Integrating AI for Human-Centric Breast Cancer Diagnostics: A Multi-Scale and Multi-View Swin Transformer Framework

Farnoush Bayatmakou*, Reza Taleei†, Milad Amir Toutounchian‡, Arash Mohammadi*
*Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, Canada
†Thomas Jefferson University Hospital, Philadelphia, Pennsylvania, USA
‡ College of Computing & Informatics, Drexel University, Philadelphia, Pennsylvania, USA

***Abstract*—Despite advancements in Computer-Aided Diagnosis (CAD) systems, breast cancer remains one of the leading causes of cancer-related deaths among women worldwide. Recent breakthroughs in Artificial Intelligence (AI) have shown significant promise in development of advanced Deep Learning (DL) architectures for breast cancer diagnosis through mammography. In this context, the paper focuses on the integration of AI within a Human-Centric workflow to enhance breast cancer diagnostics. Key challenges are, however, largely overlooked such as reliance on detailed tumor annotations and susceptibility to missing views, particularly during test time. To address these issues, we propose a hybrid, multi-scale and multi-view Swin Transformer-based framework (MSMV-Swin) that enhances diagnostic robustness and accuracy. The proposed MSMV-Swin framework is designed to work as a decision-support tool, helping radiologists analyze multi-view mammograms more effectively. More specifically, the MSMV-Swin framework leverages the Segment Anything Model (SAM) to isolate the breast lobe, reducing background noise and enabling comprehensive feature extraction. The multi-scale nature of the proposed MSMV-Swin framework accounts for tumor-specific regions as well as the spatial characteristics of tissues surrounding the tumor, capturing both localized and contextual information. The integration of contextual and localized data ensures that MSMV-Swin's outputs align with the way radiologists interpret mammograms, fostering better human-AI interaction and trust. A hybrid fusion structure is then designed to ensure robustness against missing views, a common occurrence in clinical practice when only a single mammogram view is available. Experimental evaluations on single-view and dual-view mammography based on CBIS-DDSM dataset demonstrate the superior performance of MSMV-Swin, highlighting its potential for improving breast cancer diagnosis in diverse clinical settings.**

***Index Terms*—Breast Cancer, Transformer, Multi-view Mammograms.**

I. Introduction

Breast cancer is one of the leading causes of cancer-related deaths among women worldwide, making early detection of significant importance [1]. Multi-view mammography [2] has emerged as a vital approach to breast cancer diagnosis, where radiologists analyze multiple views (e.g., CranioCaudal (CC) and MedioLateral Oblique (MLO)) to detect abnormalities that may not be visible in a single view. This has inspired the development of multi-view-based Computer-Aided Diagnosis (CAD) systems, which utilize information from different views to improve diagnostic accuracy. In recent years, Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs) [3], [4] and Transformers [5], [6], have significantly enhanced the performance of multi-view CAD systems. Despite recent advancements in this domain, however, there are still key challenges ahead, including reliance on detailed annotated Region of Interest (ROIs) and susceptibility to missing views, especially at the test time. Addressing these limitations is critical to advancing CAD system reliability and robustness. This paper proposes a hybrid, multi-view, and multi-scale framework that leverages state-of-the-art DL techniques to overcome these challenges. The proposed approach enhances diagnostic precision and robustness, providing a reliable tool for breast cancer diagnosis in diverse clinical contexts.

A. Related Works

Initial efforts in multi-view mammogram classification were dominated by CNN-based approaches. For instance, Reference [7] was among the first to apply multi-view analysis by training CNN models separately on CC and MLO views and combining features from these views using multinomial logistic regression. Such initial works demonstrated that multi-view analysis has the potential to outperform single-view models by leveraging the inherent correlations between different views. Subsequent research improved on these ideas by developing more effective fusion strategies. For example, Reference [8] introduced a view-wise feature merging strategy using tailored ResNet models to process CC and MLO images separately before averaging their predictions during inference. In another notable effort, Reference [3] proposed a two-stage multi-view fusion strategy for extracting ROI from four mammogram views using CNNs, which further enhanced the classification performance by integrating multi-view information.

More recent developments have shifted towards using Transformers for multi-view mammography due to their superior ability to model long-range dependencies and capture inter-view correlations. For instance, Reference [9] introduced a hybrid approach that integrated CNN feature extraction with global cross-view attention via Transformers. The cross-view fusion focused on correlating CC and MLO views to improve diagnostic accuracy, although it did not fully explore bilateral

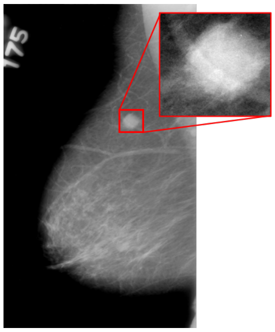

Fig. 1: Sample image with the cropped area around the designated ROI [15].

asymmetry. In another noteworthy study, Reference [6] used a pure Vision Transformer (ViT) architecture for multi-view breast cancer detection, which showed promising results. This model processed views at a later stage of the network, focusing on retaining local features while integrating global multi-view information, though it still missed some early correlations essential for optimal classification. The introduction of the Swin Transformer [10] architecture, specifically designed for computer vision tasks, has been transformative in multi-view breast cancer classification. Recently, Reference [11] proposed a pure Transformer-based multi-view architecture built upon the Swin Transformer, which incorporated novel shifted window attention mechanisms to integrate multi-views at the spatial feature level. By fusing information from the CC and MLO views early in the network, the model captured critical cross-view correlations, significantly outperforming traditional CNN and hybrid models.

Despite the above-mentioned surge of interest in DL-based multi-view breast cancer detection and achieving exceptional results, the following two major challenges have been overlooked in the literature, which is the focus of this study: noitemsep, nolistsep

(i) Most of the existing models work best when provided with cropped mammograms focused on the tumour region, which ignores the spatial properties of the tissues surrounding the tumor. Such properties play a crucial role for radiologists and should be incorporated into a DL pipeline.

(ii) While multi-view DL architectures showed superiority against their single-view counterparts, it is quite common in practice to have cases with only a single view. Robustness against missing views at test time has been mostly overlooked in the literature.

The paper targets addressing these issues as outlined below.

B. Contribution

Capitalizing on the above discussion, the paper introduces a hybrid, multi-scale, and multi-view Swin Transformer-based architecture, referred to as the MSMV-Swin framework. The multi-scale nature of the MSMV-Swin framework is introduced following the success of our multi-scale learning architecture, the 3D-MCN [12] in lung nodule malignancy. The multi-scale learning is incorporated to take into account spatial characteristics of tissues surrounding the tumor, addressing the first identified issue. The MSMV-Swin framework is further designed to tackle the second challenge by introducing robustness against missing views during test time through a feature zero padding strategy. More specifically, the proposed framework incorporates different scales of each mammogram. These scales are obtained using the Segment Anything Model (SAM) [13], capable of segmenting any discernible feature within a mammogram. This capability is crucial for extracting detailed representations of breast tissues in both CC and MLO views. As shown in Fig. 1, two scales are considered: (i) *Masked scale*, which is obtained by segmented bounding boxes around the identified breast lobe, and (ii) *Cropped scale*, which focuses on regions around the pre-existing ROI with tumors. Cropped images provide enhanced details and can improve accuracy in the analysis of critical tumor regions. Conversely, the utilization of a segmented scale allows for an extensive examination of both the localized tumor areas and the broader anatomical features, including their position, shape, and relative size within the breast anatomy. Such a multi-scale radiomics approach provides essential insights into the interactions between tumors and their surrounding tissues, which is critical for delivering precise diagnostic outcomes. To optimize feature extraction from both CC and MLO views, we employ a pre-trained model, Shifted Windows (Swin) Transformer [14]. These models facilitate robust feature extraction, which is further enhanced through various feature fusion techniques. The fused features are then channelled through a sequence of Multi-Layer Perceptrons (MLPs) to achieve final classification results.

II. Problem Formulation and Proposed MSMV-Swin Framework

The MSMV-Swin framework leverages a dual Swin Transformer architecture to process distinct scales of mammographic image inputs—specifically, cropped and segmented variants. The processing pipeline is designed to handle separate streams for CC and MLO views for each of the two underlying scales. This dual-path approach ensures comprehensive feature extraction tailored to the specific characteristics of each image type.

A. Model Architecture and Training

As the pre-processing step, we utilized a well-trained SAM for the segmentation of original mammograms. This approach is beneficial in enhancing the learning process of the MSMV-Swin framework by removing the background areas from mammogram images to isolate the breast lobe.

The extracted breast lobe is then passed into two parallel paths of the proposed MSMV-Swin framework. More specifically, extracted breast areas after pre-processing are downsampled to $(224, 224)$ pixels to reduce the complexity and memory allocation without significant loss of information. Afterwards, we leverage a pre-trained Swin Transformer for feature extraction, which is known for its efficacy in processing

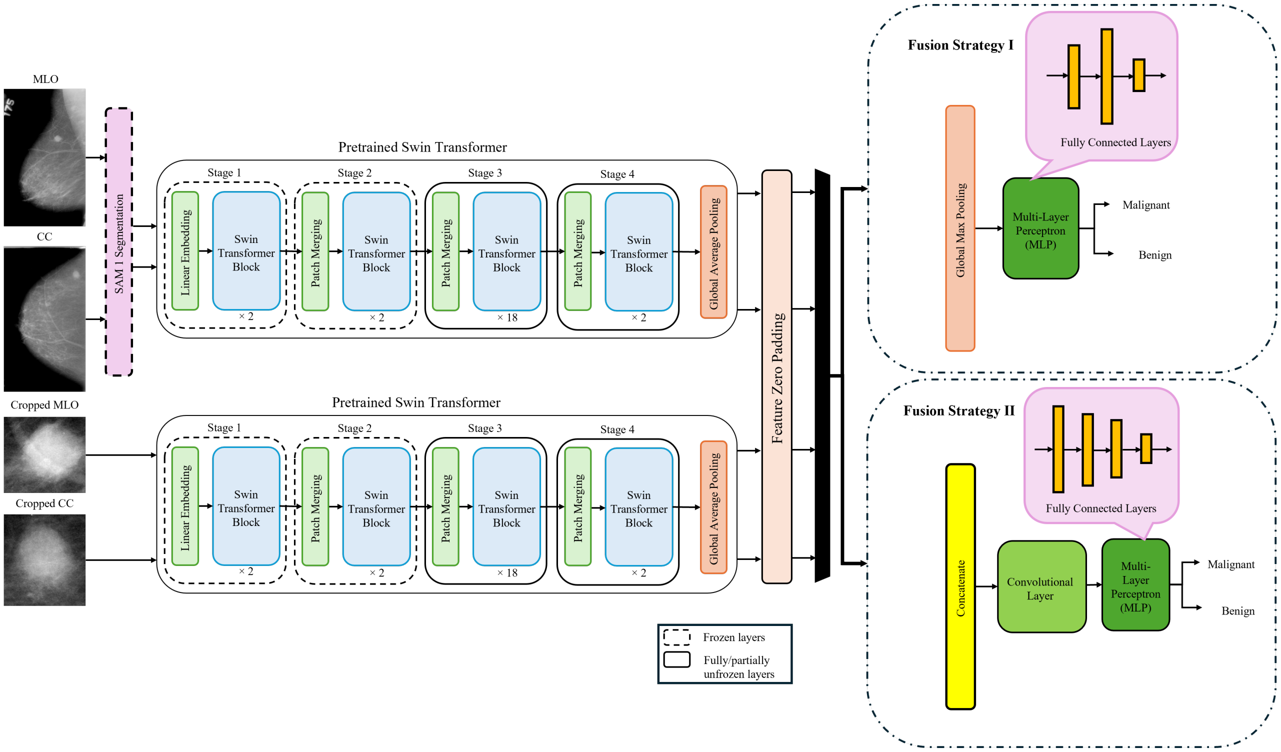

Fig. 2: The proposed MSMV-Swin framework with feature fusion through (I) max-pooling, and (II) Convolutional and MLP.

complex image structures. We have also used two different feature fusion strategies to combine the extracted features to be fed to an MLP for the final classification step.

The feature extraction and processing pipeline is divided into two distinct paths: (i) *Segmented Image Path*: This branch processes features from segmented CC and MLO views, forming a comprehensive feature set that provides a complete representation of each segmented image, facilitating a detailed analysis across various diagnostic image types, and; (ii) *Cropped Image Path*: Similarly, this pathway focuses on features extracted from specifically cropped areas within the CC and MLO views. Considering these features ensures that critical areas of interest are emphasized, enhancing the scrutiny and diagnostic potential for regions that may exhibit pathological features.

B. Feature Extraction Framework

Pre-trained Swin Transformer is deployed in dual configurations given its capabilities to process images hierarchically and capture long-range dependencies efficiently through a shifted window strategy, which is essential for integrating both local and global contextual features effectively. The architecture integrates dual instances of Swin Transformers, each fit to the unique characteristics of the corresponding scale and set to independently process segmented and cropped images. Such an approach facilitates precise feature extraction from essential regions, making it particularly suitable for detailed medical diagnostics of the two views. Initially, all parameters in each Swin Transformer instance are frozen, except the top layers, which are selectively unfrozen during fine-tuning to preserve learned general features while adapting to the underlying dataset's specific characteristics. Strategic unfreezing of a subset of the last layers during the training phase customizes the model to the special features of CC and MLO images and the targeted areas in cropped images.

After feature extraction using Swin Transformers, each of the two instances associated with the two scales processes two views, resulting in a total of four feature sets of dimension $1,024$. Let these features be denoted by: (i) $\boldsymbol{F}_{\text{CC}}^{Seg}$: Features from segmented CC images; (ii) $\boldsymbol{F}_{\text{MLO}}^{Seg}$: Features from segmented MLO images; (iii) $\boldsymbol{F}_{\text{CC}}^{Crop}$: Features from cropped CC images, and; (iv) $\boldsymbol{F}_{\text{MLO}}^{Crop}$: Features from cropped MLO images. Moreover, we applied a feature zero padding approach before the feature fusion step to handle cases with missing views. In scenarios where either the CC or MLO view is unavailable, the corresponding feature vector is set to zero. This ensures that no invalid data is propagated through the network while still allowing the available view to contribute fully to the final feature representation. By preserving the integrity of the feature fusion process, this approach maintains the robustness of the framework even when dealing with incomplete input data, a common occurrence in clinical practice. The final outputs of the feature extraction framework are given by

$$\boldsymbol{F}_{\text{CC}}^{Seg} = \begin{cases} \text{Swin}\left(\boldsymbol{I}_{\text{CC}}^{Seg}\right) & \text{if CC view exists} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\boldsymbol{I}_{\text{CC}}^{Seg}$ represents the segmented mammogram image from the CC view. The other feature vectors, after applying feature zero padding, are defined in a similar manner.

C. Fusion Framework

After extracting the features from the existing views, the four feature vectors are subsequently fused to form a unified feature map for downstream processing, given by

$$\boldsymbol{F}_{\text{combined}} = f\left(\boldsymbol{F}_{\text{CC}}^{Seg}, \boldsymbol{F}_{\text{MLO}}^{Seg}, \boldsymbol{F}_{\text{CC}}^{Crop}, \boldsymbol{F}_{\text{MLO}}^{Crop}\right), \qquad (2)$$

where function $f(\cdot)$ is the fusion function. The following two fusion strategies are implemented.

1) Max-Pooling Fusion: Under this fusion strategy, the feature vector $\boldsymbol{F}_{\text{combined}}$ is constructed by concatenating the four features extracted from both cropped and segmented mammograms across CC and MLO views. The resultant feature map undergoes a reduction process where the maximum values are selected across the concatenated features. This operation is performed across the dimension that holds different views (CC and MLO) and segmentation statuses (segmented and cropped), effectively capturing the most critical features. This max-pooling step ensures that only the most relevant features, which have the highest values post concatenation, are passed forward, enhancing the model's focus on potentially critical diagnostic information.

The reduced feature map is then directed through a robust MLP classification head, which comprises successive layers of linear transformations and non-linear activation functions. Specifically, LeakyReLU is used to facilitate smooth gradient flow together with dropout layers to mitigate overfitting. The architecture of the MLP features layer sizes that progressively decrease from 1,024 to 512. The idea is to ensure the refinement of the feature set into highly discriminative elements essential for accurate classification. Following the feature compression down to 512 units, the model's architecture culminates in a single output layer. This layer directly projects the 512-dimensional feature vector to a single logit, indicative of the binary classification task.

2) Convolution-based Fusion: In this fusion strategy, outputs from the dual transformers are concatenated into a 1024×4 array and passed through a custom-designed convolutional block. A convolutional layer further refines extracted features by enhancing spatial relationships and reducing dimensionality, thus focusing the model's attention on the most significant aspects of the image. This combination not only maximizes the feature extraction capabilities of Swin Transformers but also leverages the robust, spatial feature processing of convolutional networks, significantly improving the accuracy and reliability of image classification tasks by ensuring a comprehensive and refined analysis of input data. The incorporated block consists of a 2D convolutional layer with a kernel size of 3×3, followed by batch normalization, a ReLu activation function, and finally, a max-pooling layer.

The refined features are then flattened and fed into an MLP layer, as previously described. It is to be noted that as the only difference, the input from the convolution block to the MLP is of size 4096.

III. Experimental Result

A. Dataset

The proposed MSMV-Swin framework is developed based on the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) dataset [15]. CBIS-DDSM is a widely recognized mammography dataset and comprises $10,239$ images, including whole mammograms, cropped images, and ROI mask images with mass and calcification. A sample of mammograms from the CBIS-DDSM and its cropped version in the designated ROI is illustrated in Fig. 1. To achieve results comparable to other studies utilizing the CBIS-DDSM dataset, we assigned patients to train and test subsets following the splitting presented in [15]. In our research, we extracted both cropped and whole mammograms from the CBIS-DDSM dataset, specifically for patients who had unique images from both the CC and MLO views. For each breast, we considered the two mammograms as a single examination, resulting in a total of 477 breast samples for training and 144 for testing.

While we only have those data having both views for training, i.e., excluding 208 one-view breast samples from the training set, 59 cases of one-view breasts are considered in the test-time. In other words, while the MSMV-Swin framework is trained in a multi-view setting, it is designed in such a way as to classify single-view cases as well at test time. To increase the training dataset, data augmentation was used where each mammogram is augmented by 90, 180, and 270-degree rotations, as well as horizontal and vertical flipping. Following this augmentation approach, five new data points are generated and added to the original breast sample. The same type of augmentation is applied on the CC and MlO views, resulting in an increase in the size of the training and test datasets by a factor of 6 (477×6, 144×6, and 59×6, respectively, for multi-view training and test datasets, and single-view test dataset).

To compensate for the missing views (missing modality) at the test-time (patients with either CC or MIO view), different approaches can be used borrowing from the works on missing modality in multi-modal learning [16]–[19]. One naive approach is to zero-pad the missing view. This, however, is expected to deteriorate the results as observed through our empirical evaluations. An alternative approach could be to simulate the missing view based on the available one. In other words, the missing data can be imputed using generative models; the imputation process, however, may introduce unrealistic information to the classification process, leading to poor performance. Instead, we opted for feature zero padding for the missing view. This method ensures that the representation remains consistent and avoids introducing potentially misleading synthetic data. Additionally, by focusing on the feature space rather than the raw input space, we mitigate the

TABLE I: MSMV-Swin framework with Swin-based back-end.

Test Data	**Accuracy (%)**	**AUC (%)**	**F1 Score (%)**	**Sensitivity (%)**	**Specificity (%)**
Swin with Max-Pooling Fusion					
Breasts with both Views	**80.32**	82.27	75.07	72.32	**85.88**
Breasts with Missing Views	77.40	**84.20**	74.36	**80.56**	75.24
All Breasts	*79.06*	*81.50*	*74.42*	*74.50*	*82.22*
Swin with Convolution-based Fusion					
Breasts with both Views	**79.05**	84.67	74.54	74.86	**81.96**
Breasts with Missing Views	77.68	**85.57**	74.76	**81.25**	75.24
All Breasts	77.50	83.73	73.19	75.10	79.17

TABLE II: Compression with state-of-the-art.

Model	Variants	Accuracy (%)	AUC (%)	F1 Score (%)	Sensitivity (%)	Specificity (%)
TMSMV-Swin Framework Convolution Fusion	Both-Views Missing-Data All Breasts	80.32 77.40 79.06	82.27 84.20 84.67	75.07 74.36 74.42	72.32 80.56 74.86	85.88 75.24 81.96
Reference [11]	Both-Views Missing-Data	68.63 -	71.37 -	- -	- -	- -
Reference [21]	Both-Views Missing-Data	73.00 ± 1.90 -	80.90 ± 0.50 -	– -	– -	71.00 ± 2.00 -
Reference [20]	Both-Views Missing-Data	72.7 ± 1.91 -	– -	– -	– -	– -
Reference [22]	Both-Views Missing-Data	70.9 -	80.5 ± 0.2 -	70.9 -	– -	– -
Reference [23]	Both-Views Missing-Data	– -	80.0 -	65.0 -	– -	– -
Reference [24]	Both-Views Missing-Data	68.24 -	74.21 -	65.38 -	62.96 -	– -

negative impact of missing modalities while preserving the integrity of the classification process.

B. Results

The model is trained using a binary cross-entropy with a label smoothing loss function. This loss function helps improve model generalization by penalizing the model less severely on hard examples, effectively smoothing the decision boundary. We used AdamW optimizer, chosen for its ability to combine the benefits of Adam optimization and weight decay regularization, making it particularly suited for this kind of complex model training. Validation is conducted in parallel with training using a separate subset of the data. This approach allows for continuous monitoring of the model's performance and generalization capabilities, ensuring robustness before deployment. Metrics such as accuracy, sensitivity, specificity, and the area under the ROC curve (AUC) are computed to provide comprehensive insights into the model's performance across various thresholds and conditions.

Table I illustrates the results obtained from Swin-based back-ends. These results compare two different fusion mechanisms and the performance of the model with missing views. As can be observed, max-pooling fusion seems to be slightly better than the alternative fusion scenarios. This is an interesting observation showing complex fusion is not required, which can contribute to the strength of the feature extraction pipeline. At the same time, it is observed that the performance degradation due to missing views is quite negligible, illustrating the robustness of the proposed pipeline. Finally, Table II illustrates comparison results with recent state-of-the-art [11], [20], [21] further corroborating superiority of the proposed MSMV-Swin framework.

IV. Conclusion

In this paper, we introduced the MSMV-Swin framework, a hybrid, multi-scale and multi-view Swin Transformer-based architecture aimed at addressing two key challenges in breast cancer diagnosis through mammography, i.e., (i) Reliance on detailed tumor annotations and (ii) Susceptibility to missing views during test time. The MSMV-Swin framework distinguishes itself by utilizing a novel approach that integrates localized and contextual characteristics of tissues surrounding tumors. In other words, by leveraging the Segment Anything Model (SAM), the framework extracts detailed multi-scale information from both CC and MLO views, effectively capturing both localized and contextual features. Additionally, two different feature fusion strategies were employed to ensure robustness by compensating for missing views without relying on synthetic data. Experimental results demonstrate that the MSMV-Swin framework consistently outperforms existing models. Notably, it achieves superior accuracy and robustness even in scenarios where only single-view mammograms are available, with a negligible performance drop compared to dual-view cases. These findings underscore the potential of the MSMV-Swin framework to advance reliable and comprehensive breast cancer diagnostics in real-world clinical settings.

References

[1] E. I. Obeagu and G. U. Obeagu, "Breast cancer: A review of risk factors and diagnosis," *Medicine*, vol. 103, no. 3, p. e36905, 2024.

[2] A. Jouirou, A. Baâzaoui, and W. Barhoumi, "Multi-view information fusion in mammograms: A comprehensive overview," *Information Fusion*, vol. 52, pp. 308–321, 2019.

[3] H. N. Khan, A. R. Shahid, B. Raza, A. H. Dar, and H. Alquhayz, "Multi-view feature fusion based four views model for mammogram classification using convolutional neural network," *IEEE Access*, vol. 7, pp. 165 724–165 733, 2019.

[4] L. Sun, J. Wang, Z. Hu, Y. Xu, and Z. Cui, "Multi-view convolutional neural networks for mammographic image classification," *IEEE Access*, vol. 7, pp. 126 273–126 282, 2019.

[5] X. Zhao, L. Yu, and X. Wang, "Cross-view attention network for breast cancer screening from multi-view mammograms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1050–1054.

[6] X. Chen, K. Zhang, N. Abdoli, P. W. Gilley, X. Wang, H. Liu, B. Zheng, and Y. Qiu, "Transformers improve breast cancer diagnosis from unregistered multi-view mammograms," *Diagnostics*, vol. 12, no. 7, p. 1549, 2022.

[7] G. Carneiro, J. Nascimento, and A. P. Bradley, "Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions," *Deep learning for medical image analysis*, pp. 321–339, 2017.

[8] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzbski, T. Fvry, J. Katsnelson, E. Kim *et al.*, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE transactions on medical imaging*, vol. 39, no. 4, pp. 1184–1194, 2019.

[9] G. Van Tulder, Y. Tong, and E. Marchiori, "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, 2021, pp. 104–113.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[11] S. Sarker, P. Sarker, G. Bebis, and A. Tavakkoli, "Mv-swin-t: Mammogram classification with multi-view swin transformer," *arXiv preprint arXiv:2402.16298*, 2024.

[12] P. Afshar, A. Oikonomou, F. Naderkhani, P. N. Tyrrell, K. N. Plataniotis, K. Farahani, and A. Mohammadi, "3d-mcn: a 3d multi-scale capsule network for lung nodule malignancy prediction," *Scientific reports*, vol. 10, no. 1, p. 7948, 2020.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

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

[15] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific data*, vol. 4, no. 1, pp. 1–9, 2017.

[16] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 878–15 887.

[17] Y. Chen, Y. Pan, Y. Xia, and Y. Yuan, "Disentangle first, then distill: A unified framework for missing modality imputation and alzheimer's disease diagnosis," *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3566–3578, 2023.

[18] F. Ma, X. Xu, S.-L. Huang, and L. Zhang, "Maximum likelihood estimation for multimodal learning with missing modality," *arXiv preprint arXiv:2108.10513*, 2021.

[19] D. Shi, L. Zhu, J. Li, G. Dong, and H. Zhang, "Incomplete cross-modal retrieval with deep correlation transfer," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, pp. 1–21, 2024.

[20] K. L. dos Santos and M. P. dos Santos Silva, "Deep cross-training: An approach to improve deep neural network classification on mammographic images," *Expert Systems with Applications*, vol. 238, p. 122142, 2024.

[21] G. I. Quintana, Z. Li, L. Vancamberg, M. Mougeot, A. Desolneux, and S. Muller, "Exploiting patch sizes and resolutions for multi-scale deep learning in mammogram image classification," *Bioengineering*, vol. 10, no. 5, p. 534, 2023.

[22] S. Yang, C. Zhang, Q. Zang, J. Yu, L. Zeng, X. Luo, Y. Xing, X. Pan, Q. Li, X. Liang *et al.*, "Mammo-clustering: A weakly supervised multi-view global-local context clustering network for detection and classification in mammography," *arXiv preprint arXiv:2409.14876*, 2024.

[23] H. Allaoui, Y. Alj, and Y. Ameskine, "Hybridmammonet: A hybrid cnn-vit architecture for multi-view mammography image classification," in *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*. IEEE, 2024, pp. 1–6.

[24] L. Liao and E. M. Aagaard, "An open codebase for enhancing transparency in deep learning-based breast cancer diagnosis utilizing cbis-ddsm data," *Scientific Reports*, vol. 14, no. 1, p. 27318, 2024.