since GradCAM was applied only to the MLO, whereas for the Baseline and MaMVT the heatmaps were separately calculated for CC and MLO views. All experiments included in this analysis were trained with random swap augmentation.

To investigate the agreement between radiologist's annotations and Grad-CAM heatmap, Figs. 15(a) and 15(b) show the distribution of the DICE scores for AGN4V, Baseline, and the two MaMVT versions, separately reported for correct and wrong predictions. AGN4V achieves the highest agreement (median $DICE_{0.6} = 0.09$ for correct predictions) because the graph-based component encourages more focused attention on specific region of interests. However, in the case of incorrect predictions, the DICE score drops (median $DICE_{0.6} = 0.002$), thus indicating that the network is unable to locate the lesion. A similar behavior is observed for the baseline for both correct (median $DICE_{0.6} = 0.07$) and incorrect predictions (median $DICE_{0.6} = 0.000$). MaMVT-v1 shows the lowest agreement due to transformers' proclivity to focus on global characteristics (median $DICE_{0.6} = 0.0$ and median $DICE_{0.6} = 0.0$). Unlike it v1 predecessor, MaMVT-v2 demonstrates a heightened attention to local features, substantiated by its median $DICE_{0.6} = 0.04$ for correct predictions and $DICE_{0.6} = 0.01$ for incorrect predictions.

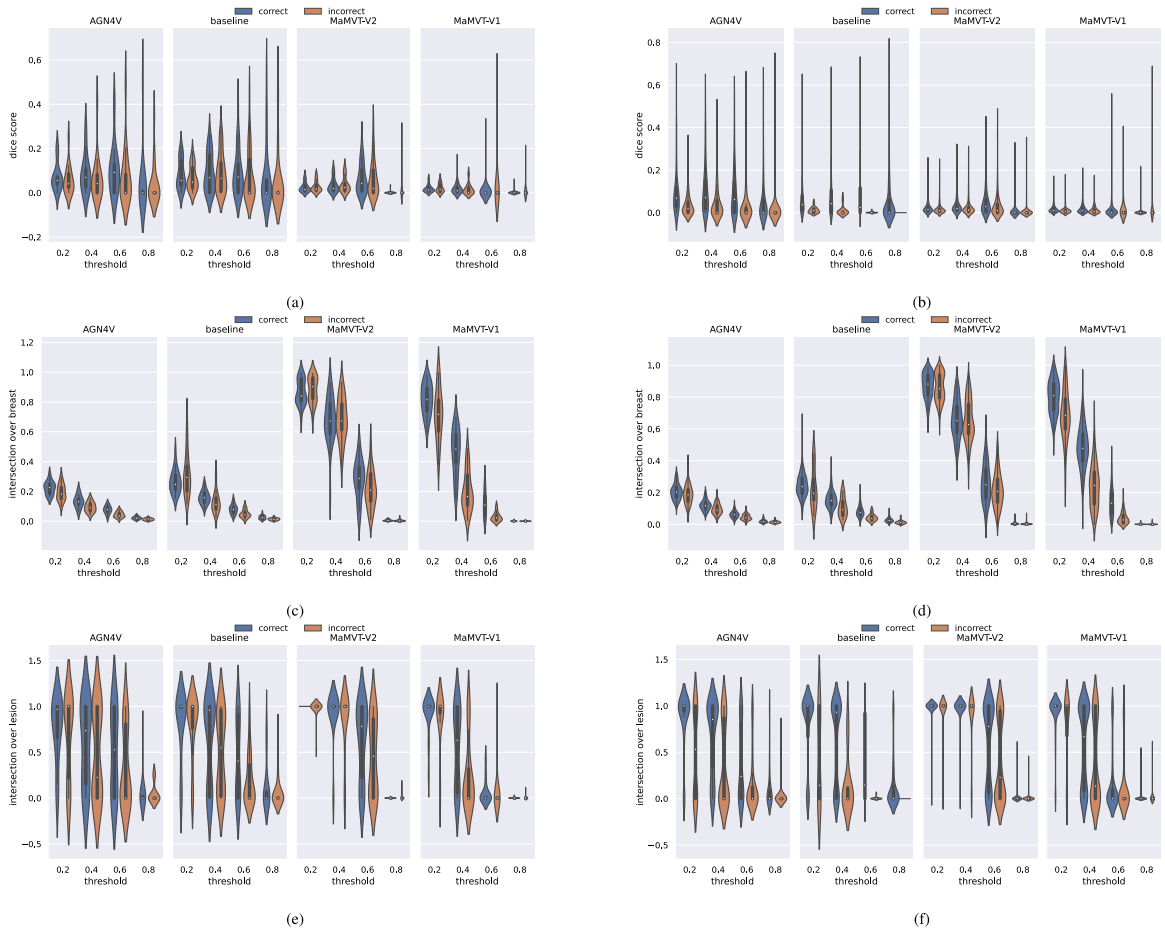


Fig. 15. DICE (a–b), IOB (c–d) and IOL (e–f) scores calculated on the cancer cases of Validation Test set (a,c,e) and Training set (b,d,f), divided by architecture and by correct and incorrect predictions (in other words, for detected and missed cancers). Each metric compares the GradCAM heatmaps with the lesion segmentation as detailed in Section 5.7. The scores at various thresholds were obtained by normalizing the GradCAM heatmaps with values between 0 and 1, and then binarizing the maps by applying the corresponding threshold.

Distributed heatmaps suggests that multi-view architectures are sensitive to global features of the breast. Compared to focused attention maps that concentrate on the lesion area, clinical interpretation of such distributed heatmaps is less clear, as well as their role in achieving good generalization to novel data. The Insertion Over Breast (IOB) is presented for the four architectures in Figs. 15(c) and 15(d). This metric quantifies *how large the area analyzed in the attention map is compared to the breast area*. Even at very low threshold values, the values found for the baseline and AGN4V are less than 30% ($IOB_{0.2} = 0.22$ and $IOB_{0.2} = 0.24$, respectively), demonstrating that the attention map is more focused on localized features. On the other hand, for MaMVT-v1 and MaMVT-v2 IOB values were closer to 1 at low thresholds ($IOB_{0.2} = 0.81$ and $IOB_{0.2} = 0.84$ respectively), denoting that transformers attend to the breast as a whole. It should be noticed, however, that this behavior may be emphasized by the Swin aggregation mechanism.

Lastly, Figs. 15(e) and 15(f) depict the distribution of the IOL_t metric. Even at high threshold values ($IOL_{0.6} = 0.526$), the AGN4V on average covers 50% of the lesion area in the case of successful predictions. Similar results are obtained for the baseline ($IOL_{0.6} = 0.405$).

Fig. 16 compares the activation maps of four images of the test set, together with the prediction scores of the four models (Baseline, AGN4V, MaMVT-v1 (ImageNet and PEAC pre-training) and MaMVT-v2), while Fig. 17 compares the activation maps of the four views of a single sample image. It can be noticed how, in Figs. 17c, 17i and 17o, the attention maps of the baseline and AGN4v are concentrated on specific structures, usually encompassing the lesion area; interestingly, there appears to be a connection between the ability to localize the lesion and the confidence of the prediction, as can be noticed in particularly by comparing Figs. 16c and 16i (both correct predictions with high positive scores) and Figs. 16i and 16j (both incorrect predictions with low positive scores). In contrast, MaMVT architectures attention maps are scattered on larger portion of the breast, regardless of the prediction score, as seen in Figs. 17e, 17f, 17q and 17r; it is also interesting to note that although the heatmaps for the PEAC pre-trained MaMVT-v1 appear even more scattered compared to their ImageNet pre-trained counterparts, the scores for the former appear to be more accurate, showing a similar behavior to those of MaMVT-v2. In the event of incorrect predictions, as well as on normal views in which no lesions are visible, the attention is evenly distributed throughout the breast (Figs. 16l and 16x), indicating that the transformer does not "see" any lesion.

7. Discussion

In this paper, we have compared three types of architectures with largely different inductive biases on multi-label classification (cancer and recall) from weakly-labeled multi-view mammographic images. Our experimental findings highlight that, although transformer-based architectures achieve higher AUC in both cancer and recall prediction, different architectures differ widely in their predictions, as well as the type of features they detect. Therefore, combining them appears to be essential especially when the amount of training data is limited.

For the Baseline architecture, we achieve similar results compared to the original paper both when trained from scratch (cancer AUC=72.4

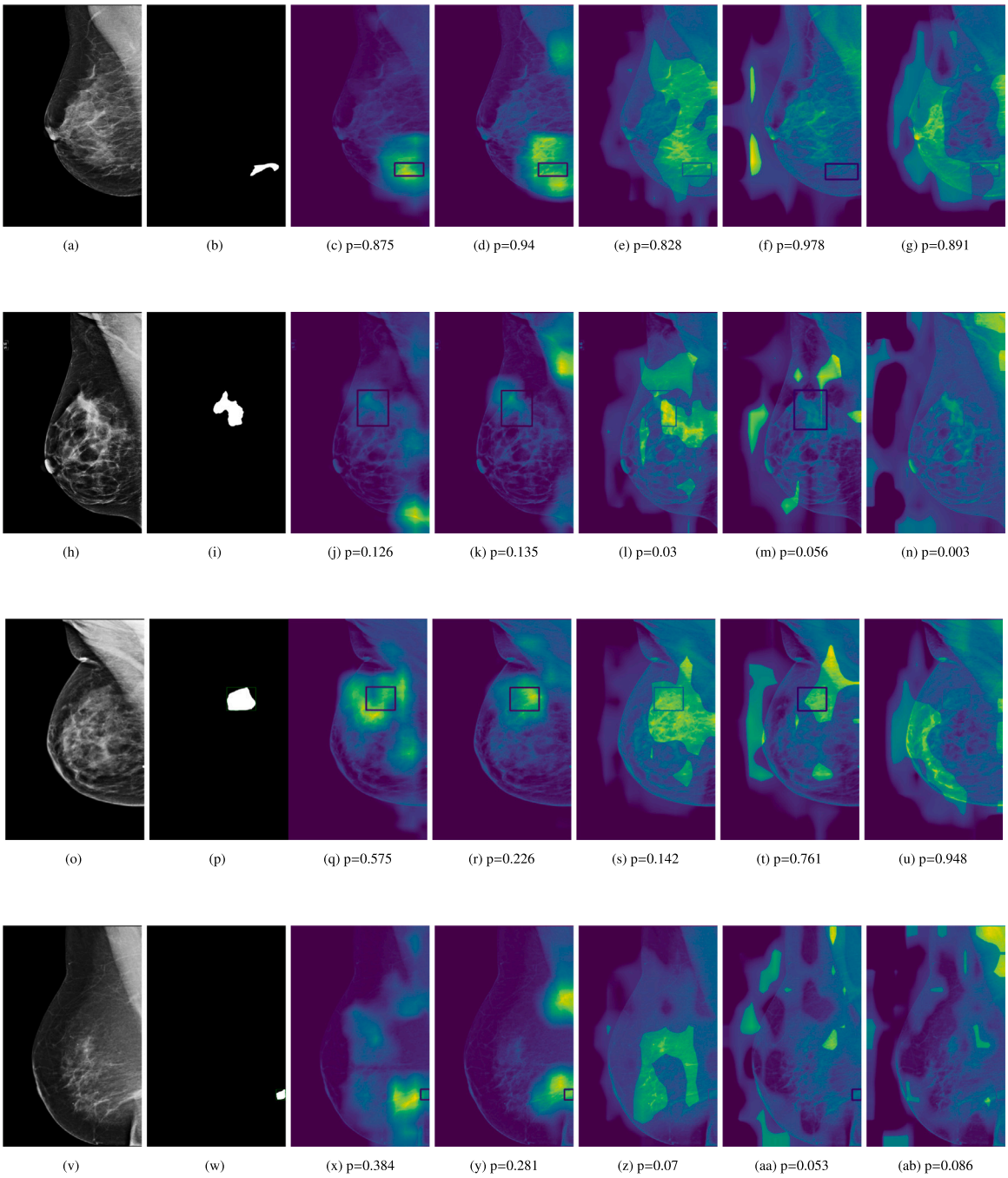


Fig. 16. Grad-CAM heatmap for the cancer prediction task. From left to right, each row displays the original image, the corresponding lesion annotation mask, and the GRADCAMs obtained from the Baseline, AGN4V and MaMVT-(v1 Imagenet, v1 PEAC and v2) architectures along with the corresponding prediction score. While the GRADCAM for the baseline and AGN4V architectures are more focused on local areas, the MaMVT architectures attend to larger portion of the breast parenchyma independently of the prediction score.

vs. 72.2) and with pre-training (cancer AUC=77.3 vs. 78.4). However, when comparing these results, we must keep in mind that the composition of the NYU dataset is quite different from CSAW, and this affects both training and evaluation. For instance, the NYU training set contains a larger number of controls (229,426 exams from 141,473 women, compared to 16,523 exams from 5,495 controls in the CSAW cohort) and biopsied cancers (750 visible lesions, compared to 509 cases). In addition, the NYU cohort includes more than 4,000 biopsied benign lesions, whereas in CSAW biopsy results are not available for non-cancer cases. On the other hand, cancers included in the NYU cohort were biopsied within 120 days of the screening mammogram, while the CSAW cohort also includes more subtle cancers detected up to 730 days after the initial screening examination. We found experimentally that enriching the cohort with additional lesions, taken from DDSM as well as synthetically generated, was key to achieve generalization. Previous replication studies achieved performance close, although lower, to NYU using a highly enriched dataset with roughly 2,000 biopsied confirmed lesions and a matched control cohort of similar size (Condon et al., 2021).

Our results qualitatively confirm previous findings on the role of pre-training for mammographic image classification (Wu et al., 2020; Miller et al., 2022; Matsoukas et al., 2022). Pre-training in the original NYU paper was supervised via BI-RADS score (Wu et al., 2020). However, BI-RADS scores are not registered in typical European screening programs; additionally, recall rates in European screening programs are usually lower than in the United States, especially when arbitration is

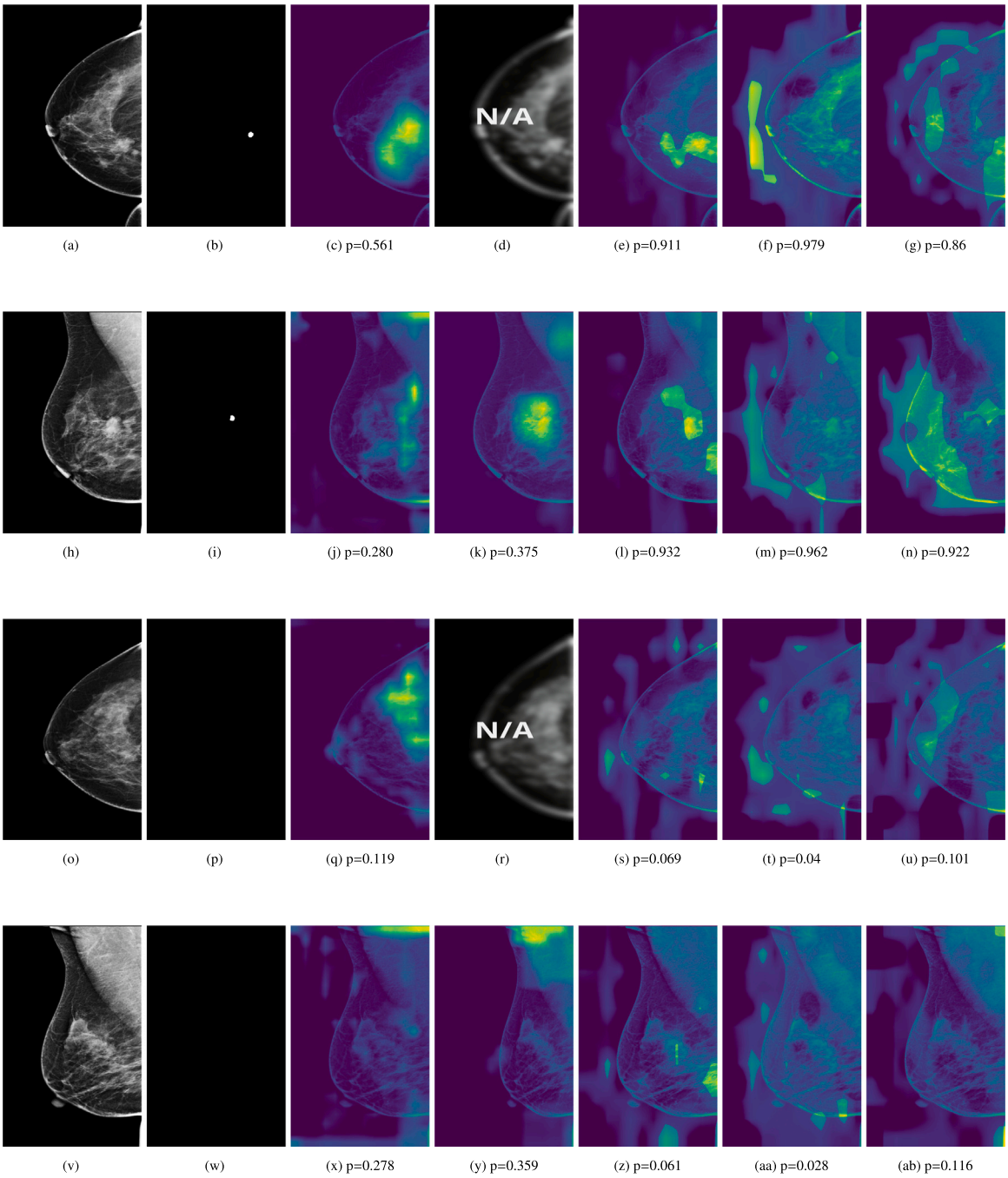


Fig. 17. Grad-CAM heatmap for the cancer prediction task on different views of the same exam. From top to bottom, each column displays the L-CC, L-MLO, R-CC and R-MLO. From left to right, each row displays the original view image, the corresponding lesion annotation mask, and the GRADCAMs obtained from the Baseline, AGN4V (when possible) and MaMVT-(v1 Imagenet, v1 PEAC and v2) architectures along with the corresponding prediction score. While the GRADCAM for the baseline and AGN4V architectures are more focused on local areas, the MaMVT architectures attend to larger portion of the breast parenchyma independently of the prediction score.

used (Domingo et al., 2016). Therefore, self-supervised pre-training is an equally effective, yet easier to implement, strategy that has also been shown to promote invariance with respect to specific vendor-related imaging characteristics (Miller et al., 2022).

Taking into account overall performance (AUC), score distribution and ability to correctly localize lesions, the four architectures have complementary strengths and weaknesses. All explainability-based metrics suggest that the various architectures are focused on regions of the breast that generally include, but extend beyond, human annotations. The Baseline and AGN4V are better at localizing lesions, with median DICE scores equal to 0.07 and 0.09, respectively. The AUC results in comparison with the attention maps provide us with evidence that the breast possesses certain predictive characteristics for identifying malignant cases, which are not always readily interpretable, such as the presence of a lesion in our particular case as visible particularly for architectures based on transformers.

On the other hand, for both versions of MaMVT the attention maps are scattered along the breast, possibly indicating a tendency to favor global over local features. However, interpreting the resulting attention maps is challenging from a clinical perspective (Salahuddin et al., 2022; Hadjiiski et al., 2023). They suggest a complex interplay between different factors possibly related to the overall breast anatomy, but also potentially depending on spurious correlations due to, e.g., vendor

differences. More studies are needed to disentangle and provide a clinically meaningful interpretation of such dispersed attention maps.

Performance-wise, our results found that both MaMVT versions outperformed convolutional and graph-based architectures (AUC=80.1 vs. 74.9, respectively). Compared to studies that focused on single-view analysis (Matsoukas et al., 2022; Miller et al., 2022; Cantone et al., 2023), we observed a higher performance benefit associated with the use of transformers. Similarly, Chen et al. (2022) reported an increase in AUC from 75.9 (view-wise model from Wu et al. (2020)) to 81.5 (DeiT-based multi-view transformer), on a small in-house dataset. These combined findings suggest that the advantage of transformers is due to cross-view attention, whereas within each view the stronger locality bias induced by CNNs provides a significant performance advantage. At the same time, caution is needed when comparing studies in different transfer settings: ImageNet pre-training (Matsoukas et al., 2022; Cantone et al., 2023), no pre-training (Chen et al., 2022) and self-supervised pre-training (Miller et al., 2022; Zhou et al., 2023). Self-supervised pre-training appears to be the most beneficial setting for CNNs since it can be easily applied to customized networks (such as those proposed by Wu et al. (2020) and used in our study and in Chen et al. (2022)), and reduces the performance gap with respect to transformers.

All proposed architectures can be applied directly to the four mammographic views without the need for previous registration, reducing computational costs and potential errors. However, unlike transformer-based architectures the AGN4V requires prior segmentation of the breast and pectoral muscle, and the identification of pseudo-landmarks. These steps are easily performed, at low computational cost, on modern full-field digital mammographies. However, we found that the accuracy of the preprocessing step was much lower on screen-film mammography (CBIS and DDSM), and for this reason the AGN4V was fine-tuned only on the CSAW cohort, starting from the Baseline backbone. We also found that the graph-based components were very sensitive to the initialization of the backbone and did not always converge. The AGN4V architecture may further improve if trained on a larger dataset.

As a final remark, our conclusions are dependent on the size of the training set. Since transformers have weaker inductive biases and thus higher capacity to scale to larger training sets, the performance gap is likely to increase as more data becomes available (He et al., 2022). Likewise, agreement between different architectures could increased when trained on a larger dataset. Validation was also performed on a single vendor and institution: external validation on different vendors and/or institutions would shed further light on the generalization capability of different architectures.

8. Conclusions

This paper presents a comparative analysis of three different multi-view architectures for breast cancer classification: a 4-view convolutional network, a graph-based architecture, and a transformer-based architecture. Given their fundamentally different inductive biases, these architectures not only achieve different performance, but also tend to focus on different areas of the breast. Even though transformer-based architectures achieve the most promising results among the three options, the results indicate that an ensemble model can improve overall performance by increasing the AUC and reducing the false positive rate (FPR). Heatmaps were used to analyze the regions of the breast that were most relevant for each model predictions. Depending on the architecture, the selected areas were not always aligned with lesion annotations, but tended to concentrate in high density regions.

Overall, the findings highlight the potential of a wide range of multi-view architectures for breast cancer classification even in datasets of relatively modest size. Further research is needed to validate these findings on larger-scale datasets, and to enhance the ability of multi-view architectures to integrate local cues to improve the detection of small and ill-defined lesions.

CRediT authorship contribution statement

Francesco Manigrasso: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Rosario Milazzo:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Alessandro Sebastian Russo:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Fabrizio Lamberti:** Writing – review & editing, Supervision, Conceptualization. **Fredrik Strand:** Writing – review & editing, Data curation, Conceptualization. **Andrea Pagnani:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Lia Morra:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lia Morra, Fabrizio Lamberti, Andrea Pagnani report financial support was provided by HealthTriage srl. Lia Morra, Francesco Manigrasso, Alessandro Russo, Rosario Milazzo has patent pending to HealthTriage srl. Fredrik Strand reports speaker fees from Lunit and Pfizer. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge funding by HealthTriage srl.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57 (1), 289–300.

Betancourt Tarifa, A.S., Marrocco, C., Molinara, M., Tortorella, F., Bria, A., 2023. Transformer-based mass detection in digital mammograms. J. Ambient Intell. Humaniz. Comput. 1–15.

Broeders, M., Moss, S., Nyström, L., Njor, S., Jonsson, H., Paap, E., Massat, N., Duffy, S., Lynge, E., Paci, E., 2012. The impact of mammographic screening on breast cancer mortality in Europe: A review of observational studies. J. Med. Screen. 19 (1_suppl), 14–25.

Cantone, M., Marrocco, C., Tortorella, F., Bria, A., 2023. Convolutional networks and transformers for mammography classification: An experimental study. Sensors 23 (3), 1229.

Carneiro, G., Nascimento, J., Bradley, A.P., 2017. Automated analysis of unregistered multi-view mammograms with deep learning. IEEE Trans. Med. Imaging 36 (11), 2355–2365.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

Chen, P., Liu, S., Jia, J., 2021. Jigsaw clustering for unsupervised visual representation learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 11521–11530.

Chen, S., Yu, T., Li, P., 2001. MVT: Multi-view vision transformer for 3D object recognition. In: British Machine Vision Conference.

Chen, X., Zhang, K., Abdoli, N., Gilley, P.W., Wang, X., Liu, H., Zheng, B., Qiu, Y., 2022. Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. Diagnostics 12 (7), 1549.

Condon, J., Oakden-Rayner, L., Hall, K., Reintals, M., Holmes, A., Carneiro, G., Palmer, L., 2021. Replication of an open-access deep learning system for screening mammography: Reduced performance mitigated by retraining on local data. medRxiv 2021-2005 Cold Spring Harbor Laboratory Press.

Dembrower, K., Lindholm, P., Strand, F., 2019. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (CSAW). J. Dig. Imag. 33, 408–413.

Dembrower, K., Wåhlin, E., Liu, Y., Salim, M., Smith, K., Lindholm, P., Eklund, M., Strand, F., 2020. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: A retrospective simulation study. Lancet Dig. Health 2 (9), e468–e474.

Domingo, L., Hofvind, S., Hubbard, R.A., Román, M., Benkeser, D., Sala, M., Castells, X., 2016. Cross-national comparison of screening mammography accuracy measures in US, Norway, and Spain. Eur. Radiol. 26, 2520–2528.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16× 16 words: Transformers for image recognition at scale.

Du, H., Feng, J., Feng, M., 2019. Zoom in to where it matters: a hierarchical graph based model for mammogram analysis. arXiv arXiv:1912.07517.

Famouri, S., Morra, L., Lamberti, F., 2020. A deep learning approach for efficient registration of dual view mammography. In: Artificial Neural Networks in Pattern Recognition: 9th IAPR TC3 Workshop, ANNPR 2020, Winterthur, Switzerland, September 2–4, 2020, Proceedings 9. Springer, pp. 162–172.

Garcea, F., Serra, A., Lamberti, F., Morra, L., 2023. Data augmentation for medical imaging: A systematic literature review. Comput. Biol. Med. 152, 106391.

Grill, J.-B., Strub, F., Altch'e, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M., 2020. Bootstrap your own latent: A new approach to self-supervised learning. abs/2006.07733, ArXiv.

Hadjiiski, L., Cha, K., Chan, H.-P., Drukker, K., Morra, L., Näppi, J.J., Sahiner, B., Yoshida, H., Chen, Q., Deserno, T.M., et al., 2023. AAPM task group report 273: Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging. Med. Phys. 50 (2), e1–e24.

He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., Girshick, R.B., 2021. Masked autoencoders are scalable vision learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 15979–15988.

He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., Shen, D., 2022. Transformers in medical image analysis: A review. Intell. Med.

Heath, M., Bowyer, K., Kopans, D., Kegelmeyer Jr., P., Moore, R., Chang, K., et al., 1998. Current status of the digital database for screening mammography. In: Karssemeijer, N., Thijssen, M., Hendriks, J., van Erning, L. (Eds.), Digital mammography. In: Computational imaging and vision, vol. 13.

Jiang, D., Li, W., Cao, M., Zhang, R., Zou, W., Han, K., Li, X., 2020. Speech SIMCLR: Combining contrastive and reconstruction objective for self-supervised speech representation learning. Interspeech.

Jiménez-Gaona, Y., Rodríguez-Álvarez, M.J., Lakshminarayanan, V., 2020. Deep-learning-based computer-aided systems for breast cancer imaging: a critical review. Appl. Sci. 10 (22), 8298.

Khan, H.N., Shahid, A.R., Raza, B., Dar, A.H., Alquhayz, H., 2019. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. IEEE Access 7, 165724–165733.

Kyono, T., Gilbert, F.J., van der Schaar, M., 2018. MAMMO: A deep learning solution for facilitating radiologist-machine collaboration in breast cancer diagnosis. arXiv arXiv:1811.02661.

Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L., 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. Scient. Data 4 (1), 1–9.

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K., 2023. Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. Med. Image Anal. 102762.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Liu, J., 2010. Fuzzy modularity and fuzzy community structure in networks. Eur. Phys. J. B. 77, 547–557.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al., 2022. Swin transformer v2: Scaling up capacity and resolution. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021a. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. ICCV.

Liu, Y., Zhang, F., Chen, C., Wang, S., Wang, Y., Yu, Y., 2021b. Act like a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass detection. IEEE Trans. Pattern Anal. Mach. Intell. 44 (10), 5947–5961.

Lotter, W., Diab, A.R., Haslam, B., Kim, J.G., Grisot, G., Wu, E., Wu, K., Onieva, J.O., Boyer, Y., Boxerman, J.L., et al., 2021. Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach. Nat. Med. 27 (2), 244–249.

Maaz, M., Rasheed, H., Gaddam, D., 2021. Self-supervised learning for fine-grained visual categorization. abs/2105.08788, ArXiv.

Maqsood, S., Damaševičius, R., Maskeliūnas, R., 2022. TTCNN: A breast cancer detection and classification towards computer-aided diagnosis using digital mammography in early stages. Appl. Sci. 12 (7), 3273.

Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K., 2021. Is it time to replace cnns with transformers for medical images? arXiv preprint arXiv:2108.09038.

Matsoukas, C., Haslum, J.F., Sorkhei, M., Soderberg, M.P., Smith, K., 2022. What makes transfer learning work for medical images: Feature reuse & other factors. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9215–9224.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al., 2020. International evaluation of an AI system for breast cancer screening. Nature 577 (7788), 89–94.

Miller, J.D., Arasu, V.A., Pu, A., Margolies, L.R., Sieh, W., Shen, L., 2022. Self-supervised deep learning to enhance breast cancer detection on screening mammography. arXiv arXiv:2203.08812.

Misra, I., van der Maaten, L., 2019. Self-supervised learning of pretext-invariant representations. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 6706–6716.

Morra, L., Delsanto, S., Correale, L., 2019. Artificial Intelligence in Medical Imaging: From Theory to Clinical Practice. CRC Press.

Morra, L., Sacchetto, D., Durando, M., Agliozzo, S., Carbonaro, L.A., Delsanto, S., Pesce, B., Persano, D., Mariscotti, G., Marra, V., et al., 2015. Breast cancer: Computer-aided detection with digital breast tomosynthesis. Radiology 277 (1), 56–63.

Nawaz, M., Sewissy, A.A., Soliman, T.H.A., 2018. Multi-class breast cancer classification using deep learning convolutional neural network. Int. J. Adv. Comput. Sci. Appl 9 (6), 316–332.

Ou, W.C., Polat, D., Dogan, B.E., 2021. Deep learning in breast radiology: current progress and future directions. Eur. Radiol. 31, 4872–4885.

Perek, S., Hazan, A., Barkan, E., Akselrod-Ballin, A., 2018. Siamese network for dual-view mammography mass matching. In: Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer, pp. 55–63.

Pezeshk, A., Petrick, N., Sahiner, B., 2016. Seamless lesion insertion in digital mammography: methodology and reader study. In: Medical Imaging 2016: Computer-Aided Diagnosis, Vol. 9785. SPIE, pp. 134–139.

Pinto Pereira, S.M., McCormack, V.A., Moss, S.M., dos Santos Silva, I., 2009. The spatial distribution of radiodense breast tissue: A longitudinal study. Breast Cancer Res 11 (3), 1–12.

Rangayyan, R.M., Ayres, F.J., Desautels, J.L., 2007. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. J. Franklin Inst. 344 (3–4), 312–348.

Ren, Y., Lu, J., Liang, Z., Grimm, L.J., Kim, C., Taylor-Cho, M., Yoon, S., Marks, J.R., Lo, J.Y., 2021. Retina-match: Ipsilateral mammography lesion matching in a single shot detection pipeline. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. Springer, pp. 345–354.

Rodríguez-Ruiz, A., Lång, K., Gubern-Mérida, A., Broeders, M.J.M., Gennaro, G., Clauser, P., Helbich, T.H., Chevalier, M., Tan, T., Mertelmeier, T., Wallis, M.G., Andersson, I., Zackrisson, S., Mann, R.M., Sechopoulos, I., 2019. Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists. J. Natl. Cancer Instit..

Sacchetto, D., Morra, L., Agliozzo, S., Bernardi, D., Björklund, T., Brancato, B., Bravetti, P., Carbonaro, L.A., Correale, L., Fantò, C., et al., 2016. Mammographic density: comparison of visual assessment with fully automatic calculation on a multivendor dataset. Eur. Radiol. 26 (1), 175–183.

Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P., 2022. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput. Biol. Med. 140, 105111.

Samee, N.A., Atteia, G., Meshoul, S., Al-antari, M.A., Kadah, Y.M., 2022. Deep learning cascaded feature selection framework for breast cancer classification: Hybrid CNN with univariate-based approach. Mathematics 10 (19), 3631.

Samulski, M., Karssemeijer, N., 2011. Optimizing case-based detection performance in a multiview CAD system for mammography. IEEE Trans. Med. Imaging 30 (4), 1001–1009.

Schaffter, T., Buist, D.S., Lee, C.I., Nikulin, Y., Ribli, D., Guan, Y., Lotter, W., Jie, Z., Du, H., Wang, S., et al., 2020. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms. JAMA Netw. Open 3 (3), e200265.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Shen, T., Hao, K., Gou, C., Wang, F.-Y., 2021a. Mass image synthesis in mammogram with contextual information based on GANs. Comput. Methods Programs Biomed. 202, 106019.

Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S.G., Moy, L., Cho, K., et al., 2021b. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. Med. Image Anal. 68, 101908.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H., 2021. Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, Vol. 139. pp. 10347–10357.

Van Gils, C.H., Otten, J.D., Verbeek, A.L., Hendriks, J.H., 1998. Mammographic breast density and risk of breast cancer: Masking bias or causality? Eur. J. Epidemiol. 14, 315–320.

Van Schie, G., Tanner, C., Snoeren, P., Samulski, M., Leifland, K., Wallis, M.G., Karssemeijer, N., 2011. Correlating locations in ipsilateral breast tomosynthesis views using an analytical hemispherical compression model. Phys. Med. Biol. 56 (15), 4715.

van Tulder, G., Tong, Y., Marchiori, E., 2021. Multi-view analysis of unregistered medical images using cross-view transformers. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, pp. 104–113.

Varamesh, A., Diba, A., Tuytelaars, T., Gool, L.V., 2020. Self-supervised ranking for representation learning. arXiv arXiv:2010.07258.

Wei, J., Chan, H.-P., Wu, Y.-T., Zhou, C., Helvie, M.A., Tsodikov, A., Hadjiiski, L.M., Sahiner, B., 2011. Association of computerized mammographic parenchymal pattern measure with breast cancer risk: A pilot case-control study. Radiology 260 (1), 42–49.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S.a., Févry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L.L.Y., Ho, K., Weinstein, J.D., Reig, B., Gao, Y., Toth, H., Pysarenko, K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S.G., Heacock, L., Moy, L., Cho, K., Geras, K.J., 2020. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging 39 (4), 1184–1194.

Wu, E., Wu, K., Cox, D., Lotter, W., 2018. Conditional infilling GANs for data augmentation in mammogram classification. In: Image Analysis for Moving Organ, Breast, and Thoracic Images: Third International Workshop, RAMBO 2018, Fourth International Workshop, BIA 2018, and First International Workshop, TIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 3. Springer, pp. 98–106.

Xu, Y., Wei, H., Lin, M., Deng, Y., Sheng, K., Zhang, M., Tang, F., Dong, W., Huang, F., Xu, C., 2022. Transformers in computational visual media: A survey. Comput. Vis. Media 8, 33–62.

Yang, Z., Cao, Z., Zhang, Y., Tang, Y., Lin, X., Ouyang, R., Wu, M., Han, M., Xiao, J., Huang, L., et al., 2021. MommiNet-v2: Mammographic multi-view mass identification networks. Med. Image Anal. 73, 102204.

Zhang, Y.-D., Satapathy, S., Guttery, D., Gorriz, J., Wang, S., 2021. Improved breast cancer classification through combining graph convolutional network and convolutional neural network. Inf. Process. Manage. 58, 102439.

Zhang, Y., Yeung, D., 2012. Overlapping community detection via bounded nonnegative matrix tri-factorization. In: Proc. ACM SIGKDD Conf. pp. 606–614.

Zhou, Z., Luo, H., Pang, J., Ding, X., Gotway, M., Liang, J., 2023. Learning anatomically consistent embedding for chest radiography. In: BMVC: proceedings of the British Machine Vision Conference. British Machine Vision Conference. 2023.