International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

VIMONET: A MOBILENET AND TRANSFORMER BASED LIGHTWEIGHT HYBRID NEURAL NETWORK FOR EARLY BREAST TUMOUR DETECTION USING ULTRASOUND IMAGES

Archana Singh[1*], Surya Prakash Mishra[2]

[1 *]Assistant Professor, Department of Computer Science & I.T, SHUATS, Prayagraj, India
[2]Associate Professor, Department of Computer Science & I.T, SHUATS, Prayagraj, India

Abstract

Breast cancer is a predominant cause of mortality among women globally, thus highlighting the urgent necessity for early and precise detection. Ultrasound imaging is a commonly employed diagnostic modality. However, Ultrasound images encounter certain limitations including noise, shadowing, contrast issues, and diversity in tumour appearances. Medical image analysis has achieved remarkable outcomes with Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). Here, we offer an efficient lightweight hybrid model, VIMONET, for early breast tumour identification, by integrating MobileNetV3- Small with Vision Transformer (ViT). Our proposed model use transformers to capture long-range interactions, whereas CNNs are utilized for local feature extraction. A feature fusion module utilizing an attention method is incorporated to enhance representation learning. To substantiate the assertions, we have trained and evaluated our model on two datasets. Experimental results concerning seven parameters on the UDIAT, Spain Breast Ultrasound dataset (B), indicate that our hybrid model surpasses other state-of-the-art CNNs (ResNet50, MobileNet, EfficientNetB2 and the ViT model), achieving a Test Accuracy (98.5%), Precision (100%), Recall (97.00%), F1-Score (98.47%), Specificity (100%), AUC (99.38) and an average epoch time 9.89seconds/epoch. Whereas, on BUS, Baheya, Egypt dataset the Accuracy is (99.0%), Precision (98.00%), Recall (98.00%), F1-Score (98.00%), Specificity (98.00%), AUC (99.62) and an average epoch time 9.68seconds/epoch. In fact, our proposed model, VIMONET, fulfilling the criteria of lightweight design as it operates with very less GPU support requiring just 0.641 giga FLOPS and provides an effective means to enhance the accuracy of breast tumour classification while delivering quicker results compared to other existing heavy CNN models.

Keywords: Ultrasound, Breast Cancer (BC), Deep Learning (DL), Convolutional Neural Network (CNN), Vision Transformers (ViT), Lightweight.

1. Introduction

Cancer continues to pose a significant public health concern globally, with Breast Cancer (BC) being the most prevalent malignancy, accounting for around 12.4% of all new cancer cases annually [1,2]. In the United States, 2024 recorded around 3,13,520 new cases of invasive breast cancer and 3,10,720 new instances of non-invasive breast cancer in women, along with an additional 2,800 new cases of invasive breast cancer identified in males [3]. Breast cancer is among the most often diagnosed malignancies in American women, representing around 30% of all new cancer diagnoses [4]. Researchers have proved that timely identification is crucial for enhancing survival rates. Minimizing the morbidity linked to breast cancer, due to its heterogeneous nature is also a major concern which includes numerous entities with unique biochemical, histological, and clinical characteristics [5,6].

High-risk breast cancer is classified into two primary categories: benign and malignant. Benign breast cancer denotes non-malignant tumours inside breast tissue [7]. These tumours do not infiltrate adjacent tissues or metastasize to other regions of the body; although they may induce discomfort or anxiety, they are generally not fatal. Prevalent benign breast disorders encompass fibroadenomas and cysts [8]. Conversely, malignant breast cancer comprises cancerous tumours that can infiltrate adjacent breast tissue and spread over to other regions of the body. Malignant tumours are aggressive

Received: August 02 , 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

and need immediate intervention to avert dissemination and diminish mortality risk. This type of breast cancer includes ductal carcinoma in situ (DCIS), invasive ductal carcinoma (IDC), and invasive lobular carcinoma (ILC). [9, 10].

Conventional diagnostic methods, such as mammography, ultrasound imaging, and magnetic resonance imaging (MRI) [11], constitute the primary line of clinical screening [12]. Sometimes, non-invasive methods may not be reliable enough to identify malignant tumour, and hence, requiring a more conclusive diagnosis via biopsy. The biopsy procedure involves the acquisition of tissue samples, their preparation on glass slides, subsequent staining, and microscopic examination by pathologists to detect malignant cells [13]. This procedure, however precise, is labour -intensive and significantly depend on pathologists' knowledge, underscoring a notable deficiency in the existing diagnostic framework. This gap highlights the pressing necessity for automated, precise, and efficient diagnostic solutions that may improve user experience and offer greater insights into the diagnostic process.
Currently, machine learning has enhanced healthcare by refining illness diagnosis and patient monitoring [14]. Conventional techniques depend on manual feature extraction, which is labour -intensive and necessitates subject expertise, whereas deep learning automates this process, enabling direct learning from unprocessed data [15]. Convolutional Neural Networks (CNNs) are widely used deep learning methodologies for image-based cancer diagnosis, owing to their capacity to discern complicated patterns inside complex datasets [16]. Deep learning methodologies encounter obstacles such as the necessity for extensive label led datasets, significant processing requirements, and restricted accessibility in resource-limited environments [17]. To resolve these challenges, transfer learning [18,19]is used in pre-trained models such as Densenet121, MobileNet, ResNet50, and VGG19, therefore diminishing data needs and enhancing performance [20-22]. Although they are predictively accurate, conventional deep learning models frequently function as "black boxes," offering no transparency on the decision-making process and failing to indicate which regions of the image are most pivotal in arriving at a diagnosis [23]. The absence of interpretability is a barrier in clinical environments, since practitioners require transparency and clarity in diagnostic instruments to trust and effectively incorporate them into their practice. Moreover, some models concentrate exclusively on feature extraction, omitting interpretive layers, hence limiting their applicability in real-world scenarios where understanding the rationale behind predictions is essential.

Driven by the necessity for early and accurate breast cancer diagnosis, our research presents the Deep Neural Breast Cancer diagnosis system (VIMONET). This system utilizes advancements in deep learning, particularly Vision Transformers and Convolutional Neural Networks (CNNs) with Transfer Learning, to automate the diagnosis procedure. We created an improved model to precisely categorize benign and malignant breast tissue cases from the available Breast Ultrasound image Dataset (BUS), Egypt [24] and the UDIAT dataset(B), Spain **[**25**]**, which provides breast ultrasound images for input and training the model.

Despite many CNN-based methods attaining classification accuracy near 90%, they possess limits since CNNs address long-range dependencies by increasing the convolution kernel size, which therefore reduces system speed and improves feature representation. In practice, it is highly computationally intensive and restricts the generalization capacity of the minor resource system. Deep learning models, especially convolutional neural networks (CNNs), have demonstrated potential in automating and enhancing diagnostic precision. Vision Transformers can concurrently analyse large image patches, enhancing the identification of subtle malignant patterns. They can acquire hierarchical and intricate characteristics, rendering them resilient for histology, radiography, and many imaging modalities [26]. In the context of extensive datasets, they can surpass CNNs in both feature extraction and classification. Attention maps can identify areas that contribute to cancer diagnosis,

Received: August 02 , 2025

Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

facilitating explainable AI in healthcare. Nonetheless, substantial datasets are required to mitigate overfitting, presenting a barrier in medical imaging due to the scarcity of labelled data. Training Vision Transformers need advanced GPUs because of self-attention processes, and they lack standardized architectures as they continue to evolve. Recent improvements in Transformer designs, first designed for natural language processing, have broadened their relevance to vision problems, providing benefits in capturing long-range dependencies and contextual interactions.

Our contributions are diverse:
* In this paper, we introduce the VIMONET model, which integrates MobileNetV3-Small with ViT Vision Transformer and employs attention-based transfer learning for breast cancer detection, emphasizing the precise classification of benign and malignant tumour classification.
* Secondly, we employ an Inception CNN block to extract rich features of images utilized in diagnostic decision-making.
* Thirdly, we tackle class imbalance in the datasets by using data augmentation and augmenting all the classes with same number of images, hence enhancing detection performance for benign and malignant classes.
* Fourthly, we do a statistical analysis of our model and demonstrate its importance in comparison to other state-of-the-art models.

We also performed comprehensive assessment of our model's performance, juxtaposing it with other established approaches and cutting-edge techniques to illustrate its superiority, achieving an accuracy of 99.0% for the BUS dataset and 98.50% for the UDIAT dataset.

Further, this paper is organized in following sections: Section 2: "Literature Review" section examines the current literature on breast cancer detection, concentrating on research utilizing the BUS, UDIAT and other datasets. The next Section 3:" Proposed Methodology of hybrid VIMONET" delineates the approach utilized in the development of the hybrid system, encompassing dataset features, preprocessing techniques, and the architecture of the VIMONET model. The fourth section titled "Performance Analysis" states the experimental findings and assesses the efficacy of the VIMONET system. Going further to Section 5: "Conclusion and Future Work" aligns the constraints and prospective avenues of the offered research. The "Conclusion" section offers an analysis of the data, assessing the system's potential influence on clinical decision-making and patient outcomes along with future perspectives.

2. Literature Review

Breast Ultrasound Images Dataset (BUS)

Using ultrasound images, Munteanu et al.[27] presented an end-to-end DL model for identifying breast cancer. Their model combines CNN for classification, UNet for segmentation, and GAN-based data augmentation. Addressing data restrictions, adding synthetic images to the training set, and attaining 86% accuracy are the main contributions of their work. Their use of a single publicly available dataset, however, restricts the model's generalizability and comes with a significant computational cost.

Using ultrasound images, Pacal et al. [28] suggested a DL method for breast cancer classification. AlexNet, VGG16, Resnet, GoogleNet, EfficientNet, and Vision Transformer were among the models they assessed. With an accuracy of 88.6%, the Vision Transformer outperformed the other CNN models. Nevertheless, the study pointed out that the performance of deeper models was limited by the BUS dataset's modest size. They suggested that larger datasets and more sophisticated data augmentation methods may be advantageous for future research.

Using ultrasound images and convolutional neural networks (CNN) and image fusion techniques, Alotaibi et al. [29] presented a method for classifying breast cancer. They used a 3 - Step

Received: August 02 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

preprocessing strategy that includes RGB fusion, Region Of Interest (ROI) highlighting, and speckle noise filtering. The accuracy of the VGG19 model was 87.8% on the BUS dataset and 85.2% on the KAIMRC dataset. Their work was limited, by the BUS dataset's relatively small size, which affects the model's capacity for generalization, its high computational cost, and the imbalance between the benign, malignant, and normal classifications, which makes consistent performance difficult.

Isik et al. [30] used Model-Agnostic Meta-Learning (MAML) and Prototypical Networks (ProtoNet) to develop a meta-learning-based model for few-shot categorization of the BUS dataset. In a 10-shot scenario, the maximum accuracy of 88.9% was attained with ProtoNet with Resnet50 as its backbone, far exceeding the baseline accuracy of 83.1% with transfer learning. Their model's drawbacks was, its reliance on dataset similarity for cross-domain training and its high processing requirement while using deeper backbones like Resnet50.

Gheflati et al.[31] used the BUS dataset, which consists of 780 images, to study the use of Vision Transformer (ViT) models for breast ultrasound (US) image classification. The ViT B/32 model outperformed conventional CNNs like as Resnet50, which had an accuracy of 85.3%, with the greatest accuracy of 86.7% and an AUC of 0.95. However, due to the unique properties of ultrasound images, traditional data augmentation approaches like cropping and rotation had little effect on enhancing accuracy, and the BUS dataset's small size hampered the model's generalizability.

SCA-InceptionUNeXt, a lightweight U-shaped network for medical image segmentation, was proposed by Tagnamas et al.[32]. It included a Spatial-aware Channel Attention (SCA) module for improved feature fusion and a redesigned InceptionNeXt block for effective feature extraction. Using 26.11M fewer parameters than U-Net, the model achieved 81.66% Dice on BUS, outperforming SOTA approaches on four datasets. It had trouble performing well in several imaging modalities, though, and its model choices were not interpretable.

To diagnose breast lesions in ultrasound images, Sirjani et al. [33] created a DL model based on an improved InceptionV3 architecture. Optimizing hyperparameters and converting InceptionV3 modules to residual Inception modules were two significant enhancements done in this proposal. The model was evaluated against 24 CNN architectures after being trained on five datasets, including 2 from imaging facilities and 3 from the general population. Its results included an accuracy of 0.81, precision of 0.83, recall of 0.77, F1 score of 0.80, AUC of 0.81, and RMSE of 0.18. Although useful for classification, the study points out that more clinical validation is required and that the model's applicability is limited.

The Hierarchical Attention-guided U-Net (HAU-Net), a hybrid CNN-transformer architecture for breast lesion segmentation in ultrasound images, was first presented by Zhang et al. [34]. The model integrates a cross-attention block (CAB) in the decoder and an L-G transformer block into U-Net skip connections, combining CNNs for local detail and transformers for long-range interdependence. HAU-Net outperformed state-of-the-art segmentation techniques with Dice coefficients of 83.11% on the BUS dataset30, 88.73% on UDIAT, and 89.48% on BLUI.

UDIAT Breast Ultrasound Dataset

Singh, A. et al. [35] introduced VISNET, a hybrid lightweight architecture that merges EfficientNetB0 for local feature extraction with Vision Transformer (ViT), a component for long range and global context, and also augmented an attention-based fusion module. The model was assessed using the UDIAT and Baheya datasets, comparing it to baseline models such as VGG16, ResNet50, EfficientNetB0, and ViT. On UDIAT, VISNET attained an accuracy of 96.90%, precision of 95.83%, recall of 97.73%, F1-score of 96.67%, specificity of 97.72%, and an AUC of 98.67%, surpassing the baseline metrics. The authors determined that lightweight hybrid designs can achieve optimal performance while minimizing computational expenses. However, they recognized that the limited dataset size of UDIAT and the absence of cross-center validation constrains the model's generalizability.

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Variational Mode Directed Deep Learning Framework for Breast Lesion Classification by Saini M. et.al [36] presents an innovative deep learning framework based on variational mode decomposition (VMD) for the classification of benign and malignant tumours. The system was evaluated using various public datasets, including UDIAT, and compared against AlexNet, MobileNetV2, ResNet-50, and attention-augmented ResNet-18. On UDIAT, the approach attained nearly flawless results: 98% accuracy, a specificity of 1.00, and both AUROC and AUPRC of 1.00, with only one false negative recorded. The research emphasized the benefit of integrating frequency-domain decomposition with CNN-based categorization. The authors warned that the very high findings may indicate overfitting owing to UDIAT's limited test set, highlighting the necessity for bigger and more diversified validation cohorts.

A Self-Supervised Framework for Improved Generalisability in Ultrasound B-mode Image Segmentation was presented by Ellis et al. [37] that investigated self-supervised learning (SSL) for breast lesion segmentation, with the objective of enhancing generalization in data-scarce environments. Utilizing UDIAT as one of the assessment datasets, they contrasted supervised segmentation models with SSL-enhanced models. With constrained labeling (20% and 50% of training data), SSL enhanced UDIAT segmentation Dice scores by 6.4% and 3.7%, respectively. The research indicated that SSL techniques are especially beneficial for inadequately labeled medical datasets. Nonetheless, the scientists observed that when comprehensive labeled datasets are accessible, supervised models maintain superiority, whereas SSL increases further complexity.

Xie et al.[38] presented research US-Net: Ultrasound Network with Attention Gates for Segmentation that introduced US-Net, a modified U-Net with attention gates to emphasize prominent lesion characteristics. On UDIAT, US-Net attained a Dice coefficient of 94.38%, markedly surpassing conventional U-Net baselines. The scientists ascribed this enhancement to the capacity of attention gates to mitigate extraneous background areas and highlight lesion margins. They highlighted the clinical significance of enhanced segmentation precision for later classification and diagnosis. They acknowledged that segmentation quality may not consistently correlate with enhanced classification results, and the applicability to other imaging centers remains ambiguous.

Merging CNN and Transformer Architectures for Breast Ultrasound Image Segmentation was presented by Huaikun Zhang et.al.[39]. The FET-UNet system amalgamates ResNet34-based convolutional blocks with Swin Transformer modules in a parallel architecture, unified by a sophisticated feature aggregation module. Assessed using BUS, BLUI, and UDIAT, FET-UNet surpassed traditional U-Nets and Attention U-Nets. On UDIAT, it attained a Dice score of 88.9%, above baseline segmentation techniques. The research illustrated the efficacy of hybrid CNN–Transformer architectures in collecting both local details and long-range relationships in ultrasound images. In this approach, the scientists observed heightened computational complexity and ongoing challenges increased with tiny or low-contrast lesions.

BUS-Set Benchmark for Breast Ultrasound Segmentation by Thomas et al. [40] developed BUS-Set, a standardized benchmark that integrates many public breast ultrasound datasets, including UDIAT. They assessed several segmentation designs, including U-Net, Attention U-Net, Swin-U-Net, Trans-U-Net, DeepLabV3+, and Mask R-CNN. On UDIAT, the performance exhibited significant variability: the regular U-Net attained around 0.80 Dice, but enhanced topologies such as Swin-U-Net and Sk-U-Net reached around 0.85 Dice. The research highlighted that irregular preprocessing and non-standardized dataset divisions resulted in diversity among previous studies. They promoted standardized criteria to facilitate equitable comparisons.

Other Datasets

The Improved Quantum-Inspired Binary Grey Wolf Optimizer (IQI-BGWO) was presented by Bilal et al. [41] in order to improve the Support Vector Machine (SVM) for the detection of breast cancer. Generalizability is hampered by the study's use of only the Mammographic Image Analysis Society (MIAS) dataset. In spite of this, IQI-BGWO-SVM enhances feature selection and

Received: August 02 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

classification accuracy. This approach uses optimization inspired by quantum mechanics to improve medical imaging.

A DL -based Raman spectroscopy model was presented by Zeng et al.[42] to diagnose triple-negative and HER2-positive breast cancer. The study's shortcomings include a limited sample size (75 samples), no external validation, no direct comparison with clinical procedures, and possible spectrum variability brought on by outside influences, despite its excellent accuracy (CNN: 91.11%). Furthermore, the method could need to be further optimized for resilience across a range of patient demographics and real-world clinical integration.

Ma et al. [43] created a method for early breast cancer diagnosis based on Surface-Enhanced Raman Spectroscopy (SERS) using a composite Ag NPs PSi Bragg reflector SERS substrate. According to the study, it was inexpensive and had good diagnostic accuracy (95%), specificity (96.7%), and sensitivity (93.3%). It was limited, by the short dataset (60 serum samples), the lack of outside validation, and possible spectrum variability. To confirm its therapeutic application in bigger and more varied groups, additional study is required.

For automated breast cancer metastasis identification, Vulli et al. [44] suggested a refined DenseNet-169 model that makes use of FastAI and the 1-Cycle policy. The model outperformed current techniques with an accuracy of 97.4% after being trained on an improved PatchCamelyon (PCam) dataset from Camelyon16. In order to facilitate early detection, a mobile application was released. Even while the model improves sensitivity and specificity, it still has drawbacks such high computing costs, overfitting risks, manual hyperparameter adjustment, and considerable data augmentation, which may restrict its applicability to clinical settings for non-experts.

Srinivasu et al.[45] used explainable AI (SHAP) with ANOVA-based feature selection to create a CatBoost+MLP model for breast cancer detection. They obtained a 99.3% accuracy rate using the Breast Cancer Wisconsin dataset, which consisted of 569 records. The model decreased overfitting and handled categorical data well. It performed better than traditional methods. increase robustness and interpretability, the authors proposed voting and stacking strategies as future developments. Through their efforts, AI-driven diagnostic algorithms for more precise and open breast cancer prediction have advanced. However, the study's generalizability was impacted by the quantity of the dataset.

In this survey, we investigated several deep learning and machine learning techniques for diagnosing breast cancer using datasets from ultrasound, histopathology, and other sources. A thorough summary of the work we have investigated in this area is given in Table 1. Deep transfer learning techniques have been investigated to boost classification; Sirjani et al. [33] and Tagnamas et al. [32] improved InceptionV3 for ultrasound image classification, while Gupta et al.38 used multi-layered Resnet features. Isik et al. [30] presented meta-learning (ProtoNet-Resnet50) for few-shot classification, whereas Pacal et al. [28] and Gheflati et al. [31] studied Vision Transformer-based models. There have been studies on models that focus on segmentation, such as Zhang et al. [34] HAU-Net, combined CNNs and transformers for the segmentation of breast lesions, and Munteanu et al. [27], coupled GAN-augmentation based on data, segmentation via UNet, and classification using CNN. Alotaibi et al. utilized image fusion and CNN for ultrasound-based diagnosis. Raman spectroscopy and other techniques have also been examined. Zeng et al. [42] utilized CNNs for the detection of HER2-positive and triple-negative breast cancer, whereas Ma et al. [43] applied SERS-based silver nanoparticles for early-stage diagnosis. Bilal et al.[41] employed IQI-BGWO for the optimization of SVM, enhancing feature selection and classification. Explainability and model interpretability have garnered attention, as Srinivasu et al. [45] used CatBoost+MLP and SHAP-based explainability for breast cancer detection. Vulli et al. [44] optimized Densenet-169 employing FastAI and the 1-Cycle strategy, while also creating a mobile application for prompt identification.

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Notwithstanding these gains, obstacles persist, including dataset restrictions like class imbalance and fewer sample numbers, which impede the model's generalization capabilities. Our suggested approach seeks to resolve these challenges by using measures to enhance performance and rectify data imbalance, hence assuring equitable representation of all classes. We prioritize the interpretability and transparency of our model through the application of hybrid CNN methods. These methodologies elucidate the decision-making processes of the model, facilitating its clinical implementation. Future research should continue to investigate hyperparameter tuning, feature selection, scalable architectures, and explainable AI to improve transparency and reliability in breast cancer diagnosis.

Table 1. Summary of related research on breast cancer detection and classification.

Author	Model	Dataset	Accuracy (%)	Limitations
Munteanu et. al.[27]	UNet+CNN	BUS[24]	86	Small dataset size and high computation cost
Pacal et al.[28]	Vision Transformer	BUS[24]	88.6	Small dataset size
Alotaibi et. al.[29]	VGG19	BUS[24], KAIMRC	87.8 (BUS[24])	Class imbalance and high computation cost
Isik et al.[30]	ProtoNet+Resnet50	BUS[24]	88.9	Dependency on dataset similarity
Gheflati et. al.[31]	ViT B/32	BUS[24]	86.7	Small dataset affects generalizability
Tagnamas et. al.[32]	SCA-InceptionUNeXt	Multiple datasets	81.66 (BUS[24])	Low performance and lacked interpretability
Sirjani et al.[33]	Enhanced InceptionV3	Multiple datasets	81 (BUS[24])	Model generalization issues
Zhang et al.[34]	HAU-Net	Multiple datasets	83.11 (BUS[24])	Model generalization issues
Singh et al.[35]	ViT+EfficientNet	BUS[24], UDIAT[25]	96.90	Small dataset size
Manali Saini et.al[36]	(VMD)decomposition	Multiple datasets	98.0(UDIAT[25])	Overfitting due to small dataset
Ellis et al.[37]	Self Supervised CNN	Multiple datasets	97.2 (UDIAT[25])	Not suited for multiple clinical domains
Wang et al.[38]	Modified-UNET	UDIAT[25]	94.38 (UDIAT[25])	Results ambiguous for other imagings
Huaikun Z. et. al,[39]	ResNet + Swin Transformer	Multiple datasets	88.9 (UDIAT[25])	Computational complexity with low contrast problems
Thomas et al.[40]	Multiple CNN	UDIAT[25], others	85.00	Bias due to the lesion size variation
Bilal et al.[41]	IQI-BGWO+SVM	MIAS	99.25	Model complexity constraints
Zeng et al.[42]	CNN	75 serum samples	91.11	Small dataset affects generalizability
Ma et al.[43]	SERS	60 serum samples	95	Small dataset affects generalizability
Vulli et al.[44]	Fine-tuned Densenet	PatchCamelyon	97.40	Model complexity constraints
Srinivasu et. al.[45]	CatBoost+MLP	Wisconsin	99.3	Small dataset affects generalizability

3. Proposed Methodology of hybrid "VIMONET" model

In clinical practice, the integration of DL techniques and computer-aided diagnostic (CAD) methods provides medical practitioners and clinicians with improved speed, efficiency, cost-effectiveness, and precise diagnostic outcomes [46]. Thus, CNNs play an important role in medical imaging tasks, including extracting features and classifying tumour lumps [47]. The proposed model utilizes a hybrid

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

of Vision Transformer (ViT) and MobileNetV3-Small architectures to categorize images from both datasets for binary (benign vs. malignant) classification, using all available data from multiple sources. The MobileNetV3Small [48] and Vision Transformer (ViT) [49] designs are the two networks that are combined to generate the hybrid network forming the basis of the proposed technique.

MobileNetV3-Small: MobileNetV3Small [48] is a lightweight convolutional neural network that is specifically designed for efficiency on mobile and embedded devices while maintaining strong performance in classification tasks. Here is a summary of its key characteristics:

1. ***Efficient Architecture Design*** – Combines depth wise separable convolutions with advanced architectural optimizations to reduce computation without sacrificing accuracy.

2. ***Lightweight Model Size*** – With approximately 2.5 million parameters, it is significantly smaller than traditional CNNs, making it highly efficient and deployable in resource-constrained environments.

3. ***Inverted Residual Blocks with Squeeze-and-Excitation (SE)*** – Uses Mobile Inverted Residual Blocks (MBConv) enhanced with SE modules to improve channel attention and strengthen feature representation.

4. ***Hard-Swish Activation Function*** – Employs Hard-Swish, a computationally efficient approximation of Swish, improving non-linearity and convergence while maintaining low latency.

5. ***Optimized for Mobile/Low-Power Devices*** – Achieves competitive accuracy with minimal computational cost (only 65 million FLOPs), making it highly suitable for medical imaging applications where efficiency is crucial.

6. ***Transfer Learning Friendly*** – Pre-trained on ImageNet, enabling effective transfer learning for specialized domains like breast ultrasound image classification.

MobileNetV3-Small offers an excellent lightweight trade-off between accuracy, efficiency, and deployment feasibility, serving as a strong backbone for deep learning tasks in medical imaging.

Vision Transformer (ViT): The self-attention mechanisms of Vision Transformer (ViT) models make them powerful for image analysis tasks. ViT models are widely used to achieve cutting-edge results [49]. After receiving 2D images as input, the ViT splits the image into multiple equally sized patches. A standard transformer encoder processes the vector sequence obtained from the linear embedding of these patches. An additional learnable “classification token” is appended to this sequence of vectors. The integration of the transformer model into our approach is motivated by its ability to capture long-range dependencies and contextual relationships within images. This is particularly vital in the analysis of histological images, where spatial arrangement and cellular morphology play a crucial role in accurate diagnosis.

An EffNetV2-ViT model proposed by M. Hayat et al. [50] focused on the BreakHis dataset of split open biopsy (SOB) images only. In contrast, our proposed hybrid model combines the strengths of both MobileNetV3-Small and the transformer-based ViT encoder for robust BUS image classification. The following steps outline our hybrid model design for image classification:

3.1 Data Collection and Preprocessing

We have utilized two datasets to demonstrate the accuracy and efficiency of our approach.

The first dataset model employed is from Baheya Hospital for Early Detection and Treatment of Women's Cancer, Cairo, Egypt [24]. Data accessibility is available at the website https://scholar.cu.edu.eg/?q=afahmy/ pages/dataset. The Baheya Breast Ultrasound Dataset (BUS) consists of labelled pretrained weights derived from ImageNet database images, including 487 benign images and 210 cancerous images.

Our second dataset comprises the UDIAT dataset (B) from Spain. The UDIAT dataset samples were collected at the UDIAT diagnostic center of the Parc Tauli Corporation in Sabadell, Spain [25]. This dataset has 163 ultrasound images accompanied with pretrained weights from ImageNet pertaining to

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

breast cancers. In these samples, benign and malignant breast cancer instances are 109 and 54 respectively. In the dataset, the ultrasound images had a resolution of 760x570 pixels. The dataset is accessible at this location: [URL https://www.kaggle.com/datasets/jarintasnim090/udiat-data (accessed on 28th May, 2025)]

The data preprocessing pipeline comprises of following subsequent sequencing:
• Loads and preprocesses images from the designated directory. Each of the .png images are resized to a specified dimension and verifies the existence of a corresponding mask file (designated by _mask in the filename).
• The ultrasound images are initially resized to uniform dimensions of 224 × 224 pixels to prepare them for training.
• When a mask is detected, it normalizes the mask and employs it to accentuate the lesion region, designated as the Region of Interest (ROI), while diminishing the backdrop using the image's average colour.
• The processed images are added to a list and returned as output.
• This facilitates the preparation of lesion-specific input images for DL models in medical imaging applications, such as breast ultrasound analysis.

A crucial strategy for artificially expanding the dataset is data augmentation [51], which enhances the model's accuracy and robustness while preventing overfitting in image categorization.
Subsequently, we apply image augmentation techniques to each magnification level independently to ensure uniform data representation. Employing random rotation of 15%, vertical 10% and horizontal flipping 10%, zoom adjustments of 15%, and contrast modification of 10%.

3.2 Model Architecture of VIMONET

The proposed hybrid architecture with CNN and ViT is shown in Figure-3. This hybrid model integrates the advantages of both CNN and ViT along with an enhanced Self-Attention (SA) mechanism for medical image analysis. Below is the description of modules:

I. ***CNN Module****:* The diagram in Figure-1 shows the architecture of MobileNetV3-Small [48] which is a lightweight convolutional neural network (CNN) designed for efficient performance on mobile and embedded devices for image classification. It begins with a standard 3×3 convolution to extract initial features from the input. The core of the network is built from a sequence of bottleneck layers, which are compact blocks that expand the input channels, apply depth-wise convolutions, and then project back to a lower dimension. Some of these bottlenecks include SE (Squeeze-and-Excitation) modules, which help the model focus on the most informative channels, and most of them use the h-swish activation function, a computationally efficient variant of swish that improves accuracy. After passing through several stacked bottleneck layers that progressively capture richer representations, the network applies a 1×1 convolution to adjust the feature space before a global pooling operation followed by a final 1×1 convolution for classification. Overall, this architecture balances accuracy and efficiency by combining bottleneck blocks, SE attention, and h-swish activations, making MobileNetV3-Small well-suited for resource-constrained environments.

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

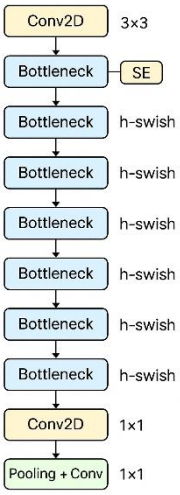


Figure 1: MobileNetV3Small block diagram

Local features are recovered using convolutional layers, with only the last 40 layers of model layers are unfreeze to support transfer learning for the initial layers of the model. To mitigate overfitting and preserve spatially invariant characteristics, the feature map is then sent to the Global Average Pooling 2D layer, which averages the spatial dimensions of height and breadth before relaying it to the final classification layer in the CNN. The output of the model is then integrated with an induced inception layer for a better feature selection.

II. ***Inception Module****:* The Inception module functions as a feature enhancement mechanism that enriches representational capacity by applying parallel convolutional and pooling operations to the input feature map (Figure 2(a)). This architecture captures information at multiple spatial scales, thereby producing more discriminative feature representations. The input features, typically derived from a CNN backbone prior to the Transformer stage, have dimensions of (7, 7, 2048). Before fusion, both feature sets are projected into a common dimensional space using fully connected layers with ReLU activation to ensure comparability in scale and size.
The Inception block comprises four parallel branches:
- A 1×1 convolution (256 filters), which performs dimensionality reduction while preserving spatial structure.
- A 3×3 convolution (256 filters), responsible for extracting medium-scale local patterns and textures.
- A 5×5 convolution (256 filters), which captures broader contextual information across a larger receptive field.
- A 2×2 max pooling operation (stride = 1), followed by a 1×1 convolution, which emphasizes strong activations and provides translation invariance without reducing spatial resolution.

The outputs from these four branches are concatenated along the channel axis, producing a combined feature map with 1024 channels (256 per branch). By integrating filters with different receptive fields and pooling operations, the Inception module captures both fine-grained details and global contextual cues. This diverse feature representation is crucial for enhancing the robustness of downstream Transformer-based processing.

III. ***Transformer Module****:* This module will process CNN-extracted features as sequences and applies Multi-Head Self-Attention (MHSA) to capture global context.

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

For the ViT vision transformer, we follow the following procedure:
i.We create patches of 16x16 pixel dimensions for our image thereby, creating a total of 196 patches per image.
ii.Perform patch + position encoding

Patch Embedding:
Step 1: An image is split into fixed-size patches (e.g., 16×16 pixels).
Step 2: Each patch is ***flattened*** into a 1D vector.
Step 3: These vectors are ***linearly projected*** into a fixed-dimensional embedding space (like 768D). So, an image of shape H×W×C (Height × Width × Channels), and a patch size of P×P is changed into (H/P) × (W/P) patches. Now each patch becomes a token (like words in NLP).

Positional Encoding: Transformers need a way to inject spatial information. This is accomplished by adding (or concatenating) a learned or fixed sinusoidal positional embedding to each patch embedding. The weakness in terms of expressive power that Transformers exhibit due to order- and proportion-invariance has motivated the need for including information about the order of the input sequence by other means; in particular, this is often achieved by using positional encodings [49]. This encoding tells the model where in the image each patch came from.

Use of transformer encoder X4: The transformer block is shown in Figure 2(b) where, each encoder block usually contains Multi-Head Self-Attention (MHSA), Add & Layer Normalization, Feed-Forward Network (MLP), Add & Layer Normalization.

Finally, a layer Normalization step is applied to the fused feature vector. This helps stabilize training, improves convergence, and maintains consistent feature distributions. The final output is a clean, normalized feature vector of shape (batch_size, projection_dim) that serves as an effective combination of the two input representations.
\

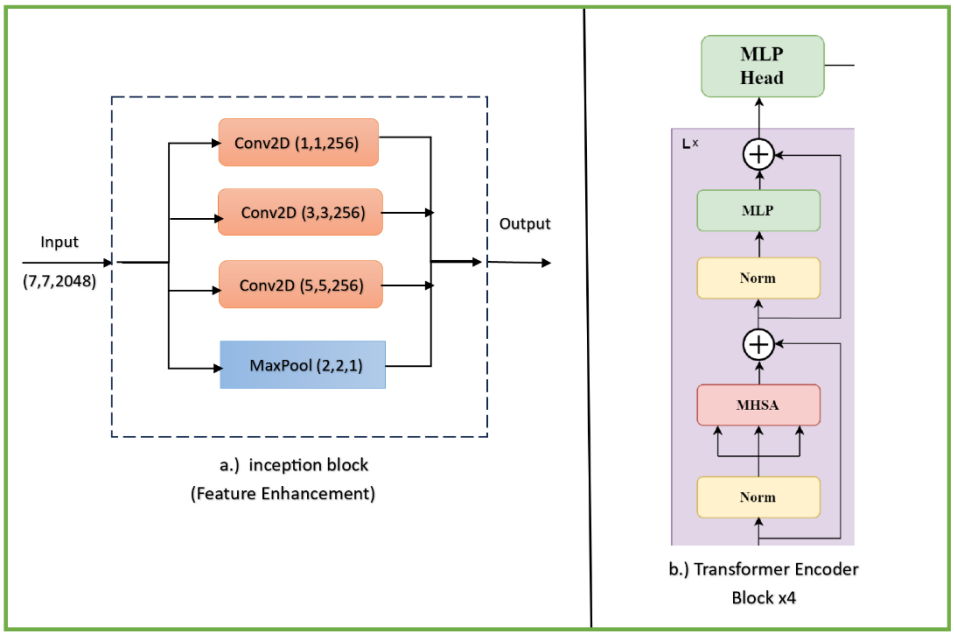


Figure 2: (a) Inception block
(b) Transformer encoder X4 block

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

IV. ***Perform Normalization:*** Normalization is performed finally in both modules of CNNs and Transformers. However, in CNNs batch normalization is done after Conv or Dense layer whereas, in Transformers Layer normalization is performed after residual and attention or MLP layer.

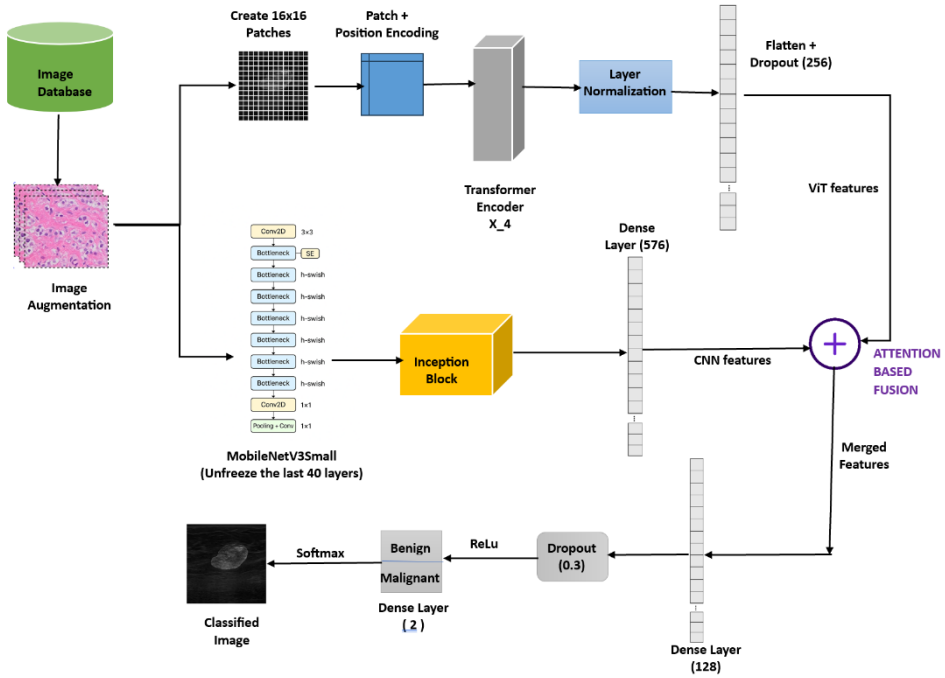


Figure-3: Proposed Hybrid model VIMONET extracting features using a self-attention based enhanced hybrid model using ViT and MobileNetV3Small for image classification.

3.3 Proposed ViT + MobileNetV3_Small ensemble-based hybrid model (VIMONET):

The diagram of the proposed VIMONET model as shown in Figure-3 above representing a hybrid DL model that combines Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for medical image classification (benign or malignant).

Pipeline summary of the model: The proposed model follows a 4-stage pipeline designed to effectively leverage both global contextual features and localized spatial information for breast ultrasound classification.

i). Image Input and Augmentation: To reduce overfitting and mitigate class imbalance, image augmentation was performed on both benign and malignant samples. For consistency, each class was augmented to include 500 images in both the BUS and UDIAT datasets, resulting in a total of 2000 images used in the study. This strategy ensured balanced class representation while enhancing model generalization.

ii). Parallel Feature Extraction Ensemble: The network adopts a dual-path ensemble to capture complementary feature representations:

a). Vision Transformer (ViT) feature extraction: Each image is partitioned into 16×16 patches, which are linearly embedded and enriched with positional encodings. These embeddings are processed through a 4 -block Transformer encoder to capture global dependencies. Following normalization, flattening, and dropout, the ViT pathway produces a 256-dimensional feature vector.

b). CNN feature extraction (MobileNetV3-Small with Inception Block): In parallel, the augmented image is processed through MobileNetV3-Small, fine-tuned by unfreezing the last 40 layers for improved adaptation. An Inception block is appended to this pathway, enabling the extraction of multi-scale representations through 1×1, 3×3, and 5×5 convolutions combined with pooling operations. This path yields a 576-dimensional feature vector.

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

iii). Self-Attention–Based (SA) Fusion: To effectively integrate the complementary information from both pathways, the model employs a self-attention–based fusion mechanism. This approach allows the network to assign higher importance to the most discriminative features, ensuring that both, the global as wee as local features to contribute to the final representation.
iv). Final Classification Head: The fused features are processed by a classification head composed of a Dense layer (128 units), followed by Dropout (rate = 0.3), and a final Dense layer with two units. A Softmax activation generates the final prediction, classifying each input image as either benign or malignant.
This model leverages the strength of CNNs in capturing local spatial features (via MobileNetV3 + Inception) and the strength of Transformers in modeling global dependencies (via ViT). The attention-based fusion ensures complementary information from both is effectively integrated, leading to better classification performance in medical imaging tasks. The detailed algorithm of the model has also been explained in the section 3.5.

3.4 Training Strategy: We have adopted the following mechanism for the training of our model:-
o **Data Split**: In the above-described module summary we've stated the technique for data augmentation in order to reduce overfitting. After data augmentation, the imbalance in data class was remitted by assigning 500 images for each class (Benign and Malignant) and per dataset viz BUS and UDIAT.
We have split the dataset in the ratio of 70% for Training, 20% of unseen data for Testing and 10% of random data for the Validation of our model. Therefore, the total number of images that have been processed are 1200 (BUS)+1200 (UDIAT) = 2400 Total.

o **Loss Function:** Binary Cross-Entropy (BCE) has been used as a loss function because it is appropriate for binary classification tasks, such as distinguishing between benign and malignant tumours. It measures the difference between the predicted probabilities and the actual binary labels.
Binary Cross-Entropy Formula:

$$\text{Log loss} = \frac{1}{N} \sum_{i=1}^{N} -\left(y_i * \log(p_i) + (1-y_i) * \log(1-p_i) \right)$$

Where:
- y_i: Ground truth label (0 for benign, 1 for malignant).
- p_i: Predicted probability for class .
- N: Total number of samples.

o **Optimizer:** In our model, we have employed an adaptive optimizer to repeatedly update the model parameters, minimize the loss function, accelerate convergence, and efficiently manage gradient sparsity. We've used Rectified Adam [52] instead of the Adam optimizer, as it incorporates a rectification function that dynamically adjusts the learning rate to mitigate excessive fluctuations during initial phases and is particularly advantageous for hybrid architectures with variable learning rates.

• **Learning Rate Scheduler**: LR scheduler dynamically adjusts the learning rate based on validation loss or epoch. Here, we employed Cosine Decay Learning Rate Scheduler, which progressively diminishes the learning rate following a cosine curve instead of a precipitous decline. The parameters consist of the initial learning rate, decay steps (epochs multiplied by steps per epoch), and the final learning rate (alpha multiplied by the original learning rate), which guarantees that the learning rate does not approach zero.

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The optimizer, facilitates the stabilization of training and enhances generalization by gradually diminishing the learning rate, particularly in the latter stages of training. The Rectified Adam (RAdam) Optimizer, together with Decoupled Weight Decay, mitigates variation during initial training phases, resulting in enhanced stability as shown in the Table 2.

Table 2: Hyperparameters used

Parameter	**Method Used**
Input Size	224 × 224
Epoch	20
Batch Size	16
Learning Rate	0.00001
Patch Size	16
Loss Function	Binary Cross Entropy (BCE)
Optimizer	Rectified Adam (RAdam)
Learning Rate Scheduler	Cosine-Decay

3.5 Proposed Algorithm:

This section describes the algorithms for different modules used in our research paper. The first algorithm is stated below. It describes the **MobileNetV3-Small** CNN whose output is fed to **inception block** for rich feature extraction.

procedure MOBILENETV3_INCEPTION_FEATURE_EXTRACTION
Input: X ∈ ℝ^{H×W×C} ⟶ Input image tensor where H, W, and C represent height, width, and number of channels, respectively
 # Step 1: Load MobileNetV3-Small Backbone
Load MobileNetV3-Small pretrained on ImageNet with:
- input_shape = (H, W, C)
- include_top = False
- include_preprocessing = True

Freeze all layers of the backbone
Unfreeze the last 40 layers of the backbone to enable fine-tuning
 # Step 2: Feature Extraction
Pass X through the backbone network
x ← backbone(augmented_inputs)
⟶ x ∈ ℝ^{224X224XC} where C is the number of output channels
 # Step 3: Multi-Scale Inception Block
conv1 ← 1×1 convolution with 256 filters and ReLU activation on x
conv3 ← 3×3 convolution with 256 filters and ReLU activation on x
conv5 ← 5×5 convolution with 256 filters and ReLU activation on x
pool ← 2×2 max-pooling with stride = 1 and same padding on x
inception_features ← Concatenate([conv1, conv3, conv5, pool])
 # Step 4: Final Representation
x ← Global Average Pooling on inception_features
x ← Dense layer with 576 units and ReLU activation applied to x
Return: Inception_features
end procedure

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Algorithm of the Vision Transformer **ViT** is as follows:

```
procedure VIT CLASSIFIER
  Input: X ∈ R^{H×W×C} ⟶ Input image tensor where H, W, and C represent height, width, and number of channels, respectively
  Compute N, the number of patches, using N = H*W/P^2 where P is the patch size ⟶ Equation 1.
  Encode patches using PatchEncoder
  for i = 1 to N_transformer_layers do
    x1 ← Layer normalization on encoded patches
    Compute multi-head attention on x1 with num_heads heads and projection_dim key dimension
    x2 ← Add attention output and encoded patches ⟶ Skip connection 1
    x3 ← Layer normalization on x2
    x3 ← Multi-Layer Perceptron (MLP) on x3 with hidden units transformer_units and dropout rate
    Update encoded patches with Skip connection 2
  end for
  Perform Layer normalization on encoded patches
  Flatten the representation
  Apply dropout regularization with dropout rate
  Apply MLP to the representation with hidden units mlp_head_units and dropout rate
  Classify outputs using MLP Head layer
  Output: Y ⟶ Class predictions
end procedure
```

The final hybrid model is then created using the above discussed procedures using self- attention (SA) based feature extraction from vision transformer and enhanced convolution neural network developed. The theoretical explanation of the model has already been discussed earlier in the section 3.3.

```
procedure HYBRID_MN_ViT_CLASSIFIER
  Input: X ∈ ℝ^{H×W×C} ⟶ Input image tensor where H, W, and C represent height, width, and number of channels, respectively
  # Step 1: Data Augmentation
  X_aug ← Apply data_augmentation(X)
  # Step 2: MobileNetV3-Small Backbone
  Load MobileNetV3-Small pretrained on ImageNet with:
    - input_shape = (H, W, C)
    - include_top = False
    - include_preprocessing = True
  Freeze all layers of the backbone
  Unfreeze the last 40 layers for fine-tuning
  Pass X through the backbone network
  x ← backbone(augmented_inputs)
  # Step 3: Inception Block for Multi-Scale Features
  conv1 ← 1×1 convolution with 256 filters and ReLU activation on x
  conv3 ← 3×3 convolution with 256 filters and ReLU activation on x
  conv5 ← 5×5 convolution with 256 filters and ReLU activation on x
  pool ← 2×2 max-pooling with stride 1 and same padding on x
  inception_features ← Concatenate([conv1, conv3, conv5, pool]) along channel dimension
  # Step 4: Flatten and Project Inception Features
  x ← Global Average Pooling on inception_features
  inception_features ← Dense(576)(inception_features)
  ⟶ Projected to contribute ViT feature dimension
  # Step 5: Vision Transformer (ViT) Feature Extraction
  vit_features ← create_vit_features(X)
  ⟶ vit_features ∈ ℝ^{1×576}
  # Step 6: Attention-Based Feature Fusion
  fused ← attention_fusion(vit_features, inception_features)
  ⟶ fused ∈ ℝ^{1×576}
  # Step 7: Classification Head
  h ← Dense(128, activation=ReLU)(fused)
  h ← Dropout(h, rate = 0.3)
  y ← Dense(K, activation=Softmax)(h)  ⟶ K = number of classes
    Output: Y = y ⟶ Classified Image (Tumor)
end procedure
```

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4. Performance Analysis

4.1 Simulation Setup

To perform our training, testing and evaluation the hardware and software setup used for predicting malignancy is listed below:

Hardware Used: -

Processing Unit (CPU): Intel Core i7 11th Gen with 16 GB of RAM
Graphic Processing Unit (GPU): 4 GB Nvidia GeForce RTX3050-Ti

Software Used (but not limited to):-

Visual Studio Code editor
Python 3.11.8
TensorFlow 2.15.0
Keras 0.20

4.2 Evaluation Parameters

To evaluate the robustness of the proposed model, confusion matrix is considered and parameters like accuracy, specificity, Precision, recall and F1-Score.

1. $\text{Accuracy} = \dfrac{TP + TN}{TP + TN + FP + FN}$
2. $\text{Precision} = \dfrac{TP}{TP+FP}$
3. $\text{Recall} = \dfrac{TP}{TP+FN}$
4. $\text{F1-Score} = \dfrac{TP}{TP + 0.5(FP + FN)}$

TP represents true positive cases where malignant cases were accurately diagnosed, TN denotes true negative cases where benign cases were correctly diagnosed, FP refers to false positive cases where benign cases were erroneously identified, and FN indicates false negative cases where malignant cases were mistakenly classified as benign. The validation will encompass the area under the curve (AUC) and receiver operating characteristics (ROC).

4.3 Comparative Analysis

Table 3: Comparative Performance analysis of proposed hybrid model VIMONET on BUS and UDIAT dataset respectively

(a) Comparative Performance analysis of proposed hybrid model VIMONET on BUS dataset

Model	Total Params	Trainable Params	FLOPs (GFLOPs)	Test Accuracy	Recall	Precision	F1-score	AUC	Specificity	Total Training Time (s)	Average Epoch Time (s)
Hybrid VIMONET model	6352344	5985864	0.6416	0.99	0.98	0.98	0.98	0.9962	0.98	193.75	9.68
MobilenetV2	6571778	6537666	0.6518	0.96	0.98	0.94	0.96	0.9923	0.94	342.30	17.11
ViT Model	2306261	2106267	0.2197	0.56	0.36	0.6	0.45	0.602	0.76	169.95	7.49
ResNet50	25687938	25634818	7.7340	0.96	0.95	0.96	0.95	0.9945	0.97	957.30	47.86
EfficientNetB2	9213435	9145860	1.3446	0.98	0.99	0.98	0.99	0.9996	1.00	928.67	46.43

(b) Comparative Performance analysis of proposed hybrid model VIMONET on UDIAT dataset

Model	Total Params	Trainable Params	FLOPs (GFLOPs)	Test Accuracy	Recall	Precision	F1-score	AUC	Specificity	Total Training Time (s)	Average Epoch Time (s)
Hybrid VIMONET model	6352344	5985864	0.6416	0.985	0.97	1.00	0.9847	0.9938	1.00	197.90	9.89
Mobilenet_V2	6571778	6537666	0.6518	0.89	0.99	0.825	0.9	0.9922	0.79	348.61	17.43
ViT Model	2306261	2106267	0.2197	0.64	0.55	0.670	0.6043	0.6366	0.73	173.884	7.69
ResNet50	25687938	25634818	7.7340	0.965	0.98	0.951	0.9655	0.9981	0.95	989.86	49.49
EfficientNet_B2	9213435	9145860	1.3446	0.98	0.97	0.989	0.9797	0.998	0.99	961.04	48.05

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

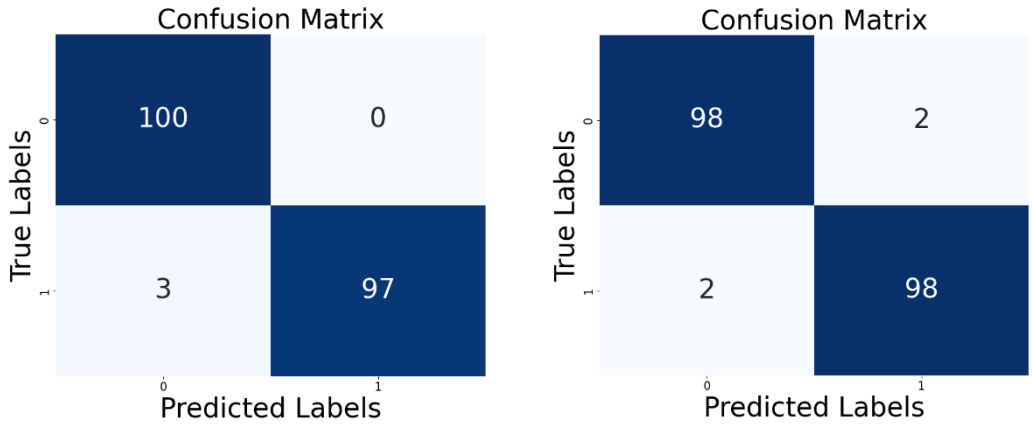
Performance analysis of the Hybrid VIMONET Model: The Hybrid VIMONET model was benchmarked against MobileNetV2, ViT, ResNet50, and EfficientNetB2 on the BUS and UDIAT breast ultrasound datasets. The experiment results from the Table 3 above yields the following summary:

On the BUS dataset (Table 3a), Hybrid VIMONET achieved the highest accuracy (0.99) with balanced performance across metrics (Recall = 0.98, Precision = 0.98, F1 = 0.98, AUC = 0.9962, Specificity = 0.98). It also demonstrated superior computational efficiency, requiring only 193.75 s training time (average 9.68 s per epoch), compared to >900 s for deeper models. ViT underperformed (Accuracy = 0.56, F1 = 0.45), suggesting limited suitability without large-scale adaptation, while EfficientNetB2 provided slightly higher Specificity (1.0) and AUC (0.9996) but at substantially greater cost.

On the other hand, the UDIAT dataset (Table 3b), Hybrid VIMONET again demonstrated strong performance (Accuracy = 0.985, Recall = 0.97, Precision = 1.0, F1 = 0.9847, Specificity = 1.0, AUC = 0.9938), achieving perfect specificity with no false positives. It exhibited enhanced computational efficiency, necessitating about 197.9 seconds of training time (an average of 9.89 seconds per epoch). MobileNetV2 yielded higher recall (0.99), but, it suffered from poor specificity (0.79), while deeper models such as ResNet50 and EfficientNetB2 achieved comparable accuracy (~0.965–0.98) but at substantially higher computational cost (>960 s training, >48 s per epoch). ViT once again lagged behind (Accuracy = 0.64, Specificity = 0.73).

Overall, Hybrid VIMONET consistently provided the best balance of accuracy, robustness, and efficiency, outperforming both CNN- and transformer-based baselines on two independent datasets. It is also obvious from the table 3 that the VIMONET model is much lightweighted as it consumes just as small as 0.6416 giga FLOPS operations as compared to existing heavy state of art models. Finally, with respect to the average training time per epoch also we can earmark that our model takes less than 10 seconds/epoch execution time next to Vit model which has a poor performance that is nowhere near to our VIMONET model.

Table 4: Confusion Matrix of BUS and UDIAT respectively



Analysis of the Confusion Matrix: The classification efficacy of the proposed Hybrid VIMONET model was further assessed by confusion matrix analysis on the BUS and UDIAT breast ultrasound imaging datasets, with class 0 denoting benign tumours and class 1 indicating malignant tumours as shown in the Table 4.

The examination of the confusion matrix further confirmed the trustworthiness of the Hybrid VIMONET model. In the BUS dataset, the model attained 98% sensitivity and 98% specificity, accurately identifying 98 benign and 98 malignant cases, with just two misclassifications each category. The model attained perfect specificity (100%) on the UDIAT dataset, with no false positives, and achieved 97% recall for malignant patients (97 out of 100 properly identified). Despite the

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

UDIAT dataset exhibiting a somewhat elevated false negative rate, the lack of false positives highlights its therapeutic efficacy in reducing over-diagnosis. These results underscore the model's equitable diagnostic proficiency, substantial applicability across datasets, and potential to facilitate sound therapeutic decision-making.

Analysis of the Validation Accuracy/Loss of the model:

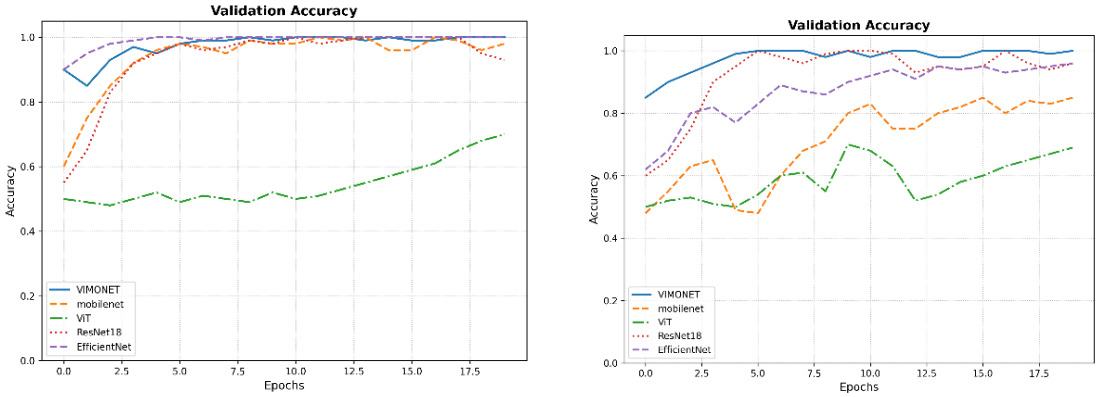


Figure 4: Validation Accuracy curve of (a)BUS and (b)UDIAT respectively

Figure 4(a) illustrates the validation accuracy curves of the proposed VIMONET model in comparison to MobileNet, ViT, ResNet, and EfficientNet on the BUS dataset. The results indicate that our Hybrid VIMONET model exhibits a steady and consistent enhancement, progressively approaching high accuracy over the epochs. While MobileNet, ResNet, and EfficientNet demonstrate marginally superior accuracy in the initial training phases, the Hybrid VIMONET model exhibits competitive performance and ultimately approaches 100% convergence. On the other hand, Figure 4(b) depicts the progression of training accuracy on the UDIAT dataset. The Hybrid VIMONET model, akin to the BUS dataset, exhibits consistent learning, gradually nearing convergence with an accuracy approaching nearly 100%. The baseline CNN models, including MobileNetV2, ResNet50, and EfficientNetB2, demonstrate comparatively lower performance, achieving near-saturation accuracy within a few epochs.

Thus, we can say that the Hybrid model (VIMONET) consistently converges across both datasets, achieving accuracy that is competitive with leading CNN-based models, and shows enhanced stability throughout the training process. The consistently subpar performance of the standalone ViT underscores the benefits of the proposed hybridization strategy, which integrates the locality-aware feature extraction capabilities of CNNs with the global context modelling of vision transformers.

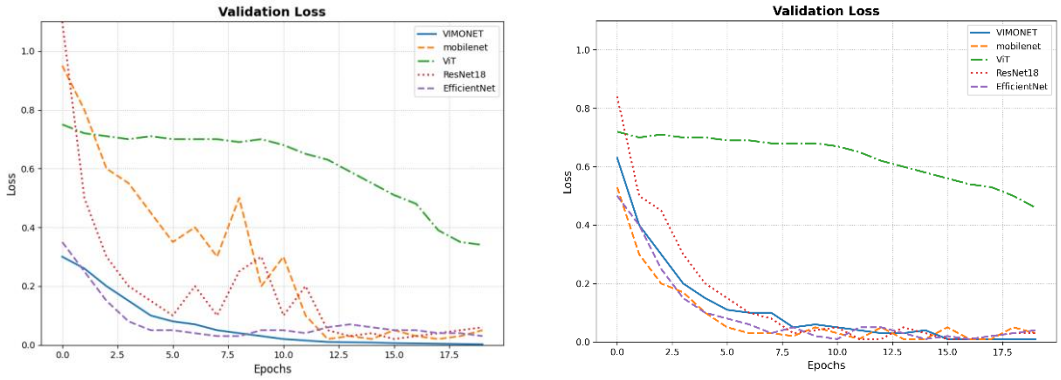


Figure 5: Validation Loss curve of BUS and UDIAT respectively

Received: August 02 , 2025

International Journal of Applied Mathematics
Volume 38 No. 8s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Validation loss patterns (Figure 5) highlight the convergence efficiency of the proposed Hybrid VIMONET model. On both the BUS and UDIAT datasets, VIMONET consistently reduced loss to near zero within a few epochs, demonstrating robust learning and stable convergence. ResNet50 and EfficientNetB2 also achieved rapid convergence, while MobileNetV2 showed occasional oscillations despite low overall loss. In contrast, the ViT model exhibited persistently high training loss, reflecting poor feature learning and limited convergence.

Overall, Hybrid VIMONET achieved the most reliable and stable convergence, reinforcing its superiority in learning discriminative features compared to standalone CNN or transformer baselines.

5. Conclusion and Future Work

In this study, we proposed a lightweighted Hybrid VIMONET model that integrates convolutional neural networks (CNN) with vision transformer (ViT) component to enhance breast ultrasound (BUS) image classification. Our model is considerably lightweighted than all of the other state-of-the-art models discussed by utilizing only 0.6416 FLOPS computations and only consuming less than 9.9 seconds per epoch as compared to other heavy models like ResNet50, EfficientNetB2 or VGG16 itself. The results across the BUS and UDIAT datasets highlight the robustness of the proposed model. In terms of training behaviour, the hybrid model achieved steady accuracy improvements, converging close to 99% while maintaining a smooth decline in training loss, outperforming MobileNetV2, ResNet50, and EfficientNetB2 in stability and efficiency. By contrast, the standalone ViT model consistently underperformed, showing both limited accuracy gains and persistently high loss values. Performance metrics derived from the confusion matrices further validate these findings. On the BUS dataset, the Hybrid VIMONET model achieved 98% sensitivity, 98% specificity, and 98% overall test accuracy, while on the UDIAT dataset it attained 97% sensitivity, 100% specificity, and 98.5% test accuracy. These results demonstrate the model's ability to balance malignant detection with the minimization of false positives, ensuring reliable diagnostic performance.
Overall, the Hybrid VIMONET model not only delivers stable convergence during training but also translates this efficiency into superior classification accuracy, establishing itself as a promising framework for automated breast cancer diagnosis from ultrasound images.

Future Work

Although the proposed Hybrid VIMONET model demonstrates strong performance and robustness across two benchmark datasets, there remain several directions for future research. First, extending the model to larger and more diverse clinical datasets would help validate its generalizability across different populations and imaging conditions. Second, incorporating multi-modal data such as mammography, histopathology, or patient metadata could enhance diagnostic reliability by providing complementary information beyond ultrasound images. Third, integrating explainable AI (XAI) methods would improve clinical trust by allowing radiologists to interpret the decision-making process of the model through visual saliency maps or feature attribution methods.

6. References

1. Bray, F. et al. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 74, 229–263 (2024).
2. Organization, W. H. World health organization (who) — https://www.who.int/ (2024). [Accessed 11-04-2025].
3. Foundation, N. B. C. Breast cancer facts & stats 2024 - incidence, age, survival, & more. https://www.nationalbreastcancer.org/bre ast-cancer-facts/#~: (2024). [Accessed 18-04-2025].
4. Breastcancer.org. Breast cancer facts and statistics 2024 — breastcancer.org. https://www.breastcancer.org/facts-statistics (2024). [Accessed 18-04-2025].

Received: August 02 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5. For Biotechnology Information, N. C. Breast cancer early detection: a phased approach to implementation. https://www.ncbi.nlm. nih.gov/pmc/articles/PMC7237065/ (2024). [Accessed 08-08-2024].
6. Swaminathan, H., Saravanamurali, K. & Yadav, S. A. Extensive review on breast cancer its etiology, progression, prognostic markers, and treatment. Med. Oncol. 40, 238 (2023).
7. Khan, S. I., Shahrior, A., Karim, R., Hasan, M. & Rahman, A. Multinet: A deep neural network approach for detecting breast cancer through multi-scale feature fusion. J. King Saud Univer.-Comput. Inform. Sci. 34, 6217–6228 (2022).
8. Stachs, A., Stubert, J., Reimer, T. & Hartmann, S. Benign breast disease in women. Dtsch. Arztebl. Int. 116, 565 (2019).
9. Zhao, H. The prognosis of invasive ductal carcinoma, lobular carcinoma and mixed ductal and lobular carcinoma according to molecular subtypes of the breast. Breast Cancer 28, 187–195 (2021).
10. Khan, M. S. I. et al. Accurate brain tumour detection using deep convolutional neural network. Comput. Struct. Biotechnol. J. 20, 4733–4745 (2022).
11. Bayram, B., Kunduracioglu, I., Ince, S. & Pacal, I. A systematic review of deep learning in MRI-based cerebral vascular occlusion- based brain diseases. Neuroscience https://doi.org/10.1016/j.neuroscience.2025.01.020 (2025).
12. Modiri, A., Goudreau, S., Rahimi, A. & Kiasaleh, K. Review of breast screening: Toward clinical realization of microwave imaging. Med. Phys. 44, e446–e458 (2017).
13. Iacob, R. et al. Evaluating the role of breast ultrasound in early detection of breast cancer in low- and middle-income countries: A comprehensive narrative review. Bioengineering 11, 262 (2024).
14. Gardezi, S. J. S., Elazab, A., Lei, B. & Wang, T. Breast cancer detection and diagnosis using mammographic data: Systematic review. J. Med. Internet Res. 21, e14464 (2019).
15. Rahman, A. et al. Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. AIMS Public Health 11, 58 (2024).
16. Dar, R. A. et al. Breast cancer detection using deep learning: Datasets, methods, and challenges ahead. Comput. Biol. Med. 149, 106073 (2022).
17. Sharafaddini, A. M., Esfahani, K. K. & Mansouri, N. Deep learning approaches to detect breast cancer: A comprehensive review. Multim. Tools Appl. https://doi.org/10.1007/s11042-024-20011-6 (2024).
18. Bilal, A. et al. Bc-qnet: A quantum-infused elm model for breast cancer diagnosis. Comput. Biol. Med. 175, 108483 (2024).
19. COŞKUN, D. et al. A comparative study of yolo models and a transformer-based yolov5 model for mass detection in mammograms. Turkish J. Electr. Eng. Comput. Sci. 31, 1294–1313 (2023).
20. Zhuang, F. et al. A comprehensive survey on transfer learning. Proc. IEEE 109, 43–76 (2020).
21. Yao, X. et al. Fusion of shallow and deep features from 18f-FDG PET/CT for predicting EGFR-sensitizing mutations in non-small cell lung cancer. Quant. Imaging Med. Surg. 14, 5460 (2024).
22. Pacal, I. A novel swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumours in MRI images. Int. J. Mach. Learn Cybernet. 15(9), 3579–3597 (2024).
23. Hassija, V. et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn. Comput. 16, 45–74 (2024).
24. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. Data Brief 28, 104863. https://doi.org/10.1016/j.dib.2019.104863 (2020).
25. Yap MH, Pons G, Martı´ J, Ganau S, Sentis M, Zwiggelaar R, Davison AK, Marti R (2017) Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform 22(4):1218–1226
26. Gheflati B, Rivaz H (2021) Vision transformer for classification of breast ultrasound images. arXiv preprint arXiv:2110.14731

Received: August 02 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

27. Munteanu, B. -Ş, Murariu, A., Nichitean, M., Pitac, L.-G. & Dioşan, L. Value of original and generated ultrasound data towards training robust classifiers for breast cancer identification. Inform. Syst. Front. 27(1), 75–96 (2024).
28. Pacal, İ. Deep learning approaches for classification of breast cancer in ultrasound (us) images. J. Instit. Sci. Technol. 12, 1917–1927 (2022).
29. Alotaibi, M. et al. Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images. Heliyon 9(11), 22406 (2023).
30. Işık, G. & Paçal, İ. Few-shot classification of ultrasound breast cancer images using meta-learning algorithms. Neural Comput. Appl. 36(20), 12047–12059 (2024).
31. Gheflati, B. & Rivaz, H. Vision transformers for classification of breast ultrasound images. in 2022 44th annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 480–483 (IEEE, 2022).
32. Tagnamas, J., Ramadan, H., Yahyaouy, A. & Tairi, H. Sca-inceptionunext: A lightweight spatial-channel-attention-based network for efficient medical image segmentation. Knowl.-Based Syst. https://doi.org/10.1016/j.knosys.2025.11316 (2025).
33. Sirjani, N. et al. A novel deep learning model for breast lesion classification using ultrasound images: A multicenter data evaluation. Physica Med. 107, 102560 (2023).
34. Zhang, H. et al. Hau-net: Hybrid cnn-transformer for breast ultrasound image segmentation. Biomed. Signal Process. Control 87, 105427 (2024).
35. Singh, A et.al. VISNET: An Efficient Light Weighted Hybrid Model for Early Detection of Breast Tumour in Ultrasound Images using Vision Transformer and Convolutional Neural Networks, Journal of Information Systems Engineering and Management, Vol 9(4s), 2024.
36. Saini, Manali & Hassanzadeh, Sara & Musa, Bushira & Fatemi, Mostafa & Alizad, Azra. (2025). Variational mode directed deep learning framework for breast lesion classification using ultrasound imaging. Scientific Reports. 15. 10.1038/s41598-025-99009-5.
37. Ellis, Edward & Bulpitt, Andrew & Parsa, Nasim & Byrne, Michael & Ali, Sharib. (2025). A Self-Supervised Framework for Improved Generalisability in Ultrasound B-mode Image Segmentation. 10.48550/arXiv.2502.02489.
38. Xie, Xiaoyu & Liu, Pingping & Lang, Yijun & Guo, Zhenjie & Yang, Zhongxi & Zhao, Yuhao. (2024). US-Net: U-shaped network with Convolutional Attention Mechanism for ultrasound medical images. Computers & Graphics. 124. 104054. 10.1016/j.cag.2024.104054.
39. Zhang, Huaikun & Lian, Jing & Ma, Yide. (2025). FET-UNet: Merging CNN and transformer architectures for superior breast ultrasound image segmentation. Physica Medica. 10.1016/j.ejmp.2025.104969.
40. Thomas, Cory & Byra, Michal & Martí, Robert & Yap, Moi Hoon & Zwiggelaar, Reyer. (2023). BUS-Set: A benchmark for quantitative evaluation of breast ultrasound segmentation networks with public datasets. Medical Physics. 50. 10.1002/mp.16287.
41. Bilal, A. et al. Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey wolf optimization. Sci. Rep. 14, 10714 (2024).
42. Zeng, Q. et al. Serum raman spectroscopy combined with convolutional neural network for rapid diagnosis of her2-positive and triple-negative breast cancer. Spectrochim. Acta Part A Mol. Biomol. Spectrosc. 286, 122000 (2023).
43. Ma, X. et al. Detection of breast cancer based on novel porous silicon Bragg reflector surface-enhanced Raman spectroscopy-active structure. Chin. Opt. Lett. 18, 051701 (2020).
44. Vulli, A. et al. Fine-tuned densenet-169 for breast cancer metastasis prediction using fastai and 1-cycle policy. Sensors 22, 2988 (2022).
45. Srinivasu, P. N. et al. Xai-driven catboost multi-layer perceptron neural network for analyzing breast cancer. Sci. Rep. 14, 28674 (2024).
46. Aggarwal R et al (2021) “Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ digital medicine 4:65

Received: August 02 , 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

47. Matsoukas C, Haslum JF, S¨oderberg M, Smith K (2021) “Is it time to replace CNNs with transformers for medical images? “ arXiv: 2108. 09038. Accessed 19 Jun 2025
48. Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam (2019), “Searching for MobileNetV3”, arXiv:1905.02244v5 [cs.CV] 20 Nov 2019
49. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
50. M. Hayat, N. Ahmad, A. Nasir and Z. Ahmad Tariq, "Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification," IEEE Access, vol. 12, pp. 184119-184131, 2024, doi:10.1109/ACCESS. 2024.3503413
51. Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled and Aly Fahmy, “Deep Learning Approaches for Data Augmentation and Classification of Breast Masses using Ultrasound Images” International Journal of Advanced Computer Science and Applications(IJACSA), 10(5), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100579
52. Liyuan Liu, Haoming Jiang y, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, Jiawei Han “On the variance of the adaptive learning rate and beyond” ICLR 2020 arXiv:1908.03265v3 [cs.LG] 17 Apr 2020

Received: August 02 , 2025