ANNALS OF "DUNAREA DE JOS" UNIVERSITY OF GALATI
MATHEMATICS, PHYSICS, THEORETICAL MECHANICS
FASCICLE II, YEAR XVII (XLVIII) 2025, No. 2
DOI: https://doi.org/10.35219/ann-ugal-math-phys-mec.2025.2.05

A survey on the use of vision transformer and custom-built CNN for classifying ultrasound images of breast tissue

Mihai Grecu[1,2], Simona Moldovanu[3,4, *]

[1]*"Sf. Andrei" County Emergency Clinical Hospital, Galați 177 Brăilei Street, 800578*
[2]*Faculty of Medicine and Pharmacy, "Dunărea de Jos"University in Galați, 800216 Galați, Romania*
[3]*Departament of Computer Science and Information Technology, Faculty of Automation, Computers, Electrical Engineering and Electronics, "Dunarea de Jos" University of Galati, 47 Domneasca Str., 800008 Galati, Romania*
[4]*The Modelling & Simulation Laboratory, Dunarea de Jos University of Galati, 47 Domneasca Street, 800008 Galati, Romania*
**Corresponding author: simona.moldovanu@ugal.ro*

Abstract
Recently, the scientific community has focused on developing convolutional neural network (CNN) algorithms to enhance the ability of medical tools to diagnose breast cancer. This study aims to evaluate a powerful new deep learning technique based on Vision Transformers (ViT), which involves pre-processing images by dividing them into patches, and a custom-built CNN. The performance of both CNNs were tested using the following class combinations: healthy/benign, healthy/malignant, benign/ malignant, and healthy/benign/malignant. The images utilized in the classification process were sourced from the BUS-BRA dataset available on Kaggle. We observed that ViT demonstrated improved performance when the benign class was included in the classification. Although the obtained accuracy is modest, it is noteworthy that in this relatively unexplored field, the classification accuracy between benign and malignant classes was 78%, and 75% for the benign class and healthy patients when the ViT was used.

Keywords: ultrasound images, Vision Transformers, BUS-BRA Dataset

1. INTRODUCTION

The latest statistics indicate that in 2025, approximately 316,950 women in the United States will be diagnosed with invasive breast cancer (www.breastcancer.org). In Europe, data suggest that from 1989 to 2025, nearly 6.8 million cancer deaths will be prevented in the EU, including over 373,000 deaths from breast cancer [1]. An estimate for the year 2050, based on global population forecasts conducted by Fu et al. [2], projects that Asia will have around 1.4 million new cases of BC and 0.5 million deaths.

Recently, researchers have developed new techniques for the early detection and monitoring of breast cancer (BC). Apart from the clinical ways, the artificial intelligence (AI) algorithms help the classical methods in prediction or classification of BC cancer. These algorithms are trained on features extracted from images of images. AI algorithms trained on images are represented by convolutional neural networks (CNN) [3, 4], graph convolutional networks [5], hybrid learning machines, and CNN techniques [6].

A special category of deep learning algorithms is vision transformers (ViT) [7-11]. For BC studies, the ViT is used for classifying the digital breast tomosynthesis [7], breast cancer histopathology [8, 9], or ultrasound (US) images [10, 11]. These advanced AI techniques have shown promising results in improving diagnostic accuracy and aiding in the early detection of breast cancer.

Features were extracted from images and their classification was subsequently made up with ensemble and machine learning algorithms. Thus, the analyzed features were extracted from regions of interest [12, 13] or the whole image [14]. A complete analysis and justification of classification involved applying the explainable artificial intelligence (XAI) method, as proposed in a recent study by Moldovanu et al. [13].

This paper involves the classification of breast US images with vision transformers. The ViT entails patching of image classification tasks; it is a new method, unexplored for all types of images and diseases [15]. A vision transformer is a specialized transformer designed for computer vision tasks. It works by breaking down an input image into a series of patches. A single matrix multiplication serializes each patch into a vector, reducing it to a smaller dimension [15, 16]. The ViT was applied to breast US images acquired from healthy patients and those diagnosed with malignant and benign cancer. All classified images belong to the BUS-BRA database [17]. It is publicly available, and images are anonymized. The BUS-BRA database serves as a valuable resource for research in medical imaging, enabling the development of advanced algorithms in breast ultrasound interpretation. By leveraging the capabilities of the Vision Transformer, this research aims to improve diagnostic accuracy and facilitate early detection of breast cancer.

The main objective was to utilize patches from an image and a deep learning algorithm to extract new features and integrate them in a classification process, to achieve the highest classification accuracy for distinguishing between images containing benign, malignant breast cancer and healthy patients. To achieve the proposed objective, a ViT deep learning algorithm was proposed; its architecture was explored to streamline the calcification process. The proposed architecture leveraged attention mechanisms to focus on critical features within the image patches, enhancing the ability of the model to differentiate between benign and malignant cases. Additionally, extensive training on a diverse dataset ensured that the model could generalize well to unseen images, ultimately improving diagnostic accuracy in clinical settings.

The current paper is organized into four sections. Section 2 provides a comprehensive description of materials and methods. Section 3 is dedicated to the results and discussions, and prospects in this field of research. Section 4 presents the conclusions of the proposed work.

2. MATERIAL AND METHODS

2.1. Hardware and software

The architecture of personal computer used in experimental process was an Apple M1 Pro, 16 GB unified RAM, Chip Apple, and 20-Core GPU.

The Python software environment, version 3.12.8 was utilized, along with the following libraries and their respective versions: Keras (3.8.0), Numpy (2.0.2), seaborn (0.13.2), Sklearn (1.6.1), Matplotlib (3.10.0).

The Google Collaboratory was used as a hosted Jupyter Notebook service, and the images in Google Drive were stored, as runtime type a L4 GPU was used.

2.2. Image Dataset

The BUS-BRA Dataset is a public collection of anonymized breast ultrasound images from 1,064 patients; it is accessible at the link https://www.kaggle.com/datasets/orvile/bus-bra-a-breast-ultrasound-dataset, last accessed 2025 May. It includes biopsy-proven tumor cases and annotations from the BI-RADS (Breast Imaging Reporting and Data System) across categories 2, 3, 4, and 5. Three classes are included; the number of images is between brackets: healthy (133), benign (437), and malignant (210). The dataset was split into 75% for the training set and 35% for the testing set. The augmentation process consists of access to the horizontal random flip method, random rotation with an angle of 0.02, and random zoom with a height factor of 0.2 and a width factor of 0.2.

2.3. Vision transformers and hyperparameters

The use of Vision Transformers (ViT) has been reported for tumor detection and classification. In this study, we propose the augmentation of images and patches, flattening of each patch, and creating multiple layers of the transformer block as normalization and multi-head attention.

Also, a layer is occupied by a multilayer perceptron (MLP), which is a feedforward neural network (FFNN) used to fully connect neurons with nonlinear activation functions. The following layers of ViT consist of a layer normalization with a variance of 1, flatten, and dropout. The output of enumerated layers is processed by MLP, and for classification, the dense layer builds deep learning models. As an optimizer, AdamW is used as a stochastic gradient descent method that incorporates adaptive estimation of first-order and second-order moments, along with a technique for weight decay. Several hyperparameters influence the performance of ViT. One critical factor is the patch size, which determines the resolution of the image patches that are input into the transformer. Smaller patches can offer more detailed information, but they also increase computational complexity. Conversely, larger patches may simplify the computation but can result in the loss of fine-grained details. This study analyzes the effects of using patch sizes of 6×6 and image sizes of 72×72, with 144 patches per image. The performance of ViT is given by 108 elements per patch and 20 epochs.

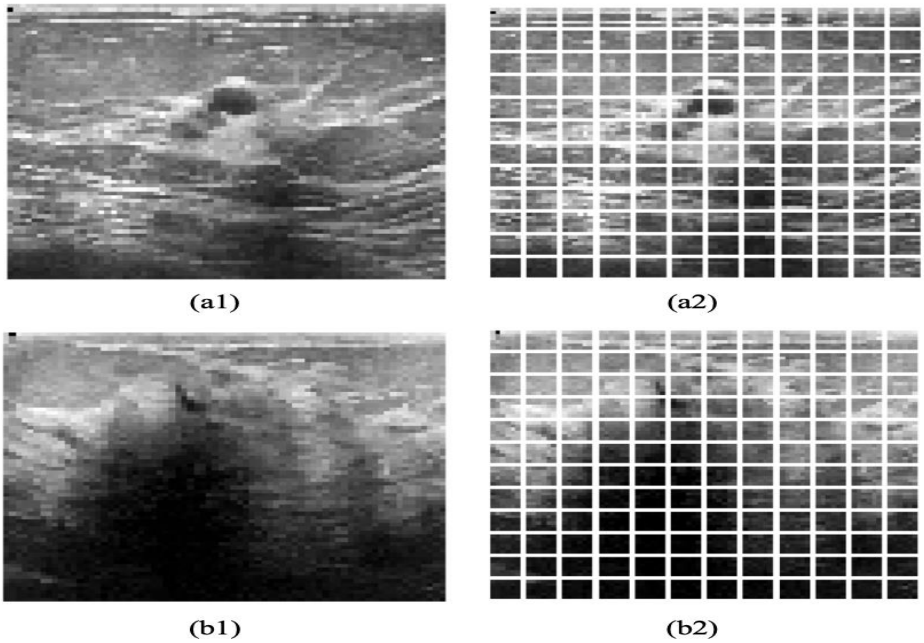

Fig. 1. Pre-processed US images (a1) Benign US image; (a2) Benign US image patches per image; (b1) Malignant US image; (b2) Malignant US image patches per image.

2.4 Custom-built CNN

In this study, a custom-built model was developed, which is an efficient model to demonstrate how a CNN-based method can classify breast cancer in ultrasound images. The experiment was conducted using a custom 5-layer CNN model. Figure 2 shows the custom CNN model, with the parameters on each layer.

Layer (type)	Output Shape	Param #
rescaling_2 (Rescaling)	(None, 80, 80, 3)	0
conv2d_6 (Conv2D)	(None, 80, 80, 128)	3,584
max_pooling2d_6 (MaxPooling2D)	(None, 40, 40, 128)	0
dropout_6 (Dropout)	(None, 40, 40, 128)	0
conv2d_7 (Conv2D)	(None, 40, 40, 64)	73,792
max_pooling2d_7 (MaxPooling2D)	(None, 20, 20, 64)	0
dropout_7 (Dropout)	(None, 20, 20, 64)	0
conv2d_8 (Conv2D)	(None, 20, 20, 32)	18,464
max_pooling2d_8 (MaxPooling2D)	(None, 10, 10, 32)	0
dropout_8 (Dropout)	(None, 10, 10, 32)	0
conv2d_9 (Conv2D)	(None, 10, 10, 64)	18,496
max_pooling2d_9 (MaxPooling2D)	(None, 5, 5, 64)	0
dropout_9 (Dropout)	(None, 5, 5, 64)	0
conv2d_10 (Conv2D)	(None, 5, 5, 128)	73,856
max_pooling2d_10 (MaxPooling2D)	(None, 2, 2, 128)	0
dropout_10 (Dropout)	(None, 2, 2, 128)	0
flatten_2 (Flatten)	(None, 512)	0
dense_4 (Dense)	(None, 128)	65,664
dense_5 (Dense)	(None, 128)	16,512

Total params: 811,106 (3.09 MB)
Trainable params: 270,368 (1.03 MB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 540,738 (2.06 MB)

Fig. 2. Proposed custom-built CNN model.

2.5. Performance evaluation

Label-based evaluation treats each label independently, effectively transforming a multi-label classifier into a binary classifier for each specific label. This approach results in four possible prediction outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The definitions of accuracy, precision, recall, and F1-score are as follows: ACC = (TP+TN)/(TP+TN+FP+FN), Precision = TP/(TP+FP), Recall = TP/(TP+FN), F1-score = (2*precision*recall)/(precision+recall).

3. RESULTS AND DISCUSSIONS

All models are trained having as inputs the classes Malignant/Benign, Healthy/Benign, Healthy/Malignant, and Healthy/Benign/Malignant. The effectiveness of the ViT models in solving the two-class and multiple-class problems with the BUS-BRA Dataset (see Section 2.2) was measured using standard performance metrics.

To validate the results, the accuracy, F1-score, recall, and precision are displayed and stored in Table 1 for ViT and Table 2 for custom-built CNN. The confusion matrices for each model are shown in Figures 3 and 4 for the same neural networks. In addition to assessing accuracy, we utilize the gradient class activation AdamW to better understand the reasoning process of our model in classifying breast US images.

Table 1. Values of precision, recall, F1-score for the ViT custom-built CNN and healthy, benign and malignant classes

Classes	Accuracy	precision	recall	F1-score
Malignant/ Benign	0.78	0.76	0.96	0.85
Healthy/ Benign	0.75	0.79	0.93	0.85
Healthy/ Malignant	0.62	0.61	0.93	0.74
Healthy/ Benign/ Malignant	0.61	0.65	0.77	0.71

Table 2. Values of precision, recall, F1-score for the custom-built CNN and healthy, benign and malignant classes

Classes	Accuracy	precision	recall	F1-score
Malignant/ Benign	0.62	0.66	0.89	0.76
Healthy/ Benign	0.68	0.78	0.82	0.80
Healthy/ Malignant	0.62	0.65	0.86	0.74
Healthy/ Benign/ Malignant	0.55	0.60	0.89	0.72

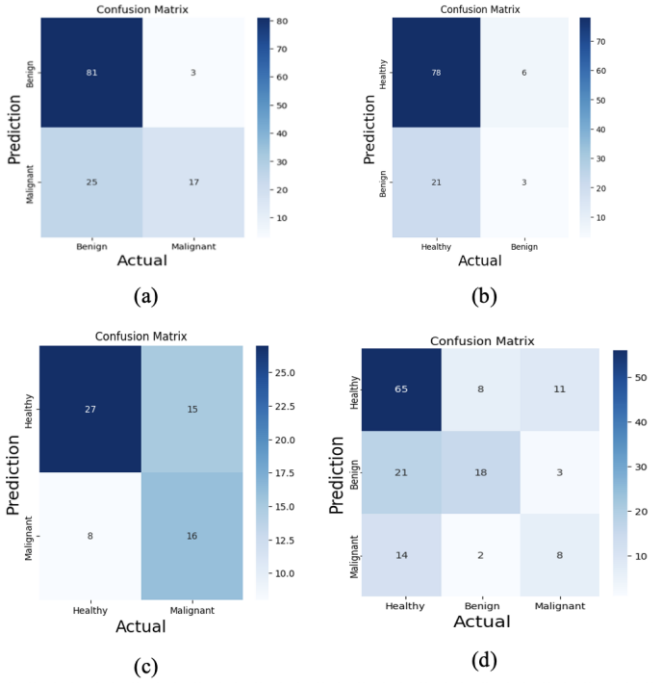

Fig. 3. Confusion matrices obtained for ViT; (a) Benign/Malignant; (b) Healthy/Benign; (c) Healthy/Malignant; (d) Healthy/Benign/Malignant

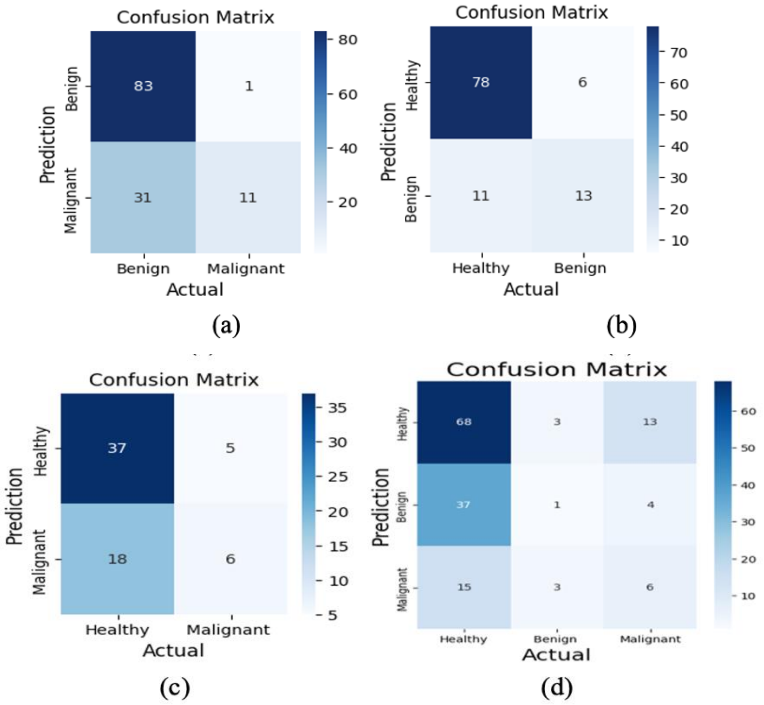

Fig. 4. Confusion matrices obtained for custom-built CNN; (a) Benign/Malignant; (b) Healthy/Benign; (c) Healthy/Malignant; (d) Healthy/Benign/Malignant

The ViT model mentioned was able to classify, in terms of accuracy, 78% of the breast US images correctly for classes Malignant/Benign, and the performance decreased to 75% for Healthy/Benign. The performance decreased very much for classes Healthy/Malignant and Healthy/Benign/Malignant. When the accuracy is higher, the benign class is implied in the classification process. For all classes, the recall metric examines the number of false negatives included in the prediction mix; it is negatively affected whenever false negative samples are identified.

A custom-built CNN with five convolutional layers yields low classification accuracy across all class combinations, indicating that the ViT is a suitable solution for this type of image and the studied pathologies. The results extracted with this CNN are incomparable to those obtained with the ViT. The superior performance of the ViT can be attributed to its ability to capture complex patterns and relationships within the image data more effectively than traditional convolutional networks. As a result, implementing the ViT enhances classification accuracy and provides deeper insights into the underlying features associated with the studied pathologies.

Throughout our research, we have encountered several intriguing findings that warrant further exploration and development. First, we note that the performance of ViT models trained on US images showed minimal improvement for disease classes benign and malignant. This result implies that the patches obtained from images contain important information used in the classification process.

In most of the previous work, the US images were classified with custom-built CNNs or pre-trained CNNs [3-6]. With reference to the US breast classification, the focus was on classifying healthy, benign, and malignant images; the ViT [7-11] was not applied to this image type.

4. CONCLUSIONS

First, while CNN shows strong generalisation, it also has significant data requirements. The necessity for extensive datasets poses challenges in the medical imaging field due to privacy concerns and the specialized nature of the images. However, increased network complexity leads to more parameters and greater computational demands, which significantly slows training speed and raises hardware requirements. To effectively implement these algorithms in clinical practice, future studies should concentrate on developing more efficient and lightweight networks.

One limitation is the small number of images from the BUS-BRA dataset used for training and testing, and the second is the large time consumption. This limitation highlights the need for more extensive datasets to improve the robustness of model training. Additionally, optimizing algorithms for faster processing without compromising accuracy will be crucial in enhancing the practicality of these solutions in real-world clinical settings.

References

1. Santucci C., Mignozzi S., Levi F., Malvezzi M., Boffetta P., Negri E., La Vecchia C., European cancer mortality predictions for the year 2025 with focus on breast cancer, Annals of Oncology 36(4) (2025) 460–468.
2. Fu M., Peng Z., Wu M. Lv M, D., Li Y., Lyu S., Current and future burden of breast cancer in Asia: A GLOBOCAN data analysis for 2022 and 2050, Breast 79 (2024) 103835.
3. Eshun R.B., Islam A.K., Bikdash M., A deep convolutional neural network for the classification of imbalanced breast cancer dataset, Healthcare Analytics 5 (2024) 100330.
4. Almaslukh B., A reliable breast cancer diagnosis approach using an optimized deep learning and conformal prediction, Biomedical Signal Processing and Control 98 (2024) 106743.
5. Bera, A. Bhattacharjee D., Krejcar O., Fluorescence microscopy and histopathology image based cancer classification using graph convolutional network with channel splitting, Biomedical Signal Processing and Control 103 (2025) 107400.
6. Qian L., Bai J., Huang Y., Zeebaree D.Q., Saffari A., Zebari D.A. Breast cancer diagnosis using evolving deep convolutional neural network based on hybrid extreme learning machine technique and improved chimp optimization algorithm, Biomedical Signal Processing and Control 87 (2024) 105492.
7. Kassis I., Lederman D., Ben-Arie G., Rosenthal M.G., Shelef I., Zigel Y. Detection of breast cancer in digital breast tomosynthesis with vision transformers, Scientific Reports 14 (1) (2024) 22149.
8. Sriwastawa A., Arul Jothi J.A., Vision transformer and its variants for image classification in digital breast cancer histopathology: A comparative study, Multimedia Tools and Applications 83 (2023) 39731–39753.

9. Hayat M., Ahmad N., Nasir A. and Ahmad Tariq Z., Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification, IEEE Access 12 (2024) 184119–184131.
10. Alruily M., Mahmoud A.A., Allahem H., Mostafa A.M., Shabana H., Ezz M., Enhancing breast cancer detection in ultrasound images: An innovative approach using progressive fine-tuning of vision transformer models, International Journal of Intelligent Systems 1 (2024) 2024.
11. He C, Diao Y., X. Ma, Yu S., He X., Mao G., Wei X., Zhang Y., Zao Y., A Vision Transformer Network with Wavelet-Based Features for Breast Ultrasound Classification, Image Analysis & Stereology 43(2) (2024) 185–194.
12. Tăbăcaru G., Moldovanu S. and Barbu M., Texture Analysis of Breast US Images Using Morphological Transforms, Hausdorff Dimension and Bagging Ensemble Method, 2024 32nd Mediterranean Conference on Control and Automation (MED), Chania - Crete, Greece, 2024, 263–267.
13. Moldovanu S., Munteanu D., Biswas K.C., Moraru L., Breast Lesion Detection Using Weakly Dependent Customized Features and Machine Learning Models with Explainable Artificial Intelligence, Journal of Imaging 11 (2025) 135.
14. Roy S.D., Das S., Kar D., Schwenker F., Sarkar R., Computer Aided Breast Cancer Detection Using Ensembling of Texture and Statistical Image Features. Sensors 21 (2021) 3628.
15. Khan S., Naseer M., Hayat M., Zamir S.W., Khan F.S., Shah M, Transformers in vision: A survey, ACM Computing Surveys 54(10) (2022) 200.
16. Bi J., Zhu Z. and Meng Q., Transformer in Computer Vision, Proceedings of 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Fuzhou, China, 24 – 26 September 2021, 178–188.
17. Gómez-Flores, W., Gregorio-Calas, M. J. & W. Coelho de Albuquerque Pereira. A. BUS-BRA: A Breast Ultrasound Dataset for Assessing Computer-aided Diagnosis Systems, Medical Physics 51(4) (2024) 3110–3123.