International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Employing Vision Transformers for the Classification of Breast Cancer in Histopathological Images

Mohamed S.M. Geoda[1], Tawfik Ezat Mousa[2]
[1,2]Department of Computer Technologies, Higher Institute of Science and Technology, Tobruk, Libya
Email Address: [1]Mhomadgg@gmail.com,[2]Tawfikezat@yahoo.com

Abstract**—***Breast cancer is one of the most common cancers affecting women worldwide. This underscores the need for early diagnosis for effective treatment. Convolutional networks (CNNs) have historically predominated in medical vision issues. The inductive bias of CNNs toward spatial locality limits their ability to extract global features from breast cancer histopathology images, resulting in suboptimal classification performance. The Vision Transformer (ViT) is a promising deep learning model for computer vision applications, recently created to overcome this constraint. This work presents a strategy for classifying breast cancer using a pre-trained Vision Transformer, a robust architecture for recognizing long-term associations. We obtained the BcTMC dataset, which consists of 332 microscopic images classified as benign and malignant, from the Histology Department of the Tobruk Medical Center and used it for model training. The findings indicated that the Vit-Large Patch16-224 transformer model excelled, attaining an accuracy of 94.32%.*

Keywords**—** *Transformers, Attention, Fine-tuning,, Embedding, Classification.*

I. INTRODUCTION

Computerized medical image processing has fundamentally revolutionized healthcare by offering a comprehensive solution for automated disease diagnosis [1]. Integrating these techniques into medical imaging reduces costs and time and enhances physicians' confidence in diagnosing patients, especially regarding breast cancer [2]. Abnormal cell development is the etiology of breast cancer, a widespread and possibly lethal condition that impacts millions of women globally. Malignant tumors characterize the most severe kind of breast cancer. A mass in the breast or beneath the axilla is the most common indicator of breast cancer [3]. Breast cancer typically does not cause pain; therefore, a painless lump is more indicative of malignancy than a painful one. Distinguishing between benign and malignant tumors often requires expertise and advanced equipment beyond visual assessment [4]. Deep learning has recently become vital across various disciplines [5]. In autonomous medical image processing applications such as image categorization, convolutional neural networks (CNNs) have emerged as the predominant networks in recent years. Due to their limited receptive fields, these models encounter difficulties acquiring long-range information, constraining their applicability in vision tasks [6]. This study enhances the application of transformers in breast tissue image analysis and explores the potential of self-attention-based structures for image classification. We assess pre-trained ViT models of different sizes and configurations by comparing their performance when adapted for alternative tasks. Subsequently, we utilize the BcTMC dataset to implement these models for our specific objective. The results indicate that ViT models are highly promising for classifying breast tissue images.

II. RELATED WORKS

Touvron et al. [7] presented a data-efficient ViT built on the regularization and data augmentation methods previously used for CNNs. They also used a Transformer-based teacher-student strategy for the image classification test, which enhanced ViT's performance. Tummala et al. [8] used Swin transformers to categorize photos from the BreaKHis database into benign and malignant categories. Different magnifications of 40X, 100X, 200X, and 400X are used to collect the images. The best results were obtained using the large version of the Swin transformer model with a magnification factor of 40X. Wen et al. [9] proposed a transformer-based method for image classification of breast cancer. The results are displayed as a probability matrix representing either a positive (malignant) or negative (benign) sample after this algorithm automatically extracts features from photos of preset sizes. Since this model was developed using a tiny dataset with only a few samples of breast tissue images at 100x magnification, it lacks generality. Elagan et al. [10] experimented with two different ViT versions, Cait and BeiT. These algorithms detect breast cancer accurately by using transfer learning techniques. Their foundation lies in attention map visualization and incorporating a delayed classification to separate learning from the portrayal of the classification challenge. The interpretability of the model concerning the attention output and latent representations is the main emphasis of this study. Chenyang He et al. [11] provide a method for classifying breast cancer using a wavelet-based vision transformer network. Discrete wavelet transform (DWT) is applied to the network input to improve the neural network's receptive fields. With this method, significant features in the frequency domain can be captured. Breast cancer can be accurately and efficiently classified thanks to the model's capacity to capture the intricate features of breast tissue.

III. METHODOLOGY

The suggested method involves using multiple ViT models to classify class (benign-malignant) breast cancer images, where we modified the pre-trained weights using hyper-parameters to perform fine-tuning.

Mohamed S.M. Geoda and Tawfik Ezat Mousa, "Employing Vision Transformers for the Classification of Breast Cancer in Histopathological Images," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 4, pp. 277-280, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

A. *Database description*

To demonstrate the effectiveness of the proposed model, we used the BcTMC dataset from the Pathology Laboratory at Tobruk Medical Center, Libya. The dataset comprises 332 high-resolution microscopic images of breast tumor tissues from various patients, which we classified into benign and malignant categories using histological imaging of breast cancer. A Leica DM2500 LED light microscope was used to capture these images with different magnification settings (40X, 100X, 200X, and 400X) and a fixed image size of 2048 × 3072 pixels. Figures 1 and 2 below show images of the BcTMC dataset for both classes (benign and malignant).

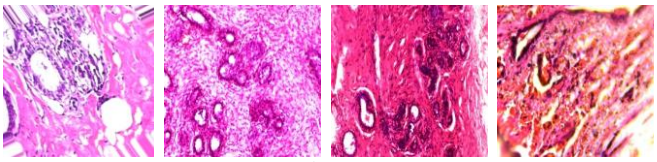

Fig. 1. Sample of benign breast tissue

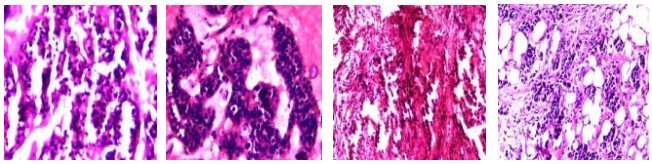

Fig. 2. Sample of malignant breast tissue

B. *Data augmentation*

We employed a data augmentation strategy on the BcTMC dataset to mitigate issues related to overfitting and model misalignment resulting from a limited number of images. This was accomplished using geometric alterations applied to the source images, including random rotation, random horizontal flipping, and random cropping. We acquired an expanded dataset of 2120 images.

C. *Model Architecture*

The Vision Transformer (ViT) is a categorization model that employs a multi-head self-attention mechanism on image patches. The model is based on the architecture of natural language processing transformers [12]. Self-attention is the most sophisticated method for processing sequences because it can express long-term semantic information [13]. A classifier and a feature extractor are the two primary components of ViT. The classifier classifies the input image, while the feature extractor aims to extract significant features from the image. A stack of transformer encoder layers makes up the feature extractor. The feature extractor comprises a position-wise feed-forward network and a multi-head self-attention mechanism [14]. You can interact with the various regions of the input image and find general relationships between them by using self-attention. The initial step in the ViT model's process is to split the input image into patches, each representing a local area of the image. After learning the spatial relationships between the patches, the model creates a list of embedding vectors by adding each patch's position embedding to its corresponding patch embedding. The model puts the embeddings into a stack of several transformer encoder layers to get useful information about the different image patches. Lastly, we run the output embedding vector through a linear classifier using a SoftMax activation function to predict the class label of the input image. Figure 3 shows an illustration of the proposed breast cancer classification system.

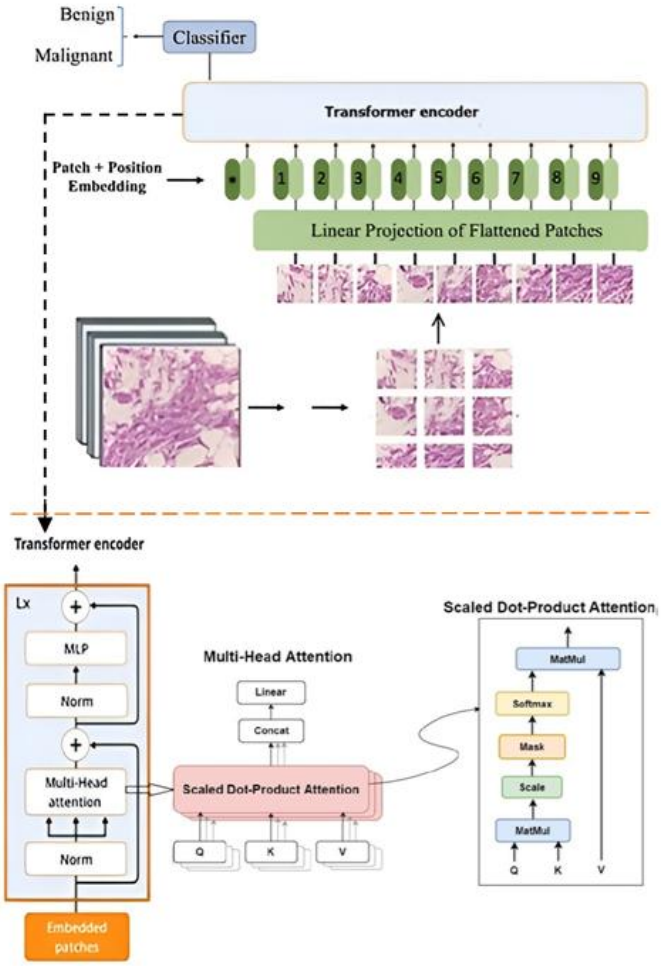

Fig. 3. Proposed model architecture

IV. Experimental Results

A. *Experimental setup*

In this work, the training dataset was split into batches of size 32, generating distinct mini-batch samples for each epoch. Using the categorical cross-entropy approach, we also determined the loss between the calculated and intended outputs at each iteration of 50 epochs. The model was trained using the Adam optimizer (adaptive moment estimation) with an initial learning rate 0.01. Tables 1 and 2 show the settings of the different hyperparameters and ViT parameter settings.

TABLE I. Hyper-Parameters Setting

Hyperparameter	Value
learning rate	0.01
optimizer	Adam
batch size	32
epochs	50
loss function	cross-entropy

Mohamed S.M. Geoda and Tawfik Ezat Mousa, "Employing Vision Transformers for the Classification of Breast Cancer in Histopathological Images," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 4, pp. 277-280, 2024.

TABLE III. ViT Parameter Settings

Setting	Value
Input size	224 X 224
Embedding size	768
Patch size	16
Number of heads	12
Head size	64
Output layer	1 neuron

B. Performance Evaluation

To assess the efficacy of the suggested models, many measures are employed, including precision, recall, accuracy, and F1 score. As shown in (1), precision is measured by dividing the total number of breast cancer images correctly identified by the total number of breast cancer images for testing. Precision, as defined in (2), represents the percentage of true positive cases that are accurately classified as positive. Recall, as shown in (3), indicates the true positive rate of predicted positive cases. The F1 score, calculated based on the harmonic mean of the precision and recall values, is defined in (4). These evaluation criteria are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\text{-}score = 2X \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

FP represents the number of test samples that belong to different classes but are incorrectly classified. FN, conversely, denotes the quantity of test samples belonging to the correct class that have been misclassified. TP and TN represent the number of test samples from different classes that the suggested models accurately predict.

C. Experimental results

The BcTMC dataset has been partitioned into training and testing in the experiments: 80% of the image samples have been used for the training stage and 20% for the tests. We investigated base, large, and huge versions of ViTs, which have several 12, 24, and 32 encoder layers, respectively. the best model is large ViT, which achieved an accuracy of 94.32%. Tables 1, 2, and 3 show the classification report for three models (base, large, and huge) of the BcTMC dataset.

TABLE IIIII. Classification report for the base ViT

Classes	Base ViT		
	Precision	Recall	F1-score
Benign	0.9144	0.9142	0.9008
Malignant	0.8977	0.8977	0.8857

TABLE IVV. Classification report for the large ViT

Classes	large ViT		
	Precision	Recall	F1-score
Benign	0.9428	1	0.9428
Malignant	1	0.944	0.9428

TABLE V. Classification report for the huge ViT

Classes	Huge ViT		
	Precision	Recall	F1-score
Benign	1	0.9250	0.9250
Malignant	0.9301	1	0.9238

V. Conclusion

In this study, we applied pre-trained ViT structures based on multi-head self-attention mechanisms to classify breast cancer in histopathology images. Since ViT models require large-scale datasets for training and the size of medical imaging is relatively small, we used a data augmentation strategy. This study demonstrates the potential of ViT models to classify global anatomical dependencies in histopathology images. In comparison to other models, the large ViT model achieved an accuracy of 94.32%.

Compliance With Ethical Standards

We used this dataset in accordance with ethical standards and laws. Tobruk Medical Center in Libya approved its use with strict assurance of patient confidentiality and privacy. We withheld all identifiable information about patients and concealed their identities.

References

[1] Rashed, Baidaa Mutasher, and Nirvana Popescu. "Critical analysis of the current medical image-based processing techniques for automatic disease evaluation: systematic literature review." *Sensors* 22.18 (2022): 7065.

[2] Sirisati, R. S., Kumar, C. S., Venuthurumilli, P., Ranjith, J., & Rao, K. S. "Cancer Sight: Illuminating the Hidden-Advancing Breast Cancer Detection with Machine Learning-Based Image Processing Techniques." 2023 *International Conference on Sustainable Communication Networks and Application (ICSCNA). IEEE*, 2023.

[3] Johnson, Karen S., Emily F. Conant, and Mary Scott Soo. "Molecular subtypes of breast cancer: a review for breast radiologists." *Journal of Breast Imaging* 3.1 (2021): 12-24.

[4] Bui, A. H., Smith, G. J., Dyrstad, S. W., Robinson, K. A., Herman, "An image-rich educational review of breast pain." *Journal of breast imaging* (2024): wbae001.

[5] Egger, J., Pepe, A., Gsaxner, C., & Li, J, "Deep learning—a first meta-survey of selected reviews across scientific disciplines, their commonalities, challenges and research impact." *PeerJ Computer Science* 7 (2021): e773.

[6] Azad, R., Heidari, M., Wu, Y., & Merhof, D, "Contextual attention network: Transformer meets u-net." *International Workshop on Machine Learning in Medical Imaging.* Cham: Springer Nature Switzerland, 2022.

[7] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H, "Training data-efficient image transformers & distillation through attention." *International conference on machine learning*. PMLR, 2021.

[8] Tummala, Sudhakar, Jungeun Kim, and Seifedine Kadry. "BreaST-Net: Multi-class classification of breast cancer from histopathological images using ensemble of swin transformers." *Mathematics* 10.21 (2022): 4109.

[9] Yu, Wen, and Yanqiu Li. "Breast cancer classification from histopathological images using transformers." *Thirteenth International Conference on Information Optics and Photonics* (CIOP 2022). Vol. 12478. SPIE, 2022.

[10] M. A. Elagan, "Breast histopathology image classification for end-toend diagnosis using transformers architecture," *in Proceedings of the 8th International Young Researchers' Conference* (IYRC 2023), pp. 1–9, SPIE, 2023.

[11] He, C., Diao, Y., Ma, X., Yu, S., He, X., Mao, G., ... & Zhao, Y, "A vision transformer network with wavelet-based features for breast ultrasound classification." *Image Analysis and Stereology* 43.2 (2024): 185-194.

[12] V. Ashish, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, p. I, 2017.
[13] C. C. Atabansi, J. Nie, H. Liu, Q. Song, L. Yan, and X. Zhou, "A survey of transformer applications for histopathological image analysis: New developments and future directions," *BioMedical Engineering OnLine*, vol. 22, no. 1, p. 96, 2023.
[14] A. M. Khan, A. Ashrafee, R. Sayera, S. Ivan, and S. Ahmed, "Rethinking cooking state recognition with vision transformers," in 2022 *25th International Conference on Computer and Information Technology* (ICCIT),pp. 170–175, IEEE, 2022.