STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, February 2026, pp 1087–1098.
Published online in International Academic Press (www.IAPress.org)

Vision Transformers for Breast Cancer Mammographic Image Classification

Elmehdi ANIQ [1,2,*], Faouzi-Ayoub EL GHANAOUI [1], Mohamed CHAKRAOUI [1]

[1] *LS2ME, Polydisciplinary Faculty of khouribga, Sultan Moulay Slimane University, Béni mellal, Morocco*
[2] *LAMIGEP, EMSI Marrakech, Marrakech, Morocco*

Abstract **Background and Objective :** The mortality rates due to breast cancer have been constantly growing and still represent one of the most common malignancies leading to death in females globally. Early and accurate detection is crucial to improve the survival rate. Recent deep learning advancements in artificial intelligence have opened a wide new avenue for further improving the results of computer-aided diagnosis. Vision transformers with their attention mechanism are among the recent promising ones, offering much-improved results for different image analysis applications, including mammography.

Methods : This study investigates the application of vision transformers and attention mechanisms for mammography image categorization. In this work, we used three publicly available datasets like the Mammographic Image Analysis Society (MIAS), Curated Breast Imaging Subset of DDSM (CBIS-DDSM), and INbreast. In the preprocessing of data, augmentation is used to enhance the generalization capabilities of models, and we have applied Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve the contrast of images, especially in situations characterized by uneven lighting or low contrast levels.

Results : The proposed approach demonstrated superior performance compared to traditional convolutional neural network (CNN)-based methods. In the evaluation of this vision transformer, we have obtained an accuracy of 0.99, an AUC of 0.99 and an F1 score of 0.98.

Conclusion : Vision transformers and attention mechanisms have great potential to boost the detection of breast cancer using CAD systems. The findings accentuate their capability to improve the precision and reliability of mammography analysis, enabling early diagnosis and minimizing false positives and false negatives in clinical practice. The research emphasizes the need to embrace these new technologies to enhance patient outcomes and streamline healthcare resources.

Keywords Breast cancer detection, Mammography, Computer-aided diagnosis (CAD), Vision transformers (ViTs), Attention mechanisms, Deep learning, Image classification, Artificial intelligence (AI), Feature extraction, Medical image analysis

DOI: 10.19139/soic-2310-5070-2539

1. Introduction

Breast cancer is still one of the major causes of cancer-related death in women worldwide, with its incidence continuing to rise across various regions. Early and accurate detection significantly improves treatment outcomes and survival rates, making imaging technologies a cornerstone in breast cancer diagnosis. Mammography is currently the standard for breast cancer screening due to its ability to detect abnormalities such as masses and

[*] Correspondence to: Elmehdi ANIQ (Email: elmehdi.aniq@gmail.com).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

microcalcifications, which are critical indicators of malignancy. However, traditional mammography is not without limitations, including variability in interpretation, false positives, and reduced sensitivity in dense breast tissues [[1], [2], [3]]. These challenges underscore the need for innovative techniques to enhance mammographic image analysis.

In recent years, deep learning has emerged as a transformative technology in medical imaging [4], offering unparalleled performance in image classification, segmentation, and feature extraction [[5, 6, 7]]. Convolutional Neural Networks (CNNs) have dominated the front in the area for many years, with excellent performance in breast cancer detection tasks[8]. However, CNNs are plagued by their focus on local features and inability to model global dependencies in images needed to understand complex structures in mammograms [[9], [10]].

Vision Transformers (ViTs), a novel deep learning model based on attention mechanisms, have recently attracted attention for their ability to successfully model long-range dependencies in images. By dividing an image into patches and taking them as sequences, ViTs apply self-attention to observe local and global patterns, thereby rendering them suitable for analyzing high-resolution medical images like mammograms [[11], [12]]. Furthermore, attention mechanisms help ViTs to concentrate on the most relevant areas of an image, which can make diagnosis more accurate and limit the cases of false positives and false negatives in breast cancer diagnosis.

The research seeks to investigate how ViTs and attention mechanisms can be applied to classify mammography images in breast cancer diagnosis. Publicly available datasets like MIAS, CBIS-DDSM, and INbreast will be used to confirm the efficacy of these models. This project will leverage the strengths of ViTs to amend the shortcomings of the existing approaches, developing a strong and versatile solution to enhance the precision and reliability of breast cancer diagnosis in medicine.

2. RELATED WORKS

Research on breast cancer detection by means of medical imaging has experienced rapid evolution due to the onset of machine learning and deep learning. In this section, related works will be discussed mainly devoted to traditional methods, convolutional neural networks, and the emerging use of Vision Transformers.

2.1. Traditional Methods in Mammography Analysis

Early approaches for breast cancer detection relied on handcrafted feature extraction techniques and classical machine learning algorithms. Methods such as histogram-based texture analysis, wavelet transforms, and support vector machines (SVMs) were commonly used to classify mammographic images into benign or malignant categories. While these methods showed promise, it must be acknowledged that accuracy and applicability of features across datasets were constrained. The lack of robustness to noise and variations in image quality further restricted their clinical applicability [[13]].

2.2. Deep Learning with Convolutional Neural Networks (CNNs)

Convolutional Neural Networks revolutionized mammography analysis by automating feature extraction and demonstrating superior performance in image classification and segmentation tasks. Studies such as those by Shen et al. [[14]] and Kooi et al. [[15]] showed that CNN-based models could outperform traditional methods in detecting breast cancer. Transfer learning, where pre-trained CNNs like VGG16, ResNet, and DenseNet were fine-tuned on mammography datasets, became a popular approach to address the limited size of medical imaging datasets [[16]].

Although CNNs have shown remarkable results, their reliance on local receptive fields and convolution operations often limits their ability for capturing global dependencies in images. This is a critical drawback for mammography, where subtle global patterns such as architectural distortions may indicate malignancy.

2.3. Vision Transformers (ViTs) in Medical Imaging

Besides that, ViTs introduced by Dosovitskiy et al. [[11]] have attracted attention as able to model long-range dependencies in images. Incorporating self-attention mechanisms does allow ViTs to focus on the most relevant regions in high-resolution images, making them viable for medical imaging tasks. Recent works have turned towards using ViTs for breast cancer detection. For instance, Wang et al. [[17]] aimed at using transformer-based architectures to classify mammographic images, and they garnered state of the art results on available datasets. Chen et al. [[18]] further proved that a good combination between the ViTs and data augmentation techniques and ensembling turns to be especially useful for improving classification accuracy and for reducing false positive rates. The studies above have shown the capability of ViTs to really shine in overcoming CNN limitations in mammography analysis.

2.4. Challenges and Future Directions

Even though ViTs show great promise, there are still challenges: huge computational requirements and a reliance on large-scale labeled datasets. Current research efforts are focused on addressing these issues through the development of lightweight transformer architectures and semi-supervised learning, making ViTs accessible for clinical applications [[12]].

3. METHODS

3.1. DATASET

The present study utilizes a dataset [1] composed of three well-known and publicly available mammography datasets [1]: MIAS [[19]], the INbreast dataset [[20]], and the Curated Breast Imaging Subset of the DDSM (CBIS-DDSM) dataset [[21]]. These datasets are appropriate for the study of mammograms for breast cancer and the training of deep learning models, each providing certain advantages.

- **MIAS** is suited for the foundational research and small studies. it make ideal for developing and testing initial models with the help of its small size and detailed annotations.
- **INbreast** due to its high-resolution digital images and comprehensive lesion annotations its optimal for advanced deep learning techniques, and also including BI-RADS assessments, mass shapes, and margins.
- **CBIS-DDSM** excels in large-scale studies, providing a diverse and extensive collection of cases that facilitate the training of robust models capable of generalizing across varied data distributions.

By combining these datasets, the study leverages their complementary strengths to address various challenges in breast cancer imaging. This integration enables the development of deep learning models that are both versatile and effective in analyzing mammographic images, improving detection and diagnosis outcomes across diverse scenarios.
To ensure label consistency across datasets, MIAS annotations were aligned with BI-RADS categories by expert-guided mapping.

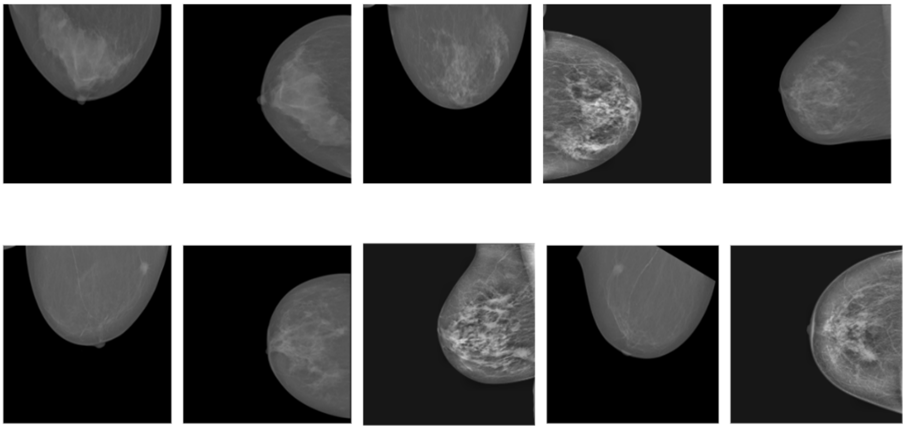

Figure 1. Examples of images dataset: the first line for the benign examples and the second one for the malignant examples

Dataset	Details
MIAS	161 cases, 322 images, MLO views, all types of anomalies, normal, benign, and malignant categories, BI-RADS classifications
INbreast	115 cases, 410 images, MLO and CC views, all types of anomalies, detailed lesion annotations, benign and malignant categories, BI-RADS classifications
CBIS DDSM	1597 cases, 3061 images, MLO and CC views, all types of anomalies, normal, benign, and malignant categories, BI-RADS classifications

Table 1. The details of the datasets

3.2. METHODOLOGY

The Vision Transformer (ViT) represents a significant advancement in deep learning for image processing tasks, introduced by Dosovitskiy et al. [[11]]. Compared to traditional convolutional neural networks (CNNs), ViTs utilize the self-attention mechanism from the transformer architecture, The original design was intended for natural language processing tasks [[22]]. This mechanism enables ViTs to capture global relationships within an image by dividing it into smaller patches, treating each patch as a sequence element.

Essentially, ViTs execute image processing in a holistic manner using self-attention mechanisms capable of modeling long-range dependencies with much effectiveness. Also, this paradigm shift has demonstrated competitively good performance on various computer vision benchmarks, showing their versatility beyond the initially intended application paradigm for transformers. They are also particularly remarkable for their scalability, with batch sizes as large as 400, allowing them to reach state-of-the-art results with adequate availability of data and computational resources [[23], [24]].

This study examined the use of Vision Transformers (ViTs) for classifying breast cancer from mammography images. Basically, ViTs process images that are divided into patches using self-attention, and apply dense layers for the classification process.

3.3. Preprocessing and Patch Generation

The input image of size 256x256x3 (height, width, and channels) is divided into N patches of size 16x16. The Number of patches N is calculated as :

$$N = \left(\frac{H}{P}\right) \cdot \left(\frac{W}{P}\right) \tag{1}$$

Where H and W are the height and width of the image, respectively, and P is the patch size. For our configuration:

$$N = \left(\frac{256}{16}\right)^2 = 256 \tag{2}$$

Each patch is flattened into a vector of size P².C = 16².3 = 768 where C is the number of channels (RGB).

3.4. Positional Encoding and Embedding

To encode spatial information, a learnable positional encoding vector is added to each patch embedding:

$$z_0^i = x^i + e^i, \quad i \in \{1, 2, \ldots, N\} \tag{3}$$

where xî is the embedded vector of the i-th patch, and eî is its positional encoding.

3.5. Multi-Head Self-Attention Mechanism

The attention mechanism identifies relationships between patches. Using 12 parallel attention heads, the attention score for each patch is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4}$$

where :

- Q, K, V are the query, key, and value matrices derived from patch embeddings.
- dk represent the dimensionality of the key vectors, ensuring stable gradients. The output of the attention layer maintains the shape (256, 768), with strong inter-patch relationships highlighted.

The query (Q), key (K), and value (V) matrices are obtained via linear projections:

$$Q = W_Q \cdot Z_0, \quad K = W_K \cdot Z_0, \quad V = W_V \cdot Z_0$$

where Z_0 is the patch embedding vector and W_Q, W_K, W_V are trainable weight matrices.

3.6. Classification Pipeline

- Flattening Layer: The attention output is flattened into a single vector of size:

$$\text{Output Vector} = N \cdot d = 256 \cdot 768 = 196,608 \tag{5}$$

- Batch Normalization and Dropout:

  - Batch normalization normalizes features to accelerate convergence and stabilize training.

$$\hat{x} = \frac{\sigma^2 + \epsilon}{x - \mu} \tag{6}$$

  - Dropout is applied with a probability p to prevent overfitting.
- Fully Connected Layers:
  - A dense layer with 2000 neurons followed by ReLU activation reduces dimensionality

$$\mathrm{ReLU}(x) = \max(0, x) \tag{7}$$

  - A second dense layer with 1000 neurons followed by ReLU activation further processes the features.
- Final Classification Layer:
  - The final layer consists of a single neuron with the activation function sigmoid to output the probability of the image belonging to the positive class.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{8}$$

3.7. Model Output

The final output is a binary classification where:

- $\sigma(x) \geq 0.5$: Positive class (malignant).
- $\sigma(x) < 0.5$: Negative class (Benign).

In summary, we presented the architecture of the Vision Transformer (ViT) used in our model[2], which is tailored for the classification of mammography images relevant to breast cancer detection. The design process starts with the input image, which is resized to 256×256×3 in dimensions and then split into non-overlapping patches of size 16×16×3. Each patch is then embedded and enriched with positional encodings to retain spatial relationships. These embeddings, when passed through a 12-head multi-head self-attention mechanism, further enable the model to build inter-patch dependencies very vital for making decisions in the process. The results are then processed through normalization, dropout layers to overcome overfitting, and dense layers of up to three hierarchical features; finally, a sigmoid output activation function yields the final result of binary classification, given whether an image is malignant or benign. This architecture emphasizes scalability and interpretability, which are in line with the state-of-the-art deep learning for medical imaging.

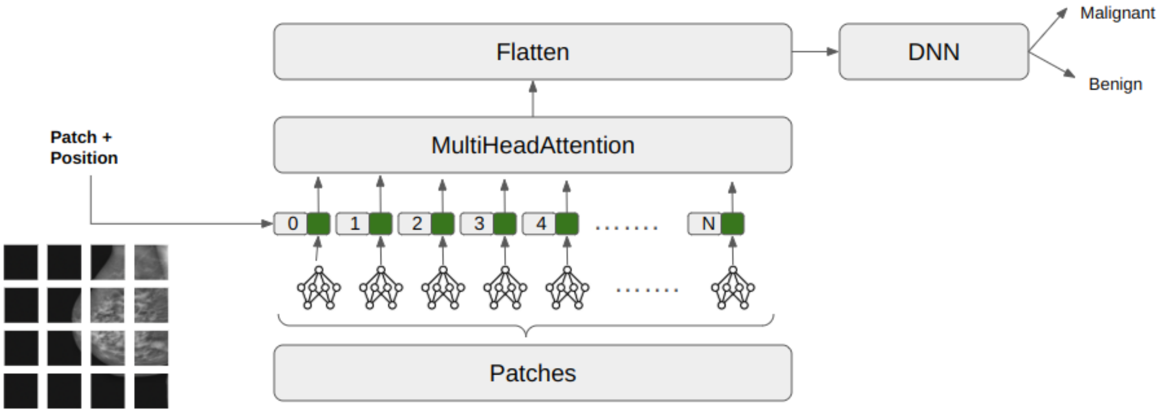

Figure 2. Architecture ViT of our model

3.8. Data preprocessing

3.8.1. Contrast Limited Adaptive Histogram Equalization Initially, after the fusion of the different datasets and during the data preprocessing stage, we applied Contrast Limited Adaptive Histogram Equalization (CLAHE)[3] to enhance the contrast of the mammography images. This technique improves local contrast while preventing over-amplification of noise, ensuring better visibility of critical features in the images. By doing so, it facilitates more effective feature extraction, which is essential for robust performance in subsequent deep learning model training. This step ensures that subtle patterns within the mammograms, which are often critical for breast cancer diagnosis, are more distinguishable and accurately detected.

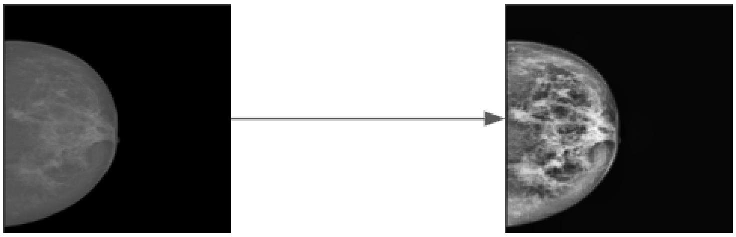

Figure 3. Application of CLAHE

3.8.2. Data augmentation The image dataset initially consists of a total of 3793 images, a quantity deemed insufficient for effective deep learning training. Consequently, we opted to commence with an augmentation phase, recognizing its crucial role in enhancing the robustness and effectiveness of the training process. seven transformations were selected for augmentation [Figure-4], including horizontal and vertical flipping, zoom, deformation, rotation, translation, and the introduction of random noise. Through this augmentation phase, we generated a significantly expanded dataset, resulting in a total of 24576 images, thereby addressing the limitations posed by the initial dataset size for more effective deep learning model training.

4. RESULTS AND DISCUSSION

To get the best performance, our model was trained on a balanced dataset with 10,000 images per class, with a 30 epochs. An epoch used 64 randomly sampled images with a stable learning rate of 0.00001. We also used the Binary Cross-Entropy loss function along with the AdamW optimizer to ensure effective and precise updates of the parameters during training. AdamW optimizer was used with β_1 = 0.9, β_2 = 0.999, and weight decay = 0.01. Similarly, 4,576 images were assigned to validation of how well the model generalizes and avoids dangers due to overfitting or underfitting.

We employed several regularization techniques in the architecture in order to at least partly mitigate intrinsic susceptibility of deep neural networks to overfitting. To this end, we placed dropout layers that randomly disabled a portion of neurons in every training cycle, therefore reducing the model reliance on particular pathways. Moreover, L1 and L2 regularizations were added to penalize large weight magnitudes, encouraging simpler, more generalizable models. These techniques together give the model better robustness and performance, which promise reliable classification results by alleviating the harmful effects of overfitting typical for deep learning tasks.

In Figure-5, we depict the various evaluations conducted throughout the training of our models. We utilized the Binary Cross-Entropy function as the loss function, coupled with accuracy, AUC, Recall and Precision as the

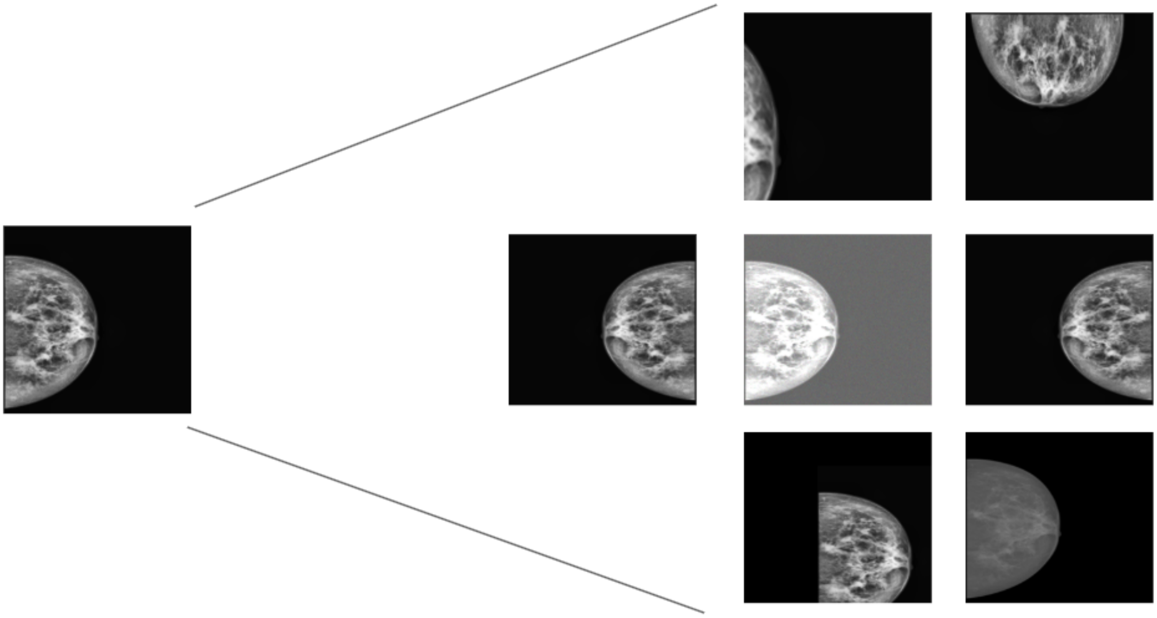

Figure 4. example of data augmentation

metrics for performance assessment. This graphical representation the evaluation of our classification model, with the techniques employed for model training.

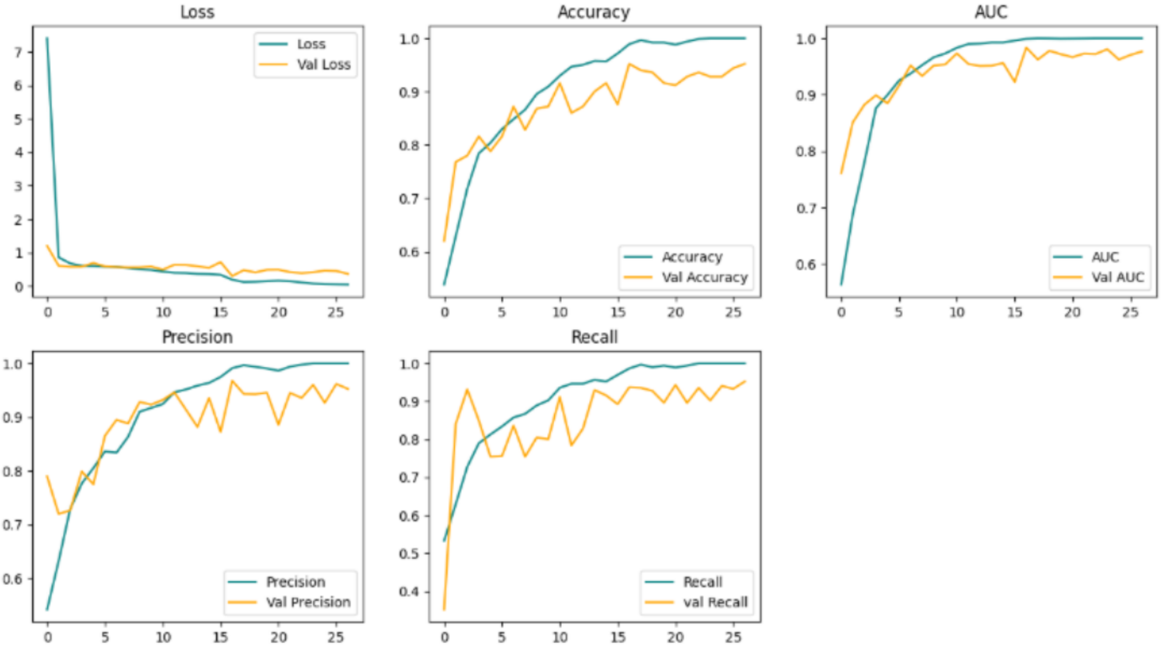

Figure 5. Assessing the status of the models during the training phase

The validity of these promising qualitative outcomes is substantiated by a comprehensive quantitative assessment conducted across the entire test set with different papers [Table-1]. This expansive evaluation encompasses a more representative sample size, ensuring a robust appraisal of models performance across diverse data instances. Such a holistic approach enhances the reliability and generalizability of the findings, providing a nuanced understanding of the model's efficacy beyond individual instances. Furthermore, the effectiveness and supremacy of the proposed method have been proved by a comparative analysis with the other research methodologies.

References	**Dataset**	**Accuracy**	**AUC**	**Recall**	**Precision**
[25]	DDSM	0.9887	0.988	0.9898	0.9879
[26]	DDSM	0.9798	0.9846	0.9763	0.9651
[27]	MIAS and DDSM	0.9793	-	-	-
[28]	INbreast	0.9564	0.9578	-	-
[29]	INbreast	0.9550	0.97	-	-
Our Model	MIAS	0.98	0.98	-	-
Our Model	DDSM	0.99	0.99	-	-
Our Model	INbreast	0.99	0.99	-	-
Our Model	MIAS, DDSM and INbreast	0.99	0.99	0.98	0.98

Table 2. Comparative analysis with other models

4.1. Ablation studies

We conducted ablation studies to assess the individual contributions of CLAHE, positional encodings, and attention heads. Removing CLAHE reduced accuracy by 3%, while switching to sinusoidal encodings dropped AUC by 2%. Reducing attention heads from 12 to 8 led to a 1.5% drop in F1-score.

Model	**Accuracy**	**AUC**	**F1-score**
CNN + Transformer	0.97	0.97	0.96
Our VIT model	0.99	0.99	0.98

Table 3. Comparative with CNN + Transformer

To improve the interpretability of the Vision Transformer model and to understand how it makes predictions, we generated the self-attention maps computed as part of the inference procedure. We used the attention weights and to emphasize the regions of the mammographies that the model considered the most important; we could visualize this process directly by producing heat maps. Figure-6 provides some examples of malignant cases in which the model identified important reflections on the masses and microcalcifications of the tissue that are standard image reflectors of breast cancer. These maps provide visual explanations that are aligned with the knowledge of the radiologist, which can help build trust and transparency in the predictions made by the model. This visual form of validation is an essential component of future practice in clinical settings and of working alongside medical professionals.

In the limitation to this study is that the available datasets used have inherent constraints, especially in their size and variety. The dataset in combination is smaller; hence, the ability for generalization of the developed model to newer or more diverse datasets may be limited when testing the model. It could also impact the real clinical performance of the models differently, as data in real different clinical settings vary substantially.

Another limitation, is the huge computational resources required for the training of vision transformers. They take up much memory and processing due to the high-resolution images, especially mammogram ones. This

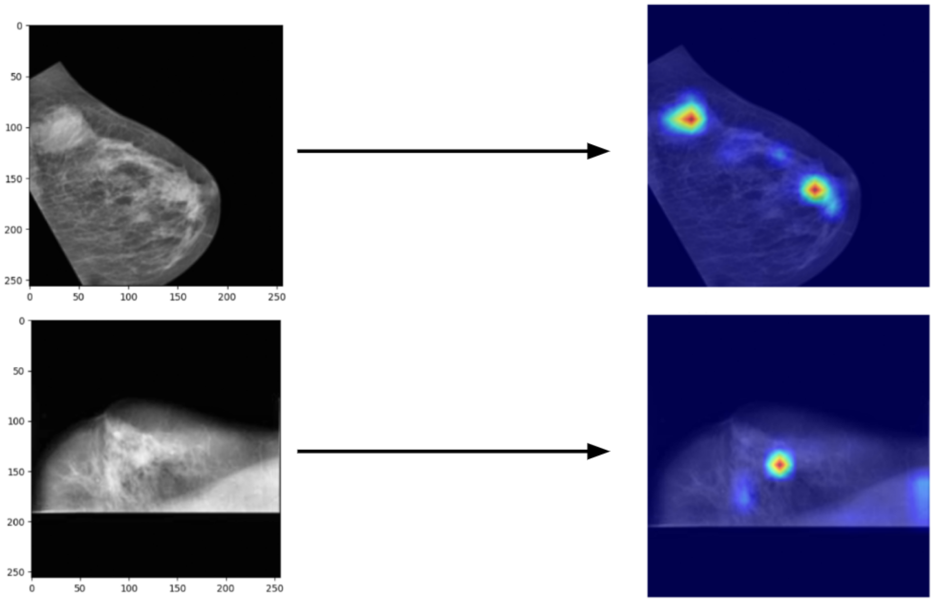

Figure 6. Examples of map attention

resource-intensive necessity in the process may sometimes lead to problems where resources are somewhat restricted.

Future work could overcome these challenges by using larger datasets of different types, possibly from several centers working in collaboration. Adding other types of imaging, like MRI or ultrasound, with mammography could make the dataset better and give a complete picture of breast cancer features. These improvements would most likely make the proposed model stronger and more useful for diagnosis, helping it to be used more widely in clinics.

5. CONCLUSION

In conclusion, this study has successfully demonstrated the utility of Vision Transformers (ViTs) in advancing the classification of mammographic images for breast cancer detection. By leveraging the patch-based representation in the multi-head attentional mechanism inherent in ViTs, the model was able to capture spatial and contextual relationships within the images, thus improving classification performance. The implementation of mature data preprocessing techniques such as CLAHE and several advanced regularization strategies enhanced the robustness and generalization capabilities of the model.

In future endeavors, researchers should strive to further enhance ViTs for medical imaging by exploring hybrid architectures that combine convolutional layers with transformer-based models to enhance the capturing of fine-grained details in mammographic images. Another productive avenue of work involves the integration of multi-modal imaging data, perhaps fusing mammography with MRI or ultrasound, giving rise to a more holistic diagnostic framework. Also, more efforts may go into the development of lightweight, very efficient, transformer architectures that are reflective of resource-constrained clinical scenarios.

Collaborations with healthcare professionals and validation of the proposed methodologies in real-world clinical settings remain paramount. In addressing these aspects, we attempt to bridge the gap between research advancement and translation toward practical applications with the ultimate goal of facilitating improved diagnostic, patient care, and personalized medicine in breast cancer management. These endeavors underscore our commitment to tackling the multifaceted challenges in breast cancer detection and treatment through innovative deep learning methodologies.

Acknowledgement

Special thanks to Professor Mohamed Chakraoui for his insights and discussions on several aspects of the material included in the article. Others whose feedback was beneficial in the early stages of the knowledge spectrum idea should also be thanked. Finally, we would like to note with appreciation the contributions of our professors-their reviews were really useful in improving this article.

REFERENCES

1. J. G. Elmore, C. K. Wells, C.-C. Lee, D. H. Howard, and A. R. Feinstein, “Variability in radiologists’ interpretations of mammograms,” *New England Journal of Medicine*, vol. 331, no. 22, pp. 1493–1499, 2005.
2. M. J. Yaffe and J. G. Mainprize, “Mammographic density: Measurement of a significant risk factor for breast cancer,” *Radiology*, vol. 258, no. 3, pp. 641–651, 2011.
3. K. Kerlikowske, D. L. Miglioretti, and C. M. Vachon, “Mammography screening: Evidence, guidelines, and clinical practice,” *The Lancet*, vol. 399, no. 10326, pp. 1856–1868, 2022.
4. M. Chakraoui, N. Mouhni, A. Elkalay, and M. Nemiche, “Deep negative effects of misleading information about covid-19 on populations through twitter,” *Ingénierie des Systèmes d’Information*, vol. 27, no. 2, p. 185, 2022.
5. E. Aniq, M. Chakraoui, and N. Mouhni, “Innovative: A novel deep learning-based semantic segmentation architecture for medical applications.,” *Ingénierie des Systèmes d’Information*, vol. 29, no. 4, 2024.
6. E. Aniq, M. Chakraoui, and N. Mouhni, “Artificial intelligence in pathological anatomy: digitization of the calculation of the proliferation index (ki-67) in breast carcinoma,” *Artificial Life and Robotics*, vol. 29, no. 1, pp. 177–186, 2024.
7. E. Aniq, M. Chakraoui, and N. Mouhni, “Ai-powered precision: breast carcinoma diagnosis through digital proliferation index (ki-67) assessment in pathological anatomy,” *Data Technologies and Applications*, 2024.
8. E. Aniq, M. Chakraoui, N. Mouhni, A. Aboulfalah, and H. Rais, “Breast cancer stage determination using deep learning,” in *World Conference on Information Systems and Technologies*, pp. 550–558, Springer, 2023.
9. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
10. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
12. S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: A survey,” *ACM Computing Surveys*, vol. 54, no. 10, pp. 1–41, 2022.
13. A. Oliver, X. Lladó, J. Freixenet, and J. Martí, “A review of automatic mass detection and segmentation in mammographic images,” *Medical Image Analysis*, vol. 14, no. 2, pp. 87–110, 2010.
14. L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, “Deep learning to improve breast cancer detection on screening mammography,” *Scientific Reports*, vol. 9, p. 12495, 2017.
15. T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. M. Mann, A. den Heeten, and N. Karssemeijer, “Large scale deep learning for computer aided detection of mammographic lesions,” *Medical Image Analysis*, vol. 35, pp. 303–312, 2017.
16. N. Tajbakhsh, J. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, “Convolutional neural networks for medical image analysis: Full training or fine tuning?,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.
17. Y. Wang, C. Song, Z. Xu, and L. Zhang, “Transformer-based neural networks for mammographic image analysis,” *Journal of Digital Imaging*, vol. 35, pp. 1011–1022, 2022.
18. H. Chen, Y. Zhang, Z. Zhou, and W. Li, “Breast cancer detection using vision transformers and ensemble learning,” *IEEE Access*, vol. 10, pp. 52856–52865, 2022.
19. J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, P. Taylor, and et al., “The mammographic image analysis society (mias) database,” 1994. Available online: http://peipa.essex.ac.uk/info/mias.html.
20. I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, J. Cardoso, and M. J. Cardoso, “Inbreast: Toward a full-field digital mammographic database,” *Academic Radiology*, vol. 19, no. 2, pp. 236–248, 2012.
21. M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. Kegelmeyer, “The digital database for screening mammography (ddsm): Cbis-ddsm subset,” 2020. Available online: https://www.cancerimagingarchive.net/.
22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.
23. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers and distillation through attention,” *arXiv preprint arXiv:2012.12877*, 2021.

24. 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,” *arXiv preprint arXiv:2103.14030*, 2021.
25. W. M. Salama and M. H. Aly, “Deep learning in mammography images segmentation and classification: Automated cnn approach,” *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4701–4709, 2021.
26. W. M. Salama, A. M. Elbagoury, and M. H. Aly, “Novel breast cancer classification framework based on deep learning,” *IET Image Processing*, vol. 14, no. 13, pp. 3254–3259, 2020.
27. M. Dong, X. Lu, Y. Ma, Y. Guo, Y. Ma, and K. Wang, “An efficient approach for automated mass segmentation and classification in mammograms,” *Journal of digital imaging*, vol. 28, pp. 613–625, 2015.
28. M. A. Al-Antari, M. A. Al-Masni, M.-T. Choi, S.-M. Han, and T.-S. Kim, “A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification,” *International journal of medical informatics*, vol. 117, pp. 44–54, 2018.
29. H. Chougrad, H. Zouaki, and O. Alheyane, “Deep convolutional neural networks for breast cancer screening,” *Computer methods and programs in biomedicine*, vol. 157, pp. 19–30, 2018.