Undergraduate Conference on Intelligent Computing and Systems (UCICS 2025)
26-27 February, 2025; Varendra University, Rajshahi, Bangladesh

Classification of Cancer from Breast Ultrasound Images using Vision Transformer

S.M.Shahriar
Dept. of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chattogram, Bangladesh
sayeem26s@gmail.com

Muhammad Junayed
Dept. of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chattogram, Bangladesh
u2008023@student.cuet.ac.bd

Tanzeem Tahmeed Reza
Dept. of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chattogram, Bangladesh
tanzeemtahmeed01@gmail.com

Md, Ryhan Uddin
Dept. of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chattogram, Bangladesh
ryhan.cuet@gmail.com

Md. Sadikur Rahman Khan
Dept. of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chattogram, Bangladesh
sadik.cuet@gmail.com

***Abstract*—** Breast cancer occurs when breast cells mutate and form tumors. While ultrasound imaging is a valuable tool for detecting and evaluating breast masses, its effectiveness is limited by operator dependency and a restricted field of view. In recent years, deep learning methods have been employed to address these limitations. Among them, Vision Transformer (ViT) models have emerged as a powerful approach for image classification, offering accurate and efficient results. This study leverages ViT to enhance breast cancer detection, ensuring reliable outcomes with reduced time requirements. In this paper, we presented tailored Vision Transformer based encoder model, to identify and effectively classify the benign, malignant and normal classes of image from Breast Ultrasound Images (BUSI) dataset. The proposed approach was compared to actual vision transformer model and it was found that our customized model performs better in every way. In our approach, we addressed class imbalance, ensuring fair evaluation across all classes in BUSI dataset. The results are measured with the performance metrics: accuracy, precision, recall and f1_score accordingly. The accuracy of the algorithm was obtained 96.98%, with 97.96% precision, 95.46% recall and f1-score as 96.71% respectively. This proposed method would be much more time efficient as well as better than other classification algorithms for breast cancer prediction.

Keywords—breast cancer, Vision transformer, encoder model

I. INTRODUCTION

Breast cancer is a leading cause of cancer-related deaths among women globally, with 670,000 fatalities and 2.3 million diagnoses in 2022 [1]. Early detection is crucial to reduce mortality and expand treatment options. Ultrasound (US) imaging has become a preferred screening method due to its accessibility, cost-effectiveness, real-time imaging, and non-invasiveness. It is particularly effective in detecting lesions missed by mammography, especially in dense breasts. However, ultrasonography has limitations, such as difficulty identifying calcifications visible in mammography and some tumors [2]. Recent advances in automated breast cancer segmentation, classification, and detection using US imaging have shown immense potential [3]. Deep learning (DL), especially convolutional neural networks (CNNs), enhances image analysis but struggles with long-range information due to limited receptive fields [4].

Driven by the success of self-attention-based deep neural networks in natural language processing, Dosovitskiy et al. [5] introduced the Vision Transformer (ViT) architecture for image classification. These models divided the input image into patches, treated each embedded patch as a word in natural language processing, and used self-attention modules to figure out how these patches relate to one another [4]. Compared to CNNs, ViTs perform better on image classification tasks because they include more global information and stronger skip connections.

In this study, we propose a tailored Vision Transformer (ViT) model designed for efficient and accurate classification on breast ultrasound images dataset [3] while minimizing the risk of overfitting.

Our primary contributions include,

- Developed an automated ViT architecture capable of classifying and detecting breast cancer stages with higher accuracy than other ViT based works.
- To tackle the inherent class imbalance in breast ultrasound datasets, we incorporated the Balanced Sparse Categorical Accuracy metric ensuring robust performance across all classes.
- An extra multilayer perceptron (MLP) leveraged the power of transformer encoder to enhance detection accuracy while significantly reducing the time and effort required for diagnosis.

The rest of the paper is structured as follows, Section 2 describes the literature related works, Section 3 describes the methodology used along with the parameters, Section 4 describes result and analysis, and Section 5 concludes the proposed work.

II. Literature Review

Several methods for breast ultrasound (BUS) image segmentation have been explored, including traditional techniques like watershed transform [6], region growth [7], and active contour [8]. In recent years, deep learning (DL)-based approaches have advanced BUS image classification [9]. CNNs have excelled due to their ability to learn semantic hierarchies. Xing et al. [10] utilized a CNN-GAN model for tumor segmentation, while Kumar et al. [11] proposed multi-UNet for mass segmentation in BUS images. However, CNNs face challenges in retaining spatial and contextual information in deeper layers. Vision Transformer (ViT) [5] and hybrid models combining CNNs and Transformers have shown promise in addressing these limitations. A hybrid CNN-Transformer leveraging anatomical priors from BUS was applied in [12]. Despite progress, capturing global relationships and local patterns in BUS images remains challenging [3].

III. Methodology

This section is about the total experiment with adequate results and further analysis. Our research aims for the research question that has been aroused, that how do we classify breast cancer with transformer models and how accurate it would be. To classify benign, malignant or normal breast cancer from images, we not only proposed a Vision Transformer architecture with proper parameters but also experimented with other parameters as well which has ensured a groundbreaking performance. We preprocessed image data and introduced a ViT architecture that outperforms others. Several deep learning models were implemented for performance comparison with the proposed ViT.

A. Dataset

For this research, the benchmark ultrasound dataset BUSI [3] was used to perform the image classification by customized Vision Transformer (ViT). Breast ultrasound images were collected in 2018 from 600 women aged 25–75, totaling 780 PNG images (500×500 pixels). Images, categorized as benign, malignant, or normal, include ground truth annotations in fig. 1.

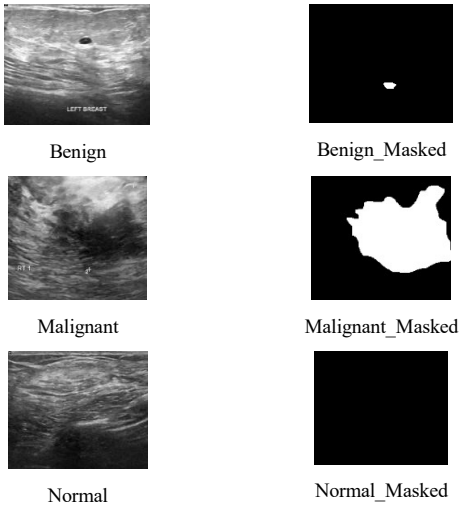

Fig. 1. Labeled Breast Ultrasound Image Dataset [3]

B. Preprocessing

The dataset, comprising three classes, was preprocessed with normalization, balancing via Balanced Sparse Categorical Accuracy, grayscale conversion, and label encoding. Training images were resized to 128×128×3 for the ViT model. The datasets employed in this suggested methodology have been categorized as training and validation components, with the appropriate percentages being 80% and 20% consecutively.

C. Proposed Approach

ViT, enhanced by transformers and attention mechanisms, has significantly improved image classification, excelling in image segmentation and prediction tasks. Our approach customizes the Vision Transformer (ViT) by utilizing only the encoder component, unlike typical encoder-decoder models. The proposed workflow is picturized in the fig. 2.

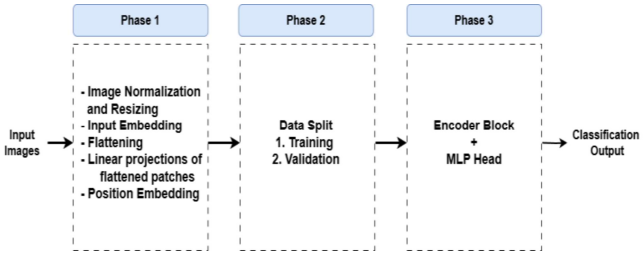

Fig. 2. Proposed Methodology in Flowchart

The performance of a Vision Transformer (ViT) model is heavily influenced by various hyperparameters, including patch size, embedding dimensions, number of transformer heads, transformer units, and Multi-Layer Perceptron (MLP) units.

In this study, two different configurations were tested: one with an embedding dimension of 64 and another with 32, transformer heads set at 3 and 2 transformer units configured as [128, 64] and [64, 32], and MLP units structured as [8192, 4096] and [2048,1024]. The embedding dimension determines the feature representation space, with higher values capturing more complex information but increasing computational demand. Transformer heads control the self-attention mechanism, where a higher number enables better context understanding but may introduce redundancy. The transformer unit sizes dictate the depth and width of the transformer layers, affecting learning capacity, while MLP units impact the model's ability to map learned features to classification outputs. A balanced combination of these hyperparameters is crucial to optimizing performance, ensuring that the model captures essential patterns without overfitting or losing important spatial details.

Fig. 3 illustrates the ViT processing steps: the input image is divided into 8×8 patches and passed through the embedded patches layer, followed by the first normalization (x1). It then passes through a multi-head attention layer, with skip connection 1 (x1, x2), and a second normalization (x3). The

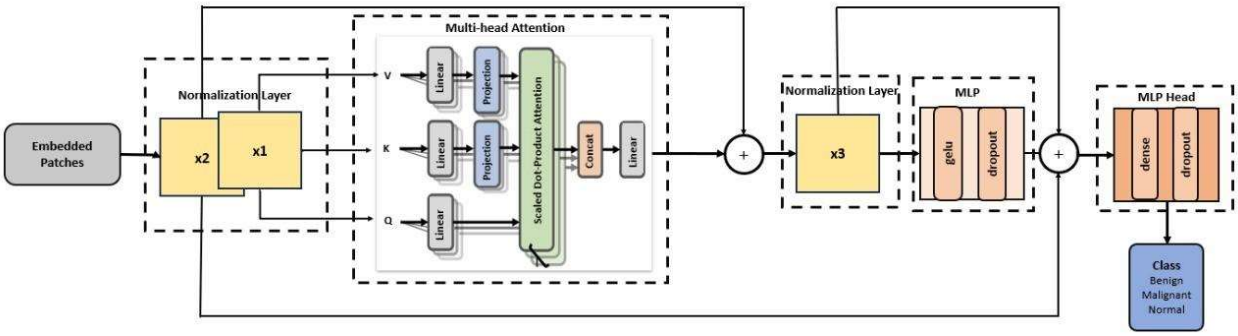

Fig. 3. Proposed ViT Architecture with added MLP head in the encoder model

result is processed through MLP layers with 'Gelu' and 'dropout', followed by skip connection 2 (x2, x3). Finally, the output undergoes the MLP head with 'dense' and 'dropout' layers, producing logits for classification.

The parameters which have been modified for the proposed transformer model is shown below in table 1 for better understanding.

TABLE I. TRAINING MODEL PARAMETERS

Parameters	***Conditions***
Patch Size	8 X 8
Activation Function	Gelu
Optimizer	'adam'
Learning rate	0.001
Maximum Iterations	20
Epochs	35

IV. RESULTS AND ANALYSIS

Our research's experimental setup includes a Nvidia GeForce RTX 3060 GPU, an Intel Core i7 processor, and 16 GB of RAM. We also used the Google Colab platform for running the programs simultaneously. To construct the proposed Vision Transformer model we used to run the program in python language with the help of the Keras library imported from TensorFlow. As a result, training time has potentially been decreased while performance improved.

A. Training and Validation Curve

The training split was 80% and the validation split was 20% as mentioned before. Therefore, while training the ViT model the training and validation accuracy along with the loss curve has been gained which is depicted in the fig. 4. It is evident from the training and validation accuracy curve that while the epochs were increasing there were several ups and downs but at last the accuracy became much higher than before after the 30th epoch.

a) Training & validation accuracy curve b) Training & validation loss curve

Fig. 4. Accuracy and Loss Curves for training & validation

The overall prediction while training was ongoing has been captured which shows a mesmerizing result in predicting the true image and the predicted image. The prediction process of training data is shown in fig. 5.

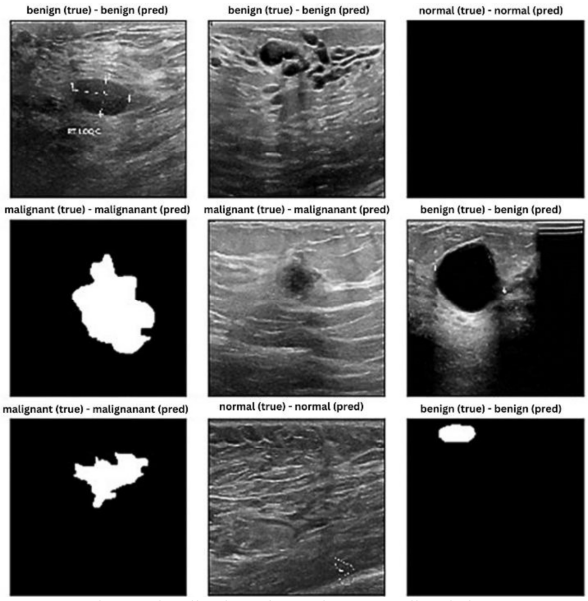

Fig. 5. Visualization of true class vs predicted class

B. Performance Evaluation

Three performance measures, F1 score, precision, and recall, were considered. Table 2 below summarizes the performance measurement matrices for each of the three classes with the performance evaluation metrices.

TABLE II. PERFORMANCE METRICS FOR EACH CLASSES

	Precision	***Recall***	***F1_Score***	***Overall Accuracy***
Benign	0.9560	0.9942	0.9747	96.98%
Malignant	0.9830	0.8923	0.9354	
Normal	0.9747	0.9827	0.9913	

C. Performance Comparision

We have performed three more ViT based transformer encoder model for the breast cancer classification: a base model with 16×16 input patch size (ViT16), a base model with 32×32 patch size (ViT32) and a ViT model implemented

TABLE III. PERFORMANCE METRICS COMPARISION WITH OTHER TRAINED VIT MODELS

Model	Optimizer	Accuracy	Precision	Recall	F1-Score	Patch size
ViT 8	Adam	96.98%	97.96%	95.46%	96.71%	8×8
ViT 16	Adam	92.25%	93.15%	92.80%	92.92%	16×16
ViT 32	Adam	76.17%	86.42%	59.67%	65.22%	32×32
ViT (from scratch)	Adam	71.48%	87.82%	60.13%	64.45%	4×4
ViT (with data augmentation)	Adam	72.82%	84.66%	59.86%	63.79%	16×16

from scratch. Also, with data augmentation of the breast cancer images, the ViT16 model has been programmed to see the results. The results are shown in the table 3.

While augmentation improves model generalizability, it poses challenges in breast cancer detection by altering tumor position, shape, or contrast. This inconsistency can reduce classification accuracy. Thus, while augmentation enhances data diversity, selective techniques are essential to preserve diagnostic integrity in sensitive medical applications.

D. Result comparison with recent works

In this part, a comparative analysis has been done to evaluate the accuracy and performance of this study. G. Ayana et. al [13] and B. Gheflati et. al [14] both have used Vision Transformer based model to classify the breast cancer and achieved 95% and 82% accordingly. A hybrid CNN with transformer model used by B. Shareef [15] achieved 82.8% of accuracy in cancer classification. Our work differs significantly from [13] and [14] in terms of ViT implementation and achieving higher accuracy of detection.

TABLE IV. RESULT COMPARISION WITH RECENT WORKS

Work	*Method*	*Accuracy*
G. Ayana [13]	ViT	95%
B. Gheflati [14]	ViT	82%
B. Shareef [15]	CNN+Transformer	82.8 %
Proposed	ViT	96.98%

V. CONCLUSION

This research proposes a Vision Transformer (ViT) model for breast cancer classification, demonstrating its effectiveness on the BUSI dataset for early diagnosis through a tailored ViT approach. The ViT model with 8 patches achieved significant improvements in classification accuracy of 96.98%. A smaller patch size, such as 8×8, captures finer image details but increases computational complexity, whereas larger patches, like 16×16 or 32×32, reduce the number of tokens while preserving global structural information. The superior performance of the 8×8 patch size model suggests that a finer resolution provides better feature representation for breast ultrasound images, likely due to the enhanced ability to detect subtle tumor features. Further research can be done by developing a hybrid transformer-based encoder decoder model which is more robust than only using vision transform based encoder models.

REFERENCES

[1] "Breast cancer," Who.int. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer. [Accessed: 02-Jun-2024].

[2] L. Wang, "Early diagnosis of breast cancer," Sensors (Basel), vol. 17, no. 7, p. 1572, 2017.

[3] W. Al-Dhabyani, M. Gomaa, H. Khaled, and F. Aly, "Deep learning approaches for data augmentation and classification of breast masses using ultrasound images," Int. J. Adv. Comput. Sci. Appl, vol. 10, no. 5,pp. 1–11, 2019.

[4] A. Hatamizadeh, D. Yang, H. Roth, and D. Xu, "Unetr: Transformers for 3d medical image segmentation," arXiv preprint arXiv:2103.10504,2021.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al.,"An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[6] W. Gómez, L. Leija, A. V. Alvarenga, A. F. C. Infantosi, and W. C. A. Pereira, "Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation," *Med. Phys.*, vol. 37, no. 1, pp. 82–95, 2010.

[7] J.-Z. Cheng *et al.*, "Computer-aided US diagnosis of breast lesions by using cell-based contour grouping," *Radiology*, vol. 255, no. 3, pp. 746–754, 2010.

[8] C.-Y. Lee, T.-F. Chang, Y.-H. Chou, and K.-C. Yang, "Fully automated lesion segmentation and visualization in automated whole breast ultrasound (ABUS) images," *Quant. Imaging Med. Surg.*, vol. 10, no. 3, pp. 568–584, 2020.

[9] Z. Zhuang, Z. Yang, A. N. J. Raj, C. Wei, P. Jin, and S. Zhuang, "Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion," *Comput. Methods Programs Biomed.*, vol. 208, no. 106221, p. 106221, 2021.

[10] J. Xing *et al.*, "Lesion segmentation in ultrasound using semi-pixel-wise cycle generative adversarial nets," *arXiv [cs.CV]*, 2019.

[11] V. Kumar *et al.*, "Automated and real-time segmentation of suspicious breast masses using convolutional neural network," *PLoS One*, vol. 13, no. 5, p. e0195816, 2018.

[12] Mo, Y., Han, C., Liu, Y., Liu, M., Shi, Z., Lin, J., Zhao, B., Huang, C., Qiu, B., Cui, Y., Wu, L., Pan, X., Xu, Z., Huang, X., Li, Z., Liu, Z., Wang, Y., & Liang, C., "HoVer-Trans: Anatomy-Aware HoVer-Transformer for ROI-Free Breast Cancer Diagnosis in Ultrasound Images," *IEEE Trans. Med. Imaging*, vol. 42, no. 6, pp. 1696–1706, Jun. 2023, doi: 10.1109/TMI.2023.3236011.

[13] G. Ayana and S.-W. Choe, "BUViTNet: Breast ultrasound detection via vision transformers," *Diagnostics (Basel)*, vol. 12, no. 11, p. 2654, 2022.

[14] B. Gheflati and H. Rivaz, "Vision transformers for classification of breast ultrasound images," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, vol. 2022.

[15] B. Shareef, M. Xian, A. Vakanski, and H. Wang, "Breast ultrasound tumor classification using a hybrid multitask CNN-transformer network," in *Lecture Notes in Computer Science*, Cham: Springer Nature Switzerland, 2023, pp. 344–353.