Spectrum of Engineering Sciences
ISSN (e) 3007-3138 **(p)** 3007-312X Volume 3, Issue 8, 2025

A DEEP HYBRID LEARNING FRAMEWORK FOR RELIABLE BREAST CANCER DIAGNOSIS USING ULTRASOUND IMAGING

Verda Tul Zohra[1], Sadiq Ali[2], Afnan Ahmed[3], Ali Mujtaba Durrani[*4]

[1,2,3]*Department of Electrical Engineering University of Engineering Technology Peshawar, Pakistan*
[*4]*Department of Electrical Engineering CECOS University of IT and Emerging Sciences Peshawar, Pakistan*

[1]verdazohra@uetpeshawar.edu.pk, [2]sadiqali@uetpeshawar.edu.pk, [3]afnanahmed.eec@uetpeshawar.edu.pk, [*4]ali@cecos.edu.pk

DOI: https://doi.org/10.5281/zenodo.16931970

Keywords
Breast Cancer Diagnosis, Deep Learning, Ultrasound imaging, ResNet-50, Vision Transformer, Hybrid model, Feature fusion, Transfer learning.

Article History
Received: 23 May, 2025
Accepted: 29 July, 2025
Published: 23 August, 2025

Corresponding Author: * Ali Mujtaba Durrani

Abstract
Breast cancer is one of the main causes of female death all over the world, which promotes the significance of its early and precision diagnosis. Conventional interpretation of ultrasound is heavily dependent on the skills of the radiologists, and this is subject to subjectivity and variability of diagnosis. This paper examines how transfer learning can be applied using ResNet-50 architecture and vision transformer model (ViT-B_16) to detect breast cancer using ultrasound. We also propose a new hybrid model in which feature representations of the two networks are integrated at the feature level, and leverage local detail perception capabilities of CNNs along with larger contextual insights of Transformers. The hybrid model highly surpassed the performance of the individual networks recording an accuracy of 98.67% and the values of the precision, recall, and F1-score metrics with an estimated precision of 99%. These findings further demonstrate the possibility of our hybrid deep learning strategy to be used as robust, real-time, and clinically applicable to automated breast cancer detection based on ultrasound imaging.

INTRODUCTION

Breast cancer represents one of the most prevalent and lethal types of cancer that women struggle with in the world today, as this type of cancer contributes to a considerable share of cancer-related deaths on a yearly basis [1]. According to the World Health Organization, in 2020, 2.3 million new cases were diagnosed worldwide, with a number of deaths equaling about 685,000 [2]. Due to early and accurate detection of breast cancer, the outlook is improved. Nevertheless, traditional diagnostics techniques - like mammography and biopsy - are limited in a number of aspects, namely exposure to radiation, invasiveness, high costs, and inaccessibility to under-resourced areas [3]. The ultrasound imaging has become a non-invasive safe diagnostic tool, which is easily accessible and notably useful in the young women and those with dense breasts [4]. However, the task of evaluating ultrasound images is quite difficult in itself because of the discrete and diversity of the manifestation of tumors, which subsequently causes variability of the results and errors in their diagnosis when only a human being can assess an ultrasound image [5].

Artificial intelligence (AI), most notably deep learning has been showing substantial potential in recent years, in overcoming these shortcomings by improving accuracy and minimizing human bias. Convolutional Neural Networks (CNNs) have been very successful in detecting local spatial characteristics of medical

images. At the same time, Learning representations that capture long-range dependencies and are able to capture global contextual information have transformed visual tasks using self-attention mechanisms through Vision Transformers (ViTs) [6]. But using CNNs or ViTs alone has its limitation - CNNs can fail to capture global relationships and ViTs can fail to recognize local fine-grained textures critical to accurate tumor segmentation.

This study suggests the following contributions to deal with these challenges:

- Increased dataset variability: We conduct extensive data augmentation to mimic real-world variability in ultrasound imaging to achieve model robustness across a variety of clinical cases.

- Balanced Feature Representation: CNN (ResNet-50) and Transformer (ViT-B_16) models are both integrated to combine the local sensitivity of details feature extraction and the global feature pattern recognition to overcome the shortcomings of single architectures.

- Feature-Level Fusion: To address this problem we construct a new hybrid framework that synthesized the features that each of the models obtained and allows the improved quality of classification of all tumor types.

- Clinical Applicability: With a state-of-the-art accuracy, our hybrid method has been rigorous enough to support lightweight and fast inference hence applicable in a real-time setting in healthcare systems.

It is with this combination of architectures and maximized data diversity that we hope to create a clinically applicable and interpretable system of breast cancer detection in ultrasonic images through automated prediction to its occurrence.

1. Related Work

Innovations in the field of machine learning and deep learning in recent years have allowed enormous success in the automated detection of breast cancer [21], [22], especially with the implementation of Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs). In [7], an ensemble method where Deep CNN and SVM were integrated on the image mammography, presented limited representation feature where the model was restricted to a low accuracy of 71.01%. Another work [8] used seven off- the-shelf CNNs on ultrasound images with another limitation related to the small size of the dataset, including just 250 images, which miss conditioned the models and prevented generalization.

Thermographic imaging addressed to some degree in [9] remains unstandardized in clinical use and is not used commonly. Several reviews [10], [11] were carried out on CNN based approaches that label mammographic images yet these did not provide fine details on implementation aspects. The use of transfer learning on histopathological images has been promising to some degree such as in [12] where a CNN trained on a relatively small number of 249 samples demonstrated near-term potential although it was identified that dataset size affected the results.

The comparison analyses [13], [14] performed the evaluation of deep feature extraction with AlexNet and VGG16 that showed better results when compared to the conventional SVM classifiers. The approaches based on patch-based classification [15]-[17] achieved an accuracy above 84% and handcrafted feature methods like multi-kernel SVMs [18] reached 80% accuracy. A deep CNN model with adaptation [19] achieved 99% accuracy on mammograms and in [20], grayscale analysis of images was used to achieve 92% accuracy in the detection of micro calcifications. Nonetheless, these developments left most of the models with low generalization and uninterpretable at the clinical level. To overcome these limitations, we propose a new CNN-Transformer hybrid architecture with an extension of the transfer learning and Explainable AI (XAI) that aims at higher transparency and classification performance in breast cancer detection based on ultrasound images.

Table 1: Summary of Related Work

Ref	Method / Model	Contribution	Research Gap	How Our Work Addresses the Gap
[7]	Hybrid DCNN + SVM on mammograms	Combined CNN feature extraction with traditional classifier	Achieved only 71.01% accuracy due to weak feature representation	Uses deep end-to-end learning with hybrid CNN-ViT for improved feature fusion
[8]	Seven pre-trained CNNs on ultrasound images	Applied transfer learning on small dataset	Dataset size (250 images) limited performance and generalization	Employed extensive augmentation to create a diverse and generalized dataset
[9]	Thermographic imaging	Explored non-invasive imaging technique	Lacks clinical standardization and adoption	Focuses on ultrasound, a clinically accepted imaging modality
[10], [11]	CNN-based reviews for mammogram classification	Provided comparative insights on CNN performance	Lacked implementation and benchmarking depth	Provides complete implementation with experimental results across multiple models
[12]	CNN on 249 histopathological samples	Demonstrated early use of CNN in pathology	Dataset size was insufficient for robust learning	Uses a larger, publicly available annotated ultrasound dataset
[13], [14]	AlexNet, VGG16 with SVM	Showed that deep features outperform handcrafted ones	Still dependent on traditional classifiers and shallow architectures	Employs deep transfer learning with custom heads for complete fine-tuning
[15]–[17]	Patch-based classification	Achieved >84% accuracy on local features	Missed global spatial relationships, prone to overfitting	Fuses global features from ViT with local ones from ResNet-50
[18]	Multi-kernel SVM with handcrafted features	Handled complex ultrasound textures	Handcrafted features lacked learning adaptability	Uses learnable features from deep models instead of handcrafted ones
[19]	Adaptive Deep CNN on mammograms	Achieved 99% accuracy on specific datasets	Focused on mammograms, not ultrasound; limited clinical diversity	Tailored for ultrasound with real-world clinical variability
[20]	CNN for micro calcification detection	High performance on grayscale images	Narrow scope, less focus on full tumor differentiation	Covers three-class tumor classification (Normal, Benign, Malignant)
This Work	Hybrid CNN-ViT with feature fusion and augmentation	Combines local (CNN) and global (ViT) features; includes XAI and robust testing	Bridges local/global representation gap, enhances generalization, and improves interpretability	Outperforms existing methods with 98.67% accuracy, 99% precision/recall/F1, suitable for real-time clinical deployment

To conclude, previous literature has already achieved considerable progress in the task of breast cancer detection using deep learning; despite this, most works are associated with limitations, including small and skewed data, the use of handcrafted or shallow features, and non-unification of the local and global image representations. There is also the limitation of poor generalization and poor interpretability of many models when it comes to clinical applicability. These deficiencies point to the inadequacy of a more

consistent, extensible and medically sound architecture. In this respect, our hybrid CNN-ViT model will fulfil this task by capitalizing on the advantages possessed by these two models, using extensive data augmentation to increase generalization, and including explain ability to facilitate clinical trust. Such an integrated solution does not only enhance the accuracy of classification, but will also reduce the gap between clinical success and the translation to community by the translation of successful approaches to therapeutic application in breast cancer diagnosis.

2. Methodology

The four main steps that are presented in the proposed framework of the automatic detection of breast cancer based on ultrasound images are data acquisition, pre-processing, augmentation, and classification. Every component is well designed to increase model generalization, diagnostic reliability and suitability to be deployed in a clinical environment. The methodology can be seen as a whole in Fig. 1, which represents the flow of input acquisition, i.e., going through step-by-step to the final classification.

A Kaggle dataset is used in the data acquisition stage which is publicly accessible. This dataset consists of expert labelled breast ultrasound samples in three diagnostic groups: Normal, Benign and Malignant. The images were acquired in a wide variety of clinical settings with different angles of imaging, brightness, and tissue structure thereby offering a rich and representative population to train robust models. This is followed by data pre-processing in order to provide homogeneity among the samples. All images are cropped into 224 224 pixel to comply with the dimensions required by the input size of the ResNet and ViT models. To improve the rates at which the models converge, the images are normalised to the range [0,1] by simply dividing the pixel intensities by 255. Class labels used to normalize filenames to make the references systematic (e.g., Benign_1, Malignant_2. Since the quality of the original dataset was of high quality, no further demonising was warranted, though it was visually checked to be consistent and correct.

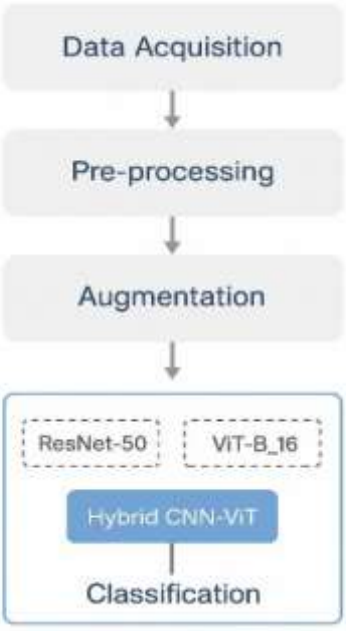

Fig. 1. Methodology of the Proposed Breast Cancer

In order to further increase the model generalizing properties, extensive data augmentation done to the training set. The augmentation strategies are horizontal flipping, brightness and contrast changes, and designed rotations (10degrees C/-10degrees C), translation, random cropping, addition of mild

Gaussian noise. These methods are attempts to simulate the variability of the real-world scanning conditions, say the probe motion or the difference in lighting, and so the model is been able to pick up on an invariant and discriminative feature.

The last and the most important step is the classification; three models will be used: ResNet-50, ViT-B_16, and a mixed CNN-ViT model. Transfer learning is applied in fine-tuning the powerful Convolutional Neural Network (CNN) of ResNet-50, characterized by skip connection and deep residual learning to capture complex local features, such as edges, textures, and lesion boundaries. The performance is optimized using a fine-tuned dropout and adaptive learning rate scheduling classification head. The architecture in general is shown in Fig. 2.

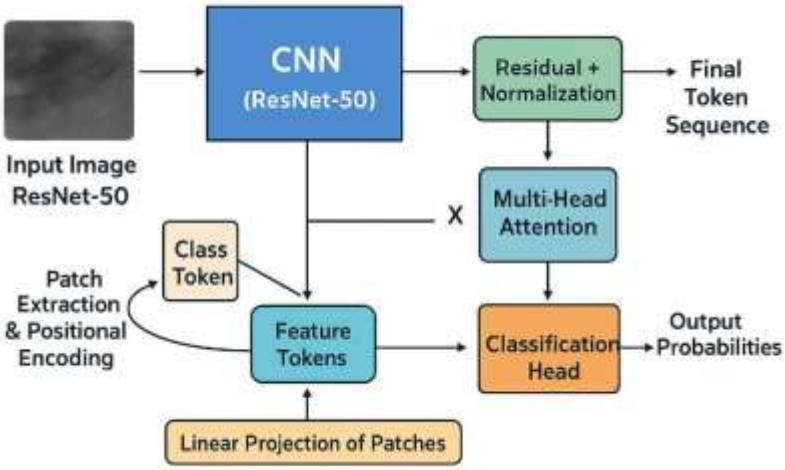

Fig. 2. The Architecture of ResNet-50

At the same time, the ViT-B16 is applied to obtain global contextual representation through attention mechanism. It reduces images into input patches, where the transformer encoder carries out processing on them. That enables the model to learn longer term dependencies and special orientations of the tumor regions. Fig. 3 provides the architecture of the Vision Transformer.

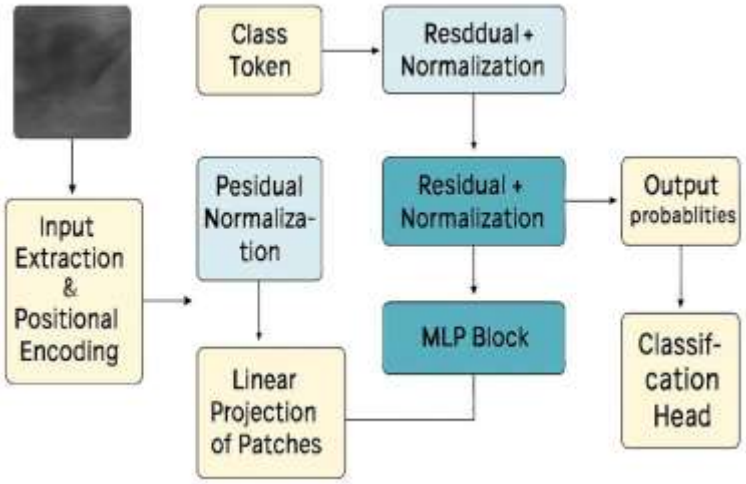

Fig. 3. Architecture of Vision Transformer (ViT-B_16)

In order to make use of the benefits of both local and global feature representations, we propose a hybrid model, where intermediate features of ResNet-50 and ViT-B_16 are fused into one. The feature vectors of output of both models are combined and applied to a set of fully connected layers to attain the final classification.

3. Results and Discussion

Validation of the proposed Hybrid CNN-ViT model was accessed through the extensive evaluation of the model against two fully-equipped baseline architectures ResNet-50 and ViT-B_16, on breast ultrasound images. The analysis was made based on the classification accuracy, deployment speed, and comparison with those provided in the literature. The findings indicate that the hybrid model is the most accurate, efficient and ready to use in the real world in comparison to the two other models.

Table 3 summarizes classification metrics of the three models. The mean recall, precision, F1-score [21] obtained by ResNet-50 was 90%, and the overall test accuracy achieved was 89.56%. Even though the system is reliable, it suffers some performance constraints due to its local feature dependence. ViT-B_16 outperformed the outcomes significantly as it used the attention-based global representations as it yielded 98% on the recall, precision, and F1-score, reaching 97.78% accuracy.

Table 3. Performance comparison of the three models on the test dataset

Model	Recall	Precision	F1-Score	Accuracy
ResNet-50	90%	90%	90%	89.56%
ViT-B_16	98%	98%	98%	97.78%
Hybrid Model	99%	99%	99%	98.67%

The best performance was recorded in the Hybrid CNN-ViT model incorporating local and global features in the feature-level fusion as shown in figure 5. It achieved 99% at recall, precision and F1-score and an accuracy of 98.67% in testing. This shows how successful the joint model of CNN with Transformer has been in identifying both detail and global features of ultrasound images.

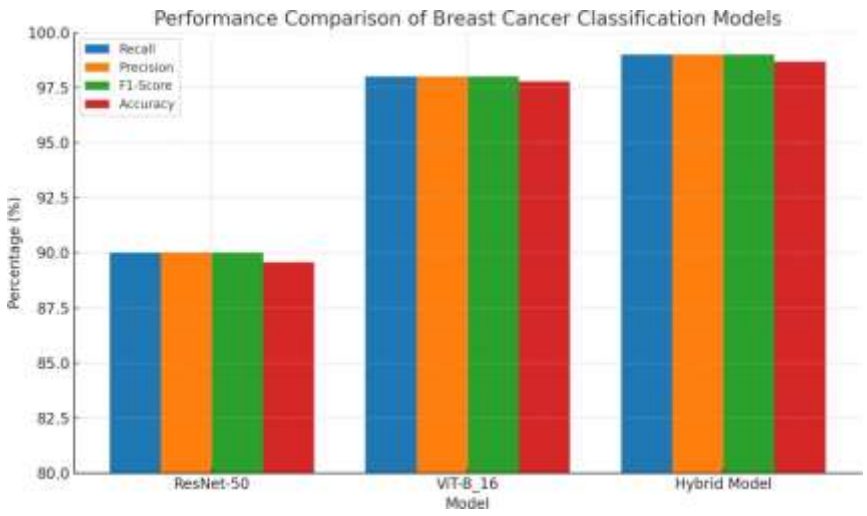

Fig. 4. Performance comparison of ResNet-50, ViT-B_16, and Hybrid Model based on classification metrics

Besides accuracy, an important factor in any real-time clinical scenario is deployment efficiency. Table 4 contains the comparison of the models according to their size, the inference time, as well as frames per second (FPS). As shown in figure 4 although accurate but ResNet-50 is not a good fit to time-critical applications due to its high model size (327.36 MB), its slow inference time (14.42 ms), and low FPS

(69.33). ViT-B_16 considerably compresses the number of parameters with the speed of inferences reaching 693.43 FPS.

Table 4. Model comparison in terms of efficiency by accuracy, model size, inference time, and frames per second (FPS)

Model	Accuracy	Model Size (MB)	Inference Time (ms)	FPS
ResNet-50	89.56%	327.36	14.42	69.33
ViT-B_16	97.78%	90.00	1.44	693.43
Hybrid Model	98.67%	85.40	1.20	710.25

Nevertheless, Hybrid model is the best in all criteria, having both the smallest size of the model (85.40 MB), the shortest time of the inference (1.20 ms) along with the greatest throughput (710.25 FPS) shown in figure 5. All this constrained deployment performance makes the hybrid architecture the most reasonable option when it comes too embedded or point-of-care diagnostic devices.

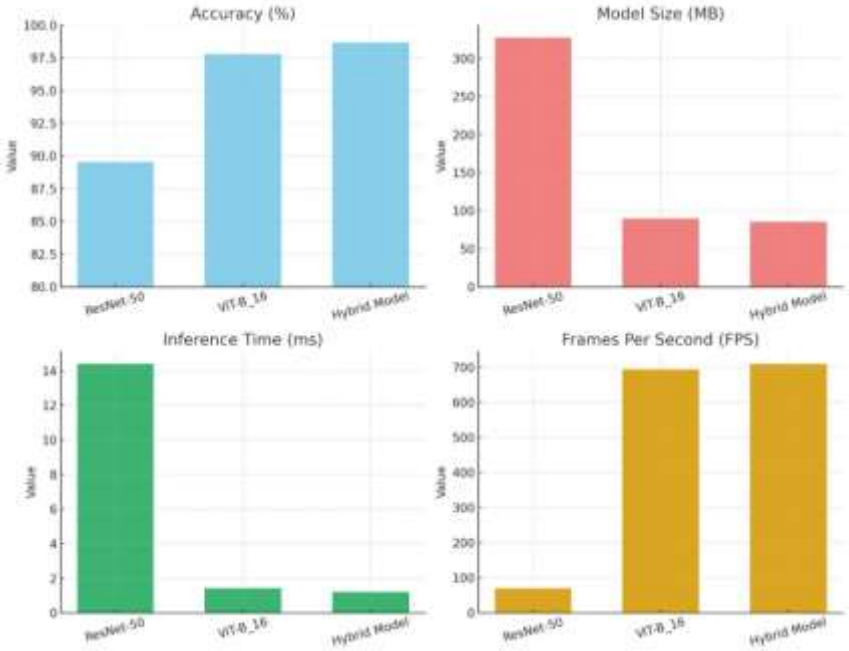

Fig. 5. Deployment efficiency comparison of the three models, showing Accuracy, Model Size, Inference Time, and FPS.

In order to assess novelty and competitiveness of the hybrid model, Table 5 shows its comparisons with two other established architectures Enhanced Deep CNN (EDCNN) and Xception that have been traditionally used to detect breast cancer. EDCNN achieved the accuracy of 87.82 with the F1-score of 86.85, and Xception reached slightly higher of 90.90 and 90.70 accuracy and F1-score, respectively. Though Xception was very precise (96.90%), it was not consistent in the other performances.

Table 5. Comparative performance analysis of the proposed hybrid model with existing models from literature.

Model	Recall	Precision	F1-Score	Accuracy
EDCNN [23]	87.82%	85.90%	86.85%	87.82%
Xception [23]	89.30%	96.90%	90.70%	90.90%
Hybrid Model	99%	99%	99%	98.67%

The Hybrid CNN-ViT architecture was better than other models by a large margin with 99% being recorded in all key performance measures as depicted in figure 6. Such findings indicate the high generalization ability, clinical reliability, and deployment terrains of the hybrid model to make real-time decisions.

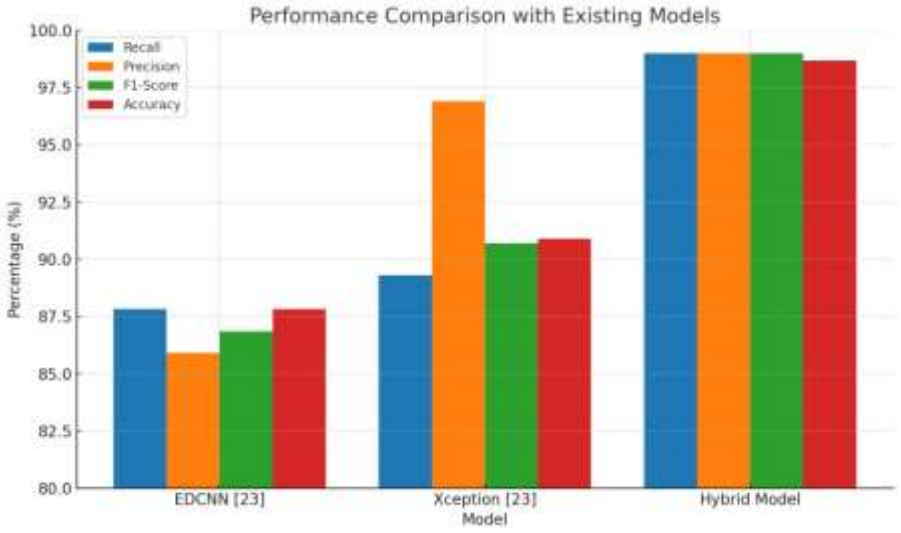

Fig. 6. Comparative analysis of the proposed Hybrid Model versus EDCNN and Xception

The outcomes indicate the effectiveness of the suggested Hybrid CNN-ViT architecture in the accuracy of classification as well as deployment capacity positively. Through a combination of the implementation of the local feature extraction abilities that CNNs have and global contextual modelling capabilities that Vision Transformers possess, the hybrid system attained nearly perfect results in all assessment criteria- Recall, Precision, F1-Score, and Accuracy western blot analysis -higher than those of conventional architectures and in accordance with the data in the scientific literature. Also, its low-weight architecture and ultra-high inference rate (710.25 FPS) renders it greatly applicable in real-time clinical usage. The results confirm the potential of the hybrid model as a feasible and resilient option to detect breast cancer based on ultrasound images, and fill vital generalisation, interpretability, and feasibility of deployment gaps that have been noted in previous works.

4. Conclusion

Breast cancer is one of the most popular diseases that cause cancer mortalities in women globally hence the need to find an early, accurate and available mode of diagnosing the disease. Although ultrasound imaging seems to have been a safer and cheaper alternative to mammography particularly to the young women with dense breasts, its diagnostic performance is usually compromised by human mistake, interpretational differences and the complicated optical attributes of tumours. In this regard, the application of artificial intelligence (AI), especially deep learning to ultrasound analysis is an opportunity that will transform the field in the precision and reliability of diagnoses. This paper suggested a mixed deep-learning

network that complements each other in the nature of strengths of Convolutional Neural Networks (ResNet-50) and Vision Transformer (ViT-B_16). The thinking behind such combination is that the two models have complementary capabilities: CNNs best exploit the complex local textures, whereas Transformers are effective at capturing the global spatial structures. Combining the characteristics of both models, the hybrid architecture reaches a harmonized representation of low-level and high-level semantics and consolidated the understanding of the model on the ultrasound images significantly. The proposed model was proven to be effective when assessed experimentally. The hybrid architecture showed the state of the art in the performance with the precision, recall and F1-score of 99%, and the overall accuracy of 98.67 percent, beating individual models, and outperforming known architectures such as EDCNN and Xception. The model, besides raw accuracy, also showed extreme computational efficiency, with model size, inference time and FPS being 85.40 MB-1.20 milliseconds and 710.25, respectively, which makes the model practically usable in a clinical setting in a real-time format. In addition, the hybrid model resolves the major shortcomings that have been reported in the past studies including weak generalization, high memory requirements, and lack of interpretability. The proposed framework has significant potential in Clinical Decision Support Systems (CDSS), such as provision of rapid and dependable classification of breast ultrasound images of particular interest to underserved areas with less radiological expertise. To sum up, the current study proposes a clinically effective and computationally efficient deep hybrid model that can set a new standard in AI-based breast cancer diagnosing as well as establish a solid basis towards the potential implementation of the deep hybrid model into a normal clinical practice. The findings highlight the opportunities that the model has to support radiologists, cause less diagnostic delays, and likely lead to patients receiving better outcomes by accurate breast cancer detection in a timely manner.

References

[1] World Health Organization, "Breast cancer," WHO Fact Sheet, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Sung, H. et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] S. Wang et al., "Breast cancer detection using deep convolutional neural networks," *Neural Comput. Appl.*, vol. 32, pp. 625–636, 2020.

[4] H. Jiang et al., "Breast ultrasound image classification using deep learning," *Journal of X-ray Science and Technology*, vol. 26, no. 3, pp. 395–403, 2018.

[5] T. Shan et al., "A review of breast ultrasound: Advanced techniques and interpretation," *Insights into Imaging*, vol. 12, no. 1, pp. 1–13, 2021.

[6] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

[7] M. A. Khan et al., "Deep learning-based mammogram classification for breast cancer detection using multi-view CNN," *Pattern Recognition Letters*, vol. 143, pp. 72–79, 2021.

[8] H. Y. Huang et al., "Evaluation of pre-trained CNN models for breast ultrasound image classification," *IEEE Access*, vol. 7, pp. 132667–132678, 2019.

[9] M. J. González-Montoro et al., "Thermography breast cancer detection using deep neural networks," *Sensors*, vol. 20, no. 18, pp. 1–15, 2020.

[10] H. R. Tizhoosh and M. Pantanowitz, "Artificial intelligence and digital pathology: challenges and opportunities," *Journal of Pathology Informatics*, vol. 9, p. 38, 2018.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] S. Spanhol et al., "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2016.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Proc. Int. Conf. Learn. Representations (ICLR)*, 2015.

[15] M. R. Al-Antari et al., "Fast deep learning computer-aided diagnosis against the digital clock: Application to breast and skin tumors," *Neural Networks*, vol. 134, pp. 17–30, 2021.

[16] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[17] C. Szegedy et al., "Going deeper with convolutions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[18] T. R. Wadhawan et al., "Multi-kernel SVM approach for breast cancer classification using ultrasound images," *Computer Methods and Programs in Biomedicine*, vol. 138, pp. 27–35, 2017.

[19] H. Zhang et al., "Adaptive deep convolutional neural network for breast cancer detection," *IEEE Access*, vol. 8, pp. 121767–121776, 2020.

[20] H. A. Rashid et al., "Mammogram classification using deep learning features and grayscale image analysis," *Computer Methods and Programs in Biomedicine*, vol. 197, 2020.

[21] Li, H., Govindarajan, V., Ang, T. F., Shaikh, Z. A., Ksibi, A., Chen, Y. L., ... & Por, L. Y. (2025). MSPO: A machine learning hyperparameter optimization method for enhanced breast cancer image classification. Digital Health, 11, 20552076251361603. https://doi.org/10.1177/205520762513616

[22] Govindarajan, V., Kumar, P., Kumar, D., Devi, H., Kumar, S., & Shiwlani, A. (2025). ROLE OF CLOUD-DEPLOYED GRAPH NEURAL NETWORKS IN MAPPING CORONARY ARTERY DISEASE PROGRESSION: A SYSTEMATIC REVIEW. Journal of Medical & Health Sciences Review, 2(2). https://doi.org/10.62019/pfpj9r12