Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

VISNET: An Efficient Light Weighted Hybrid Model for Early Detection of Breast Tumour in Ultrasound Images using Vision Transformer and Convolutional Neural Networks

Archana Singh[1*], Surya Prakash Mishra[2], Prateek Singh[3], Anshuka Srivastava[4]
[1,2,3] Department of Computer Science & I.T, SHUATS, Prayagraj, India, [4] Professor, Department of Mechanical Engineering, SHUATS, Prayagraj, India

ARTICLE INFO
Received: 03 Nov 2024
Revised: 12 Dec 2024
Accepted: 22 Dec 2024

ABSTRACT

Breast cancer is a leading cause of mortality among women worldwide, emphasizing the critical need for early and accurate diagnosis. Ultrasound imaging, a widely used diagnostic tool, presents challenges such as noise, shadow, contrast and variability in tumour presentation. Medical image analysis has seen impressive results from deep learning models, especially Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). For breast tumour detection, we provide an effective light-weighted hybrid model VISNET, which combines EfficientNetB0 with a ViT Transformer. In our model, we used transformers to snare long-range relationships, whereas CNNs are used for local feature extraction. To improve representation learning, a feature fusion module based on attention mechanism is infused. To satisfy the claims we've trained and tested our model on Two Datasets. According to experimental findings with respect to seven parameters on the UDIAT, Spain Breast Ultrasound dataset(B), our hybrid model outperforms other State-of-the-art CNNs (ResNet 50, VGG16, EfficientNet B0) and ViT model achieving Accuracy (96.9%), Precision (95.83%), Recall (97.73%), F1-Score (96.67%), Sensitivity (97.74%), Specificity (97.72%) and an AUC (98.67). Whereas, on Baheya, Egypt dataset the accuracy is 97.6%, Precision (97.51%), Recall (96.79%), F1-Score (97.14%), Sensitivity (96.8%), Specificity (96.79%) and AUC (99.82). In practice, our suggested model, VISNET satisfies the claims of light – weightiness as it can run on minimum GPU support and offers a viable way to increase the accuracy of breast tumour categorization as well as yields faster results in comparison to available heavy CNN models.

1. Introduction

1.1 Background

Breast cancer diagnosis relies heavily on imaging techniques. Early detection of cancer cells effectively reduces the mortality rate giving a better sustainability **[1]**. The first clinical test advised by doctors, ultrasonography finds its way in early detection of cancer cells. Ultrasound being a preferred modality due to its affordability, safety among pregnant and old age women, non-invasiveness **[2]**. However, accurate interpretation of ultrasound images is challenging due to artifacts, operator dependence, and subtle differences in benign versus malignant tumours **[3].** With the advent of Deep learning, a subset of machine learning, and advanced computational models such as Convolutional Neural Networks (CNNs) which are capable of handling medical imaging, disease prediction has become more accurate and efficient, leveraging vast amounts of data for early detection and prognosis **[4,5].** The growing availability of datasets like ImageNet and advancements in computational hardware further bolster the applicability of CNNs. They use layers of convolutional filters to extract features from images, making them highly effective for tasks such as tumour detection, organ segmentation, and classification of diseases from X-rays, Ultra Sonography (USG), Mammography, Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Split Open Biopsy (SOB) and Computed Tomography (CT) scans to name a few **[2].**

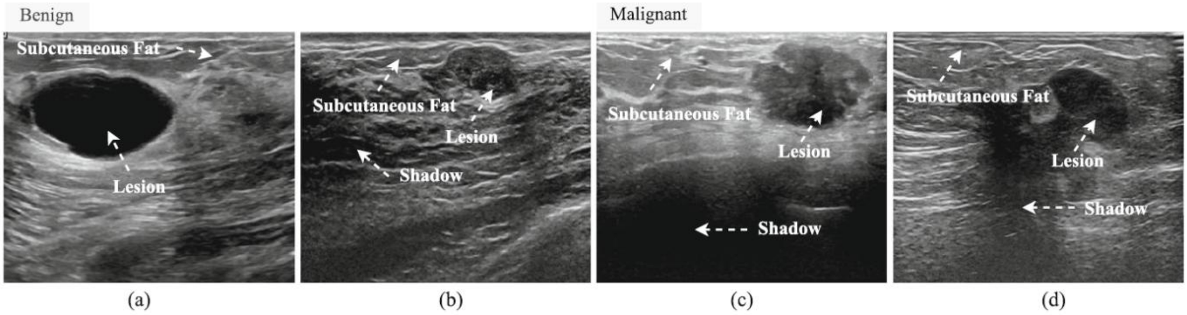


Figure-1 Benign and Malignant USG images of Brest Ultrasound tumours

Figure 1(a) shows image of Benign tumour with subcutaneous fat and lesion with smooth boundary while 1(b) benign image also contains shadow along with fat and tumour lesion. Figure 1(c) and 1(d) shows Malignant tumour image with same features having subcutaneous fat, shadow and lesion but malignancy tumour has irregular boundaries and speculation in comparison with Benign tumours which has smooth boundaries **[33].** The inconsistency in BUS images during predictions may arise as a result of low contrast, shadows and noise speckles.

Although some CNN-based solutions have achieved a classification accuracy close to 90%, they have limitations because CNNs handle the long-range dependencies by expanding the convolution kernel size while slowing down the system speed and enhancing the feature representation. In practice, it is very computationally expensive and limits the generalization ability of the minor resource system. Deep learning models, particularly CNNs, have shown promise in automating and improving diagnostic accuracy **[4].** Vision Transformers can analyse entire image patches simultaneously, improving the detection of subtle cancerous patterns. They can learn hierarchical and complex features, making them robust for histopathology, radiology, and other imaging modalities **[6}.** With large datasets, they can outperform CNNs in feature extraction and classification. Attention maps can highlight regions contributing to cancer detection, aiding in explainable AI for healthcare. However, they need large datasets to avoid overfitting, which is a challenge in medical imaging where labelled data is scarce. Training ViTs requires high-end GPUs due to self-attention mechanisms and they lack standardised architectures as they are still evolving. Recent advancements in Transformer architectures, originally developed for natural language processing, have extended their applicability to vision tasks, offering advantages in snaring long-range dependencies and contextual relationships.

1.2 Motivation

The proposed model, VISNET is based on the basic investigation of cancer diagnosis i.e. Ultrasonography which is real time in nature. Breast cancer involves other investigations also, including mammography and biopsy which can be a painful investigation, PET Scan, MRI may not be suitable for young girls, pregnant and older women because of ionization and radiations. Though ultrasounds are preferred choice but clinically doctors have to write the report which is easy to miss and miswrite and is time consuming **[7]**. Manual medical diagnosis, while important, is time consuming and places a substantial burden on pathologists. In addition, unskilled pathologists who misdiagnose diseased tissues are possible [**38**]. However, considering only ultrasonography also has its limitation like, speckle, noise, shadows, dense breast, probe not being properly used, incorrect alignments, angles while using probe, reporting etc. While CNNs excel in localized feature extraction, they often struggle to capture global contextual information. Transformers, on the other hand, provide a mechanism to analyse relationships across an entire image. This research aims to investigate the combination of these two paradigms to develop a robust light-weighted hybrid model for breast tumour malignancy prediction that can run on minimum GPU support. This model aims to provide better accuracy with less powerful machine, avoiding the use of high computational systems, saving time, cutting the diagnostic costs and making it affordable to both the diagnostic centres and common man.

1.3 Contributions

1. Propose a light – weighted hybrid architecture integrating CNN and Transformer modules for ultrasound image analysis.
2. Develop a pre-processing pipeline tailored to the unique challenges of breast ultrasound images.
3. Validate the proposed model on publicly available datasets, comparing its performance with available state-of-the-art methods.

2. Related Work

2.1 CNNs in Breast Tumour Classification

CNNs have been widely employed for breast cancer detection in medical imaging. Studies demonstrate their efficacy in extracting spatial features, yet their inability to model global context limits their performance, especially in complex datasets. [8] ***Amin M S and Ahn H (February 2023)*** in their research paper "***FabNet****: A Features Agglomeration-Based Convolutional Neural Network for Multiscale Breast Cancer Histopathology Images Classification*" proposed a deep layer CNN model architecture for cancer image classification by closely accumulating the layers together to merge semantic and spatial features. The model can learn fine-to-coarse structural and textural features of multiscale histopathological images by using accretive network architecture that agglomerate hierarchical feature map to acquire classification accuracy. [9] ***Sandler M et.al. (2018)*** in the paper "*MobileNet V2: Inverted Residuals and Linear Bottleneck*" defines a very basic network design to develop a family of an efficient mobile models that increases the state-of-the-art performance of mobile models on several tasks and bench- marks as well as over a spectrum of diverse model sizes. It demonstrates how to develop mobile semantic segmentation models through a shortened variant of DeepLabv3 which we name Mobile DeepLabv3.It is based on an inverted residual structure where the shortcut connections are between the thin bottle- neck layers. The suggested convolutional block has a unique attribute that allows to decouple the network expressiveness (encoded by expansion layers) from its capacity (represented by bottleneck inputs). it is important to reduce non-linearities in the thin layers in order to sustain representational power. [10] ***Basem S Abunasser et. al. (September 2023)*** *in "Convolutional Neural Network for Breast Cancer Detection and Classification using Deep Learning*" proposed a deep learning model BCCNN for detecting and classifying breast cancers into a total of eight classes where 4 classes were of Benign category namely benign adenosis (BA), benign fibroadenoma (BF), benign phyllodes tumour (BPT), benign tubular adenoma (BTA) and 4 were malignant namely malignant ductal carcinoma (MDC), malignant lobular carcinoma (MLC), malignant mucinous carcinoma (MMC), and malignant papillary carcinoma (MPC). They used deep Learning model with additional 5 fine-tuned models consisting of Xception, InceptionV3, VGG16, MobileNet and ResNet50 trained on ImageNet database. The preprocessing and balancing of dataset significantly boosted and helped in improving the accuracy of the fine-tuned pre-trained models and the detection and classification of breast cancer of the suggested model (BCCNN). [11] ***Tiara Lailatul Nikmah, Risma Moulidya Syafei and Devi Nurul Anisa (July 2024)*** in paper "Inception *ResNet v2 for Early Detection of Breast Cancer in Ultrasound Images*" used Inception ResNet v2 and augmented their data using zooming, rescaling and rotating which provided a better recognition of lesions. They used pre-trained Inception ResNet v2 with ImageNet big data with a diverse and large dataset to increase efficiency. the model showed some prediction errors but they were negligible as claimed it might be due to variations in the visual representations of the images**.** [12] ***Kalafi Elham Yousef et. al. (October 2021)*** in their research *"Classification of Breast Cancer Lesions in Ultrasound Images by using Attention Layer and Loss Ensemble in Deep Convolutional Neural Networks"* proposed a new architecture by augmenting attention module in modified VGG16 model to enhance feature discrimination of targeted lesions. Binary cross-entropy and Logarithm of the hyperbolic cosine loss were combined to create ensembled loss function for improving the model discrepancy between classified lesions and their labels. Their classification step involved the proposed

Attention-VGG16 model with CE-LogCosh loss function and compared it with standard pre-trained VGG16 network. [13] ***Yaozhong Luo, Qinghua Huang, and Xuedong Li (November 2021)*** in research paper "*Segmentation Information with Attention Integration for Classification of Breast tumour in Ultrasound Image*" proposed a novel segmentation-to-classification scheme using segmentation-based attention information and augmenting it to deep convolutional neural networks (DCNN). Segmentation network enhances the tumour segmented images and the features are extracted using two DCNNs. Channel attention-based feature extracted is aggregated to improve the performance. They compared their results with existing approaches which includes VGGNet16, VGGNet19, DenseNet121, DenseNet169, ResNet50, Inception V3, Xception, Inception ResNet V2. A ten-fold cross validation was performed and to evaluate the performance of the proposed method confusion matrix with metrics, Accuracy, F1-Score, Sensitivity, Specificity and AUC were considered. [14] ***Michal Byra (2021)*** in "*Breast mass classification with transfer learning based on scaling of deep representations*" proposed a novel transfer learning technique based on deep representation scaling (DRS) layers, which are inserted between the blocks of a pre-trained CNN to enable better flow of information in the network. The aim of this technique was to avoid the hassle in fine tuning when the number of trainable parameters of the pre-trained network is large and the available medical data are scarce. During training only DRS network parameter is updated to adjust the pre-trained CNN. In this approach, the last dense layer (classification layer) of the pre-trained ResNet101 was replaced with a dense layer suitable for the binary classification of breast masses, initialized with random weights. The task involved feature extraction, fine-tuning last block, full fine-tuning and DRS. This approach could simultaneously scale deep representations and modify convolutional filters of the pre-trained model. The issues in their work were they utilized linear spatial and depth-wise operations only although they could have used nonlinear scaling functions also. [15] ***Zheng Dong Fei et.al. (March 2022)*** in their research paper "*Discrimination of Breast Cancer Based on Ultrasound Images and Convolutional Neural Network*" proposed a deep learning algorithm Efficient-Det to assist in diagnosing suspicious breast lesions into benign, malignant or normal. They used Jiangsu Hospital oncology data of 1181 records determined by surgery or biopsy of 487 patients. Efficient-Det was first retrained using an exclusive public breast cancer US dataset with transfer learning techniques. A blind test set consisting of 50 benign, 50 malignant, and 50 normal tissue images was randomly picked from the patients' images as the independent test set to test its searching ability on suspicious tumour regions. [16] ***Moon W K, Lee Y-W, Ke H-H, Lee SH, Huang C-S, Chang R-F (2020)*** in their research paper "*Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks*" stated some issues with traditional CAD software (e.g., it is difficult to create handmade features; it is difficult to check overfitting issues with conventional CAD systems, etc.). The findings concluded that the tumour shape characteristic may enhance the diagnostic effect, that various picture content representations impact the CAD system's prediction performance, and that more image information enhances prediction performance. The drawbacks include ROI area and tumour contour of the B-mode US picture that are clipped according to expert definitions, might lead to varied ROI regions and tumour contours from various operators.

2.2 Transformers in Vision Tasks

Vision Transformers (ViTs) have redefined image classification tasks by treating images as sequences of patches. Their self-attention mechanism enables the capture of long-range dependencies, making them suitable for tasks requiring holistic analysis. [6] ***Behnaz Gheflati and Hassan Rivaz (2021)*** in the paper "*Vision transformer for classification of breast ultrasound images*" used an image dataset of 500*500 pixels with two datasets. They used Adam optimizer for training their model to 30 epochs. The model was fine-tuned, cross entropy loss function was used with self-attention-based mechanism. [17] ***Sudhakar Tummala, Jungeun Kim and Seife dine (2022)*** in the research paper "*BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers*" worked on the concept of Swin transformer (SwinT) with non-overlapping shifted windows. They used openly available BreakHis dataset containing 7909 histopathology images acquired at different zoom factors of 40×, 100×, 200×, and 400× for the two-class classification. [18] ***Kelei He et al (2023)*** in their paper "*Transformers in medical image analysis*" proved the use of Vision Transformers improved the performance of CNN-based classifiers for both natural and medical images. They con-ducted several experiments to compare the performance of a

CNN (i.e., ResNet50) and a ViT (i.e., DEIT-S) using different initialization strategies: (1) randomly initialized weights, (2) transfer learning using ImageNet pretrained weights, and (3) self-supervised pretraining on the target dataset with the same initialization as in (2). They evaluated these methods on the APTOS 2019, ISIC 2019, and CBIS-DDSM datasets.

2.3 Extensive Artificial Intelligence in Health care (XAI)

Artificial Intelligence in Health care is seemingly gaining importance and has preserved its roots in patient care**.** Researchers have developed various models and techniques and have successfully implemented it in various care centres**.** [19] ***Chaddad A, Peng J, Xu J and Bouridane A (January 2023)*** in their research *paper "Survey of Explainable AI Techniques in Healthcare*" provides a survey of techniques used for XAI and explains the concept behind the black box model of deep learning that reveals how the decisions are made. [8] ***Amin M S and Ahn H (February 2023)*** in their research paper "*FabNet: A Features Agglomeration-Based Convolutional Neural Network for Multiscale Breast Cancer Histopathology Images Classification*" proposed a deep layer CNN model architecture for cancer image classification by closely accumulating the layers together to merge semantic and spatial features. [20] ***Tim Hulsen (August 2023)*** in the research paper "*Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare*" talks of a clinical decision support system for aiding doctors not only in the interpretation of reports by radiologists or pathology but also in the analysis of reports.

2.3 Hybrid Architectures

Recent efforts to combine CNNs and Transformers show promising results in general vision tasks. However, their application in medical imaging, particularly ultrasound-based malignancy prediction, remains underexplored. [21] ***Juan Gutierrez-Cardenas, Carrera de Ingeniería de Sistemas, Universidad de Lima and Lima-Perú (2024)*** in their *research "Breast Cancer Classification Through Transfer Learning with Vision Transformer, PCA, and Machine Learning Models"* used pretrained ViT on ImageNet dataset for feature extraction and reduced dimensionality using Principal Component Analysis (PCA). They evaluated mammogram images for Multilayer Perceptron (MLP) and Support Vector Machines (SVM). They used DDSM and INbreast Breast Cancer Dataset for their research a dataset of breast mammography images from the Dataset of Breast Mammography Images with Masses (Huang and Lin, 2020), available at https://data.mendeley.com/datasets/ywsbh3ndr8/2. The result demonstrated Average Accuracy 98.19%, Precision 98.3%, Recall 98.36% and F1-score of 98% for the DDSM dataset with MLP as classifier. The same metrics give us an average of 95.4% for the InBreast dataset.

3. Proposed Methodology

In clinical practice, the integration of deep learning techniques and computer-aided diagnostic methods provides specialists and clinicians with more effective speed, efficiency, cost, and precise diagnostic outcomes [**39**]. Thus, CNNs have a role in medical imaging tasks, including extracting features and classifying tumour lumps [**40**].The proposed model uses a hybrid of Vision Transform (ViT) and EfficientNetB0 architectures to categorize both dataset's images for binary (benign vs. malignant) classifications, using all available images from various datasets. The model EfficientNetB0 and Vision Transformer (ViT) **[22]** designs are the two networks that are combined to generate the hybrid network that forms the basis of the suggested technique.

EfficientNetB0: EfficientNetB0 **[22]** is the baseline model of the EfficientNet family, designed for high accuracy and efficiency in image classification tasks. Here's a summary of its key characteristics:
1.Compound Scaling – EfficientNetB0 introduces a novel compound scaling method, balancing network width, depth, and resolution to optimize performance while maintaining efficiency.
2. Lightweight Architecture – It has approximately 5.3 million parameters, making it significantly smaller and more efficient than traditional CNN models like ResNet.

3. Depthwise Separable Convolutions – Uses Mobile Inverted Bottleneck Convolution (MBConv) layers, reducing computation while maintaining feature extraction quality.
4. Squeeze-and-Excitation (SE) Blocks – Implements SE blocks to enhance channel-wise feature recalibration, improving model accuracy.
5. Swish Activation Function – Uses the Swish activation function, which improves non-linearity and gradient flow, leading to better convergence than ReLU.
6. High Accuracy with Low Computational Cost – Achieves 77.1% Top-1 accuracy on ImageNet with only 0.39 billion FLOPs, making it more efficient than models like ResNet50.
7. Transfer Learning Friendly – Pre-trained on ImageNet, making it suitable for transfer learning applications in medical imaging and other domains.

EfficientNetB0 serves as a strong foundation for deep learning tasks, balancing performance and efficiency effectively.

Vision Transformer (ViT): The self-attention processes of Vision Transformer (ViT) models make them popular models. ViT models are widely used to deliver cutting-edge outcomes [22]. After receiving the 2D images as input, the ViT splits the input into multiple identically sized patches. The standard transformer encoder receives the vector sequence that results from the linear embedding of the patches. In this sequence of vectors, an extra learnable "classification token" is added. The motivation behind the integration of a transformer model into our approach is that it captures the long-range dependencies and understanding of the contextual relationships within the images. This is vital in the examination of histological pictures, where the spatial organization and shape of the cells and tissues play a critical part in the appropriate diagnosis. The EffNetV2-Vit **[37]** model proposed by M. Hayat et.al. worked on a specific dataset called BreakHis. Our suggested hybrid model combines the benefits of both EfficientNetB0 and the transformer-based ViT model encoder for image classification. The following steps were involved in our proposed hybrid model for image classification:

3.1 Data Collection and Preprocessing

We used **two datasets** for proving the accuracy and efficiency of our model. The *first dataset* model used is of ***Baheya Hospital*** *for Early Detection & Treatment of Women's Cancer*, ***Cairo, Egypt [24]*** the data accessibility can be done at https://scholar.cu.edu.eg/?q=afahmy/pages/dataset Baheya Breast Ultrasound Dataset **(BUS),** comprising labelled pretrained weights from ImageNet images with benign (487) images and malignant (210) images. Our *second dataset* includes the ***UDIAT dataset(B),*** Spain. The UDIAT dataset samples were gathered from the UDIAT diagnostic centre of the Parc Tauli Corporation, **Sabadell, Spain [25]**. This dataset consists of a total of 163 ultrasound images with a pretrained weights from ImageNet of breast tumours. In these samples, benign and malignant cases correspond to 109 and 54 breast tumours. These ultrasound images have an average resolution of 760x570 pixels. The dataset can be found here:

The preprocessing pipeline includes the following sequence.

- loads and preprocesses images from a specified directory. For each .png image, it resizes it to a given size and checks for a corresponding mask file (with _mask in the filename).
- Initially the ultra sound images are downsized to consistent 224 × 224-pixel proportions in order to get them ready for training.
- If a mask is found, it normalizes the mask and uses it to highlight the lesion area which is our Region of Interest (ROI) while dimming the background using the image's mean colour.
- The processed images are appended to a list and returned as output.
- This helps prepare lesion-focused input images for deep learning models in medical imaging tasks like breast ultrasound analysis.

An important tactic for artificially growing the dataset is data augmentation **[32]**, which improves the model's accuracy and resilience while avoiding overfitting in image classification.

▪ Then we perform image augmentation algorithms to each magnification level separately to guarantee consistent data representation. using random rotation by 10%, random flipping vertical and horizontal, zoom in-out by 10% and contrast adjusted by 10%.

3.2 Model Architecture

The proposed hybrid architecture with CNN and ViT is shown in Figure - 3. This hybrid model integrates the advantage of both CNN and Vision Transformers along with an enhanced self-attention mechanism for medical image analysis. Below is the description of modules:

I.CNN Module:

In this hybrid architecture we have used EfficientNet B0 (figure-1), the basic model among the EfficientNet family of CNN model. This is best suited for our model as it takes in fewer parameters and provides higher accuracy than other existing models. The image input provided is of 224 × 224 and it takes into account the MBConv (Mobile Inverted Convolutions) blocks with depth wise separable convolutions and an (SE) squeeze and excitation block with pretrained ImageNet weights.

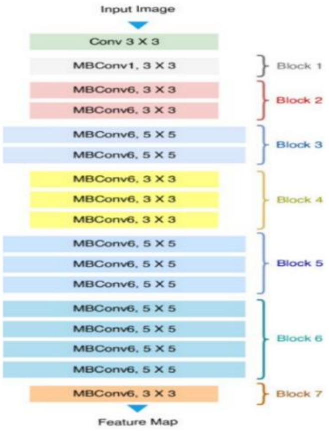


Figure 1: EfficientNet B0 block

Local features are extracted using convolutional layers with layer freezing the top five layers of MBConv blocks. For reducing overfitting and retaining the spatially invariant features the feature map is then passed on to **Global Average Pooling 2D** layer does the averaging in spatial dimensions of height and width of feature map before passing it to final classification layer in CNN. The fully connected dense layer with 1024 units is then used for feature extraction and merged with the ViT features.

II.Transformer Module: This module will process CNN-extracted features as sequences and applies multi-head self-attention to capture global context. For ViT

i.We create a 16×16 Patch
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 concatenate) 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 **[30]**. This encoding tells the model where in the image each patch came from.

iii.Use a 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.

III. **Enhanced Self Attention Module**: This is a feature fusion function (Figure 2(a)) designed to intelligently combine feature vectors from two different models namely Vision Transformer (ViT) and EfficientNetB0 into a single, dynamically weighted representation **[36]**. This is particularly useful in multimodal or hybrid architectures, where different feature sources may capture complementary information. The module learns to focus more on the relevant features for each input sample by computing attention weights.

To begin, both input features are projected into a common feature space of the same dimensionality using Dense layers with ReLU activation. This ensures that both features are comparable in size and scale, and facilitates a smooth fusion process.

Next, the projected features are concatenated and passed through another Dense layer with a Softmax activation function. This generates two attention scores—one for each feature source (Vision Transformer (ViT) and EfficientNetB0 ensuring that their sum is always 1. These scores represent the relative importance of each feature vector for the current input sample, and are split and reshaped to match the dimensions of the feature vectors.

The attention scores are then applied to the respective feature vectors via element-wise multiplication. This effectively scales each feature according to its learned importance. The resulting weighted features are added together to produce a single, fused feature vector that contains information from both inputs, with dynamic emphasis placed on the more relevant one

Finally, a Layer Normalization step is applied to the fused feature vector. This helps stabilize training, improve convergence, and maintain 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.

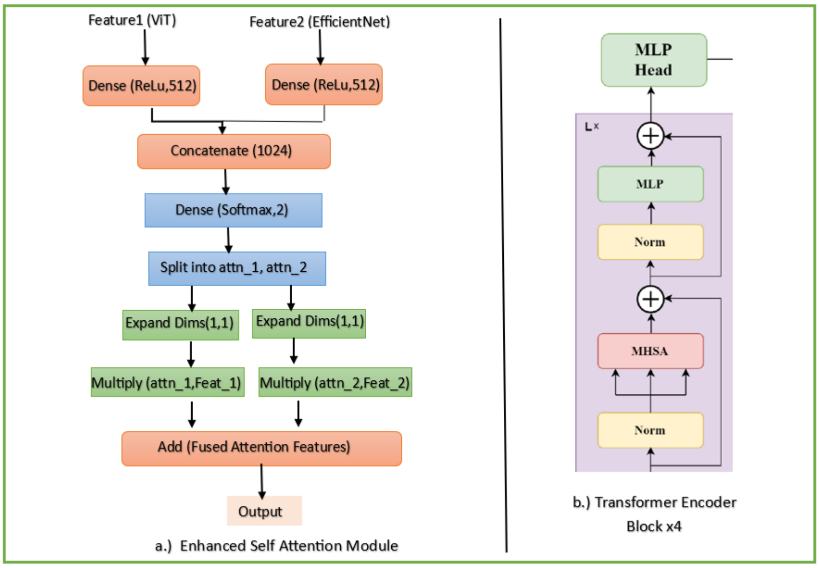


Figure 2:

(a)Self Attention block
(b) transformer encoder X4 block

III.Perform Normalization:

Normalization is performed in both 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.

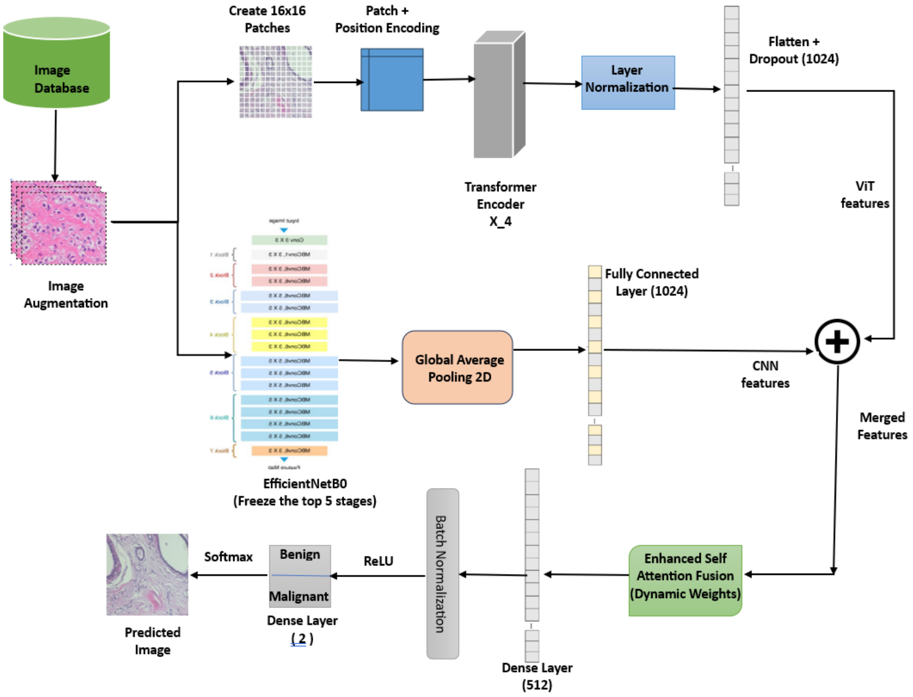


Figure-3: Proposed Hybrid model VISNET extracting features using a self-attention based enhanced hybrid model using ViT and EfficientnetB0 for image classification.

Hybrid ViT + EfficientNet

- **EfficientNet** handles local-level detail by analysing fine-grained texture, boundary sharpness, local contrast, noise. However, irregular, large scale lesions may be a problem in global context.
- **ViT** captures the global structural relationships and models global symmetry, irregular shapes, boundaries and is useful in in cases where global structure is to be considered.
- Finally fusing the features (here, self-attention) to give best extraction feature.

Final Construction:

After successfully fusing the extracted features by self -attention mechanism from both EfficientNet B0 and ViT, a compact (fully connected) dense layer of 512 units from 1024 is generated. The dense layer thus generated of 512 unit is again optimized using ReLU to transform into a 256 one. Further the output image is classified into Benign and Malignant using the Softmax activation function.

3.3 Training Strategy

- Loss Function: Binary cross-entropy.
- Optimizer: Rectified Adam with a Cosine rate scheduler.
- Class Weight: Balancing for imbalanced data in class by assigning class weights

○ **Loss Function: Binary Cross-Entropy**

Binary Cross-Entropy (BCE) 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.

○ Optimizer

To iteratively adjust the model parameters and to minimize the loss function we will use an adaptive optimizer for accelerating the convergence and handling the gradient sparsity effectively. Instead of using Adam optimizer we've used Rectified Adam **[35]** as it Introduces a rectification function that adjusts the learning rate dynamically to prevent extreme variations during early stages and is Well-suited for hybrid architectures with dynamic learning rates.

- **Learning Rate Scheduler**: Reduces learning rate dynamically based on validation loss or epoch. Here we've used Cosine Decay Learning Rate Scheduler as it gradually reduces the learning rate using a **cosine curve** rather than a sharp steep curve. The parameters include initial learning rate, decay steps (epochs*steps per epoch) and final learning rate(alpha*initial LR) which ensures that LR never reaches zero.

Helps stabilize training and improve generalization by slowly reducing LR, especially towards the end of training. Rectified Adam (RAdam) Optimizer with Decoupled Weight Decay helps in rectifies its variance in early training steps which is often more stable.

○ Assign Class Weight

The stability in our model is secured as it works well with the imbalanced data. In our dataset the number of images in Benign Class is double than the Malignant class images. In classification tasks with imbalanced labels (like benign vs malignant), it helps prevent the model from being biased toward the majority class.

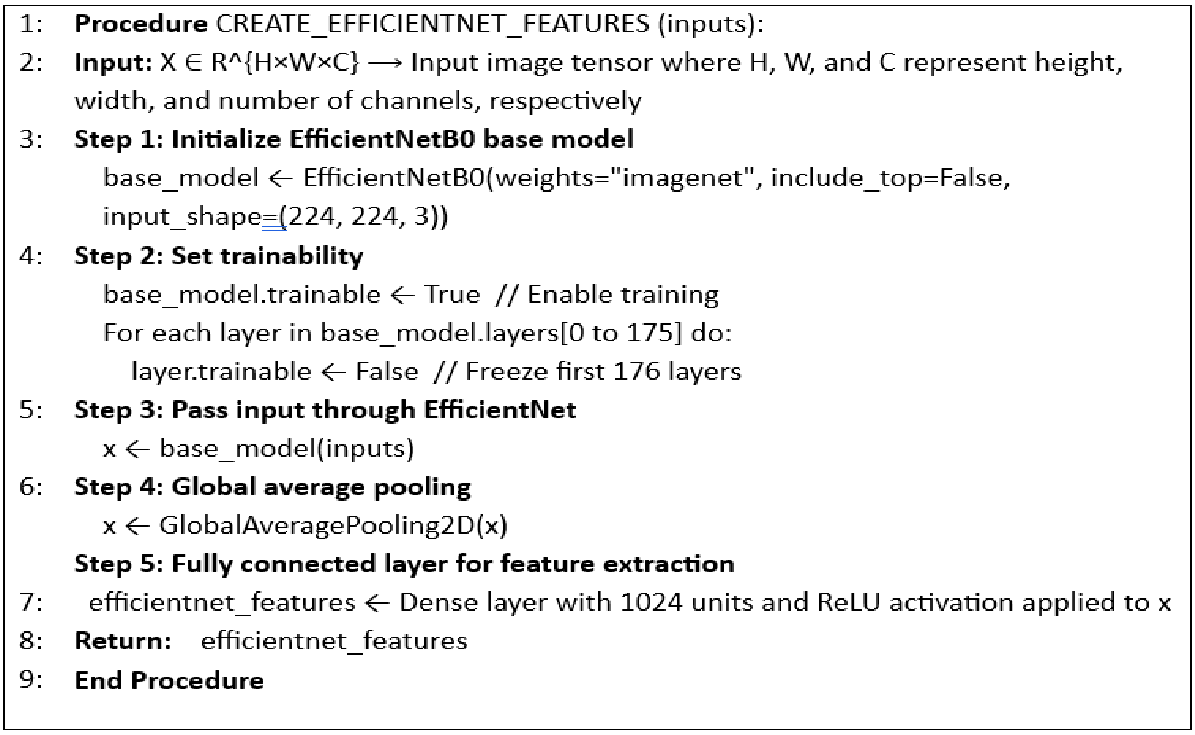
Therefore, we have assigned weight 1.0 to benign class and weight 2.0 to malignant class to give more importance to underrepresented malignant class during training.

Table 1: 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

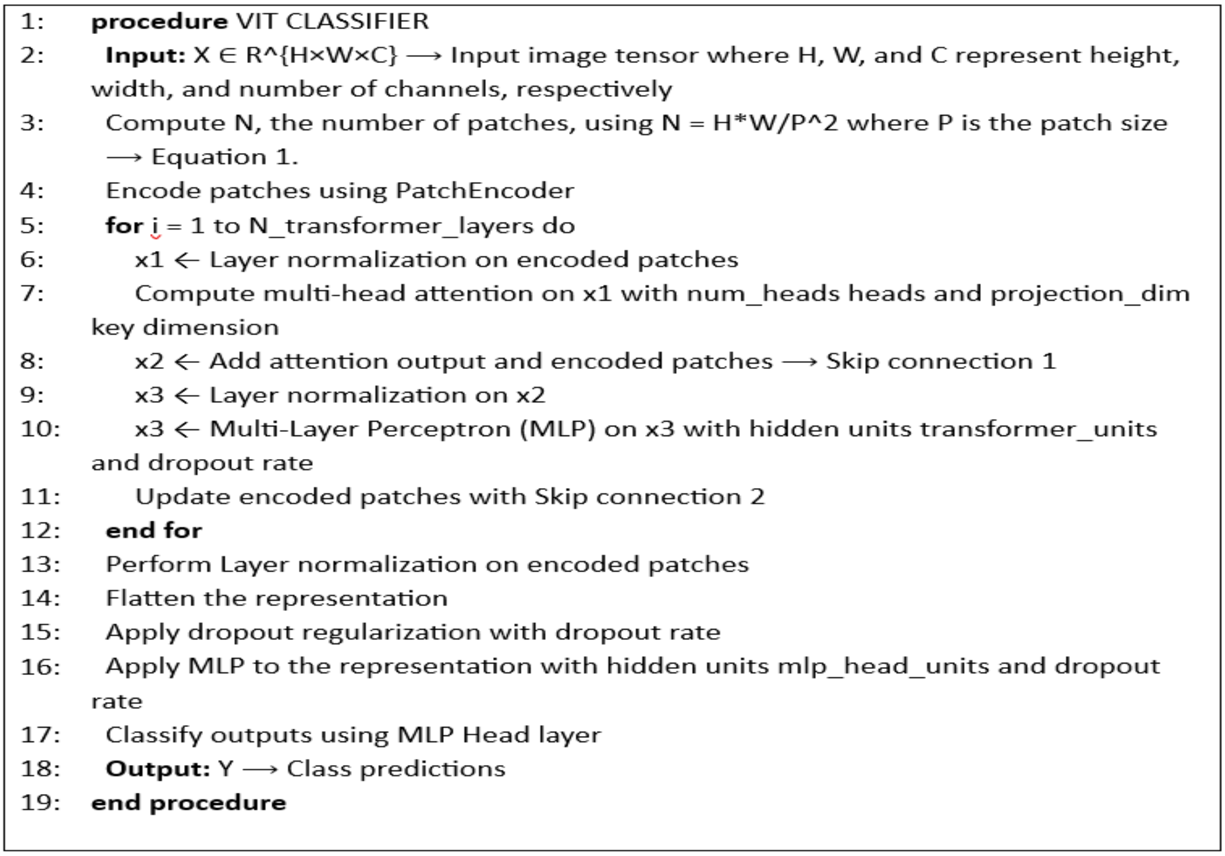
3.4 Proposed Algorithm and Flowchart

This section describes the algorithms of different modules used in our research paper. The first algorithm is of CNN **EfficientNet B0**



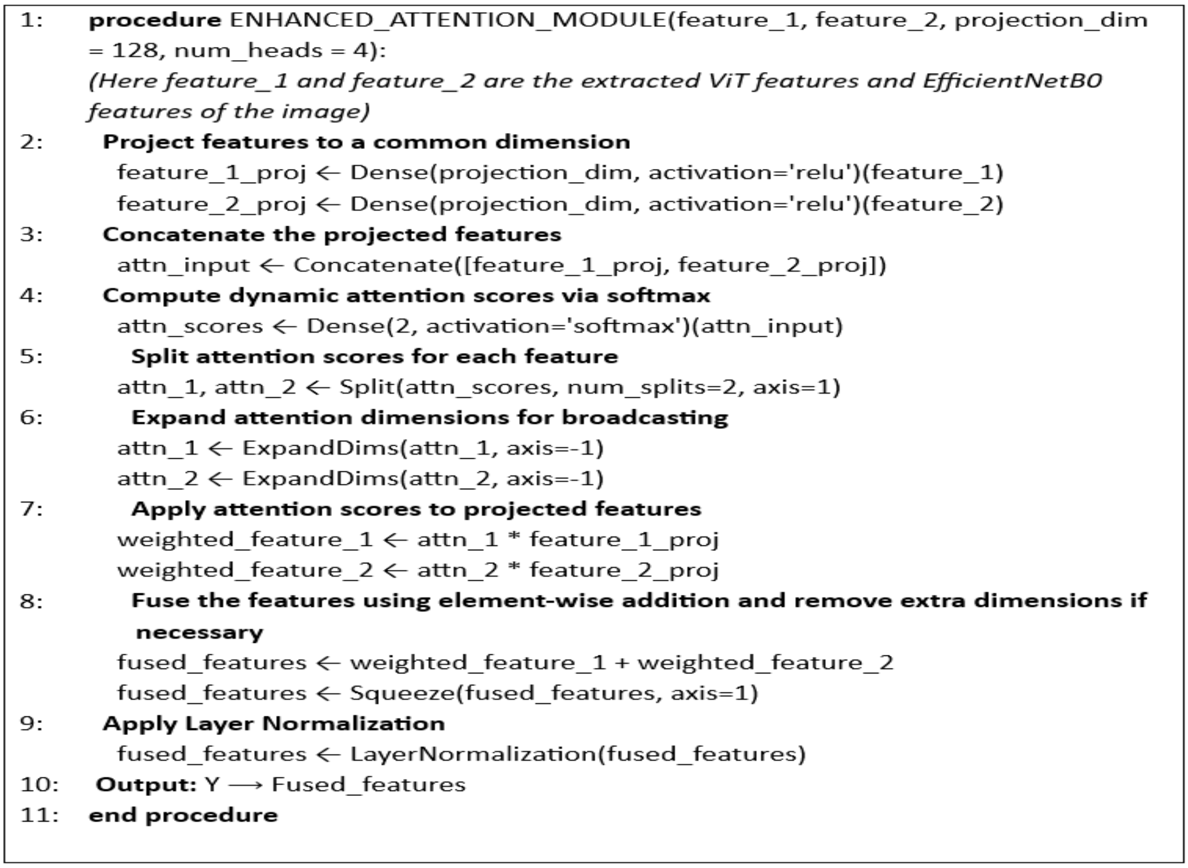
```
Procedure CREATE_EFFICIENTNET_FEATURES (inputs):
Input: X ∈ R^{H×W×C} ⟶ Input image tensor where H, W, and C represent height,
width, and number of channels, respectively
Step 1: Initialize EfficientNetB0 base model
  base_model ← EfficientNetB0(weights="imagenet", include_top=False,
  input_shape=(224, 224, 3))
Step 2: Set trainability
  base_model.trainable ← True  // Enable training
  For each layer in base_model.layers[0 to 175] do:
    layer.trainable ← False  // Freeze first 176 layers
Step 3: Pass input through EfficientNet
  x ← base_model(inputs)
Step 4: Global average pooling
  x ← GlobalAveragePooling2D(x)
Step 5: Fully connected layer for feature extraction
 efficientnet_features ← Dense layer with 1024 units and ReLU activation applied to x
Return:   efficientnet_features
End Procedure
```

The second procedure is of Vision Transformer ViT



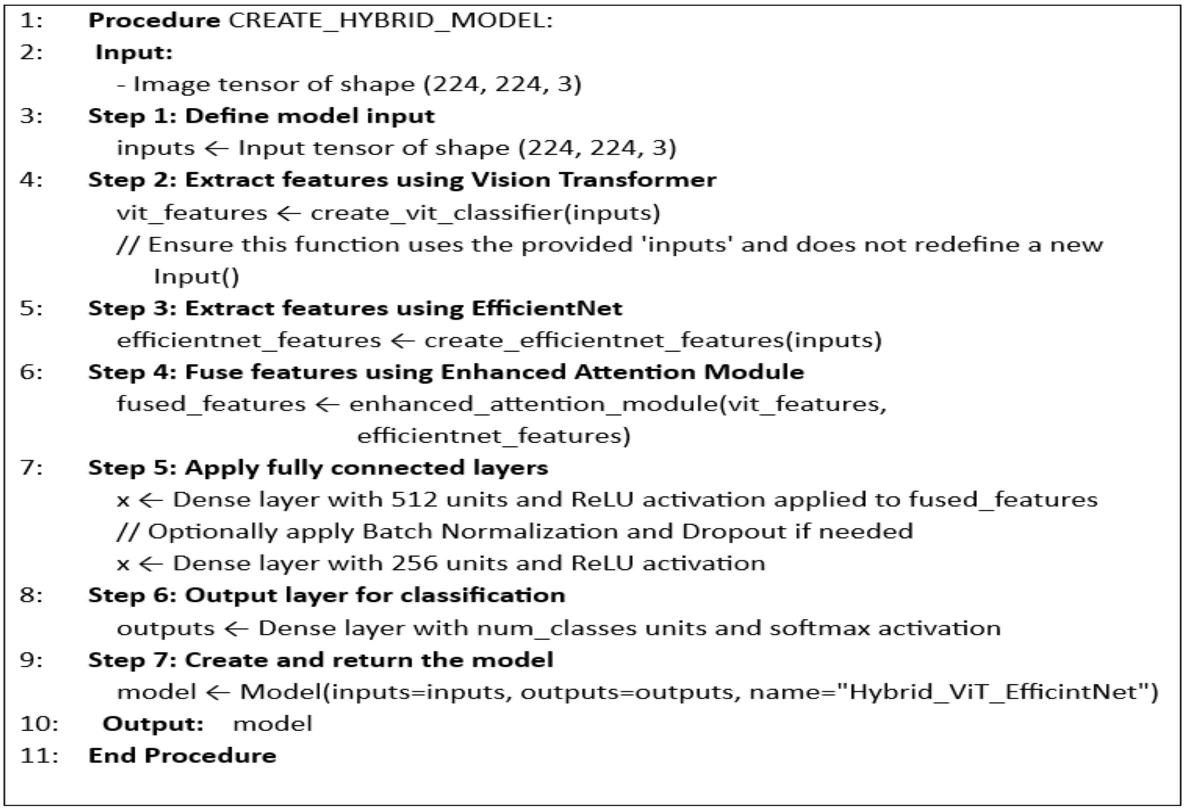
```
procedure VIT CLASSIFIER
 Input: X ∈ R^{H×W×C} ⟶ Input image tensor where H, W, and C represent height,
width, and number of channels, respectively
 Compute N, the number of patches, using N = H*W/P^2 where P is the patch size
 ⟶ Equation 1.
 Encode patches using PatchEncoder
 for i = 1 to N_transformer_layers 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 features from the above algorithms are fed in **self-attention module** below.



```
procedure ENHANCED_ATTENTION_MODULE(feature_1, feature_2, projection_dim = 128, num_heads = 4):
(Here feature_1 and feature_2 are the extracted ViT features and EfficientNetB0 features of the image)
  Project features to a common dimension
    feature_1_proj ← Dense(projection_dim, activation='relu')(feature_1)
    feature_2_proj ← Dense(projection_dim, activation='relu')(feature_2)
  Concatenate the projected features
    attn_input ← Concatenate([feature_1_proj, feature_2_proj])
  Compute dynamic attention scores via softmax
    attn_scores ← Dense(2, activation='softmax')(attn_input)
    Split attention scores for each feature
    attn_1, attn_2 ← Split(attn_scores, num_splits=2, axis=1)
    Expand attention dimensions for broadcasting
    attn_1 ← ExpandDims(attn_1, axis=-1)
    attn_2 ← ExpandDims(attn_2, axis=-1)
    Apply attention scores to projected features
    weighted_feature_1 ← attn_1 * feature_1_proj
    weighted_feature_2 ← attn_2 * feature_2_proj
    Fuse the features using element-wise addition and remove extra dimensions if
     necessary
    fused_features ← weighted_feature_1 + weighted_feature_2
    fused_features ← Squeeze(fused_features, axis=1)
  Apply Layer Normalization
    fused_features ← LayerNormalization(fused_features)
 Output: Y ⟶ Fused_features
end procedure
```

The final **hybrid model of VISNET** is then created using the above discussed procedures.



```
Procedure CREATE_HYBRID_MODEL:
 Input:
   - Image tensor of shape (224, 224, 3)
Step 1: Define model input
   inputs ← Input tensor of shape (224, 224, 3)
Step 2: Extract features using Vision Transformer
   vit_features ← create_vit_classifier(inputs)
   // Ensure this function uses the provided 'inputs' and does not redefine a new
      Input()
Step 3: Extract features using EfficientNet
   efficientnet_features ← create_efficientnet_features(inputs)
Step 4: Fuse features using Enhanced Attention Module
   fused_features ← enhanced_attention_module(vit_features,
                          efficientnet_features)
Step 5: Apply fully connected layers
   x ← Dense layer with 512 units and ReLU activation applied to fused_features
   // Optionally apply Batch Normalization and Dropout if needed
   x ← Dense layer with 256 units and ReLU activation
Step 6: Output layer for classification
   outputs ← Dense layer with num_classes units and softmax activation
Step 7: Create and return the model
   model ← Model(inputs=inputs, outputs=outputs, name="Hybrid_ViT_EfficintNet")
 Output:   model
End Procedure
```

4. Experimental Setup and results

4.1 Experimental setup

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

Hardware to be Used:-

Processor: Intel Core i7 11th gen
16Gb ram
Graphics card: 4 GB Nvidia GeForce RTX3050-Ti

Software to be 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. Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$
2. Precision = $\frac{TP}{TP+FP}$
3. Recall = TP

TP+FN

4. F1-Score = $\frac{TP}{TP + 0.5(FP + FN)}$

Where, TP is the true positive cases where the malignant cases were truly diagnosed, TN is the true negative cases where benign cases were truly diagnosed, FP is for false positive cases where benign cases were incorrectly identified and FN is for false negatives where malignant case were incorrectly identified as benign. The validation will also include the area under curve (AUC) and receiver operating characteristics (ROC).

4.3 Comparative Analysis

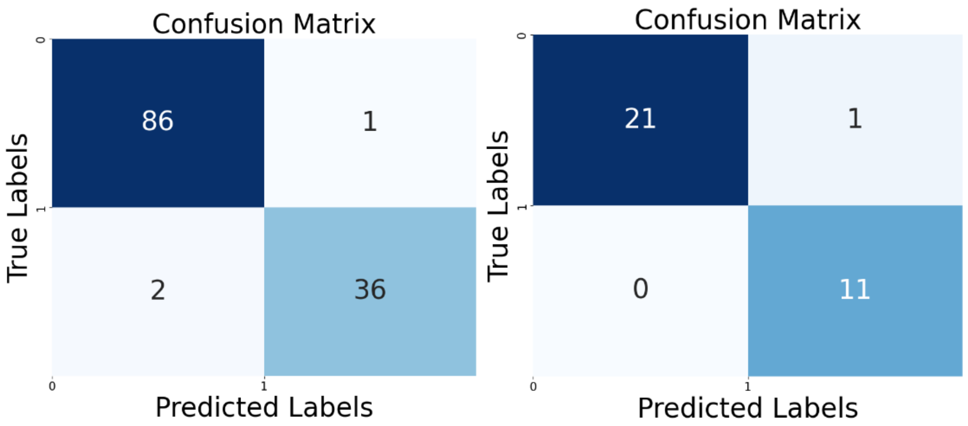
Table 2: Comparative Performance analysis of proposed hybrid model VISNET on BUS and UDIAT dataset respectively

	Baheya Hospital, Egypt Dataset (BUS)					**Parc Tauli, Spain Hospital Dataset (UDIAT)**				
Model	**VISNET (EN-VIT)**	**VIT**	**RESNET-50**	**EFFICIENT NET-B0**	**VGG-16**	**VISNET (EN-VIT)**	**VIT**	**RESNET-50**	**EFFICIENT NET-B0**	**VGG-16**
Test Accuracy	**97.60%**	71.20%	91.20%	96.80%	91.20%	**96.97%**	60.61%	84.85%	87.88%	66 67%
Precision	**97.51%**	64.69%	88.86%	96.21%	89.83%	**95.83%**	50.00%	85.74%	92.30%	33.33%
Sensitivity	96.80%	58.56%	91.46%	**96.20%**	89.21%	**97.74%**	50.10%	79.55%	81.81%	50.10%
Specificity	96.79%	58.55%	91.45%	**96.21%**	89.21%	**97.72%**	50.12%	79.54%	81.82%	50.05%
Recall	96.79%	58.56%	91.45%	**96.20%**	89.23	**97.73%**	50.11%	79.50%	81.82%	51.00%
AUC	**99.82%**	62.67%	98.75%	99.10%.	96.85%	**98.76%**	48.35%	89.66%	97.11%.	23.14%
F1 score	**97.14%**	58.58%	89.95%	96.15%	89.52%	**96.67%**	48.50%	89.52%	84.72%	40.00%
Validation Accuracy	**100.0%**	69.23%	100.00%	99.20%	98.90%	**100.0%**	80.77%	96.15%	88.46%	76.92%
Total Training Time (s)	202.34	**68.06**	580.67	388.53	1238.57	72.13	**35.38**	185.65	127.43	351.96
Avg. Training Time per Epoch (s)	10.11	**03.41**	29.03	19.43	61.93	03.60	**01.52**	09.29	06.38	17.60
Total parameters	5,470,180	926,754	25,634,818	5,363,365	15,242,050	5,470,180	926,754	25,634,818	5,363,365	15,242,050

Table 2 above shows a comparative analysis of both datasets (UDIAT and BUS) with performances highlighted in bolds. In comparison with UDIAT, Baheya dataset gained better accuracy while overall performance came out better in case of UDIAT. For Baheya Hospital dataset in the left, VISNET, gains a maximum test accuracy of 97.6% and maximum validation accuracy of 100% with an AUC of 99.82%. The Precision and Recall is measured to be 97.51% and 96.79%. The F1-Score is 97.14% while, sensitivity and specificity are 96.8% and 96.79%. The Average training time per epoch amounts to 10.11(s) with an epoch size of 20 and batch size of 16 and total parameters of 5.4 million.

For UDIAT dataset in the right, VISNET, gains a maximum test accuracy of 96.9% and maximum validation accuracy of 100% with an AUC of 98.76%. The Precision and Recall is measured to be 95.83% and 97.73%. The F1-Score is 96.67% while, sensitivity and specificity are 97.74% and 97.72%. The Average training time per epoch amounts to 10.11(s) with an epoch size of 20 and batch size of 16 and total parameters of 5.4 million. With reduced parameters and less GPU support VISNET authenticates its acclaimed light weighted model architecture.

Table 3: Confusion Matrix of BUS and UDIAT respectively



The confusion matrix of BUS and UDIAT datasets indicating the model's efficiency and prediction capability are shown in Table 3 above with Class 0 as benign (non -cancerous) and class 1 as malignant (cancerous). The figure3 left confusion matrix of Baheya Hospital test dataset shows 86 images correctly predicted of class 0 (TN), predicted 1, but the true label was 0 (FP), model predicted 0, but the true label was 1 (FN) and correctly predicted 1 when it actually was 1(TP) out of total tested 125 images. The matrix shows a sound solid predicting model with only 3 mistakes out of 125 predictions. In case of FP, 1 benign tumour was incorrectly predicted as malignant which means a patient with a benign (non-cancerous) tumour might be incorrectly told they have cancer. It may cause Anxiety, unnecessary additional testing, possibly unneeded treatment but better safe than sorry in some medical cases. In case of FN, **2** malignant tumours were incorrectly predicted as benign which is dangerous because a patient with a cancerous tumour could be wrongly told everything is fine. It may lead to missed or delayed treatment, which could let the cancer progress definitely something critical to minimize. The confusion matrix on the right for UDIAT test dataset, 21 benign tumours correctly predicted as benign (TN), 1 benign tumour incorrectly predicted as malignant (FP), **No malignant tumours** were missed (FN), 11 malignant tumours correctly predicted as malignant (TP). Out of 33, only 1 benign tumour was misclassified as malignant and no malignant was left out which gives a recall of 100% and is very important for cancer diagnosis.

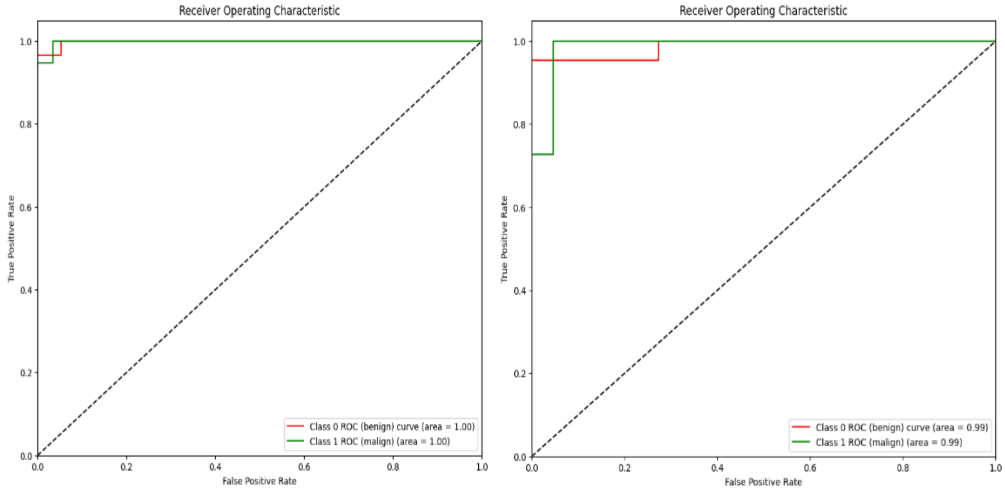


Figure 4: ROC curve of BUS and UDIAT respectively

An ROC curve (Receiver Operating Characteristic curve) **[28]** is a graphical representation used to evaluate the performance of a binary classification model with X-axis indicating False Positive Rate (FPR) = FP / (FP + TN) and Y-axis indicating True Positive Rate (TPR), also called Recall = TP / (TP + FN). Each point on the curve represents a different classification threshold. A perfect model has a curve that hugs the top-left corner (TPR = 1, FPR = 0). A random guess results in a diagonal line from (0,0) to (1,1). The Area Under the Curve (AUC) summarizes the curve into a single number. An AUC = 1, represents a perfect model while, AUC = 0.5 indicates no discrimination (random). The curve shows that our approach for both classes(Class 0 for Benign and Class 1 for Malignant) are predicted precisely.

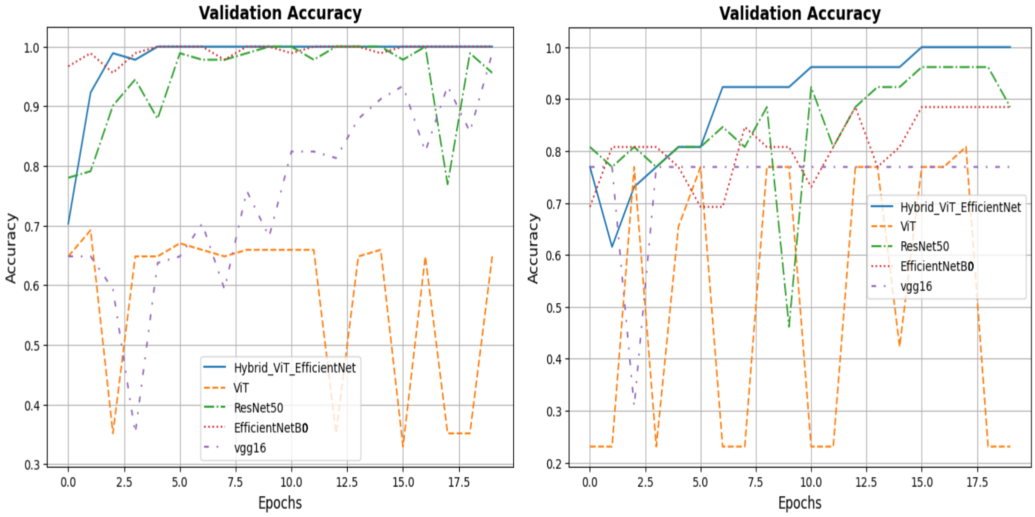


Figure 5: Accuracy curve BUS and UDIAT

Figure 5 above right shows the Validation Accuracy of Baheya (BUS) Dataset for our proposed model VISNET-hybrid of ViT and EfficientNet B0 with ViT, ResNet50 **[26],** EfficientNetB0 and VGG 16 **[27]** across epochs. Our model consistently performs the best, reaching close to perfect accuracy very quickly whereas, Vit struggles the most, with unstable and much lower accuracy. In case of Validation Accuracy Curve on the right of UDIAT dataset , VISNET still performs the best, reaching 1.0 accuracy faster and staying stable. ViT still shows a lot of instability, with big drops down to ~20–30%. ResNet 50 seems better now, but still has a few sharp dips (especially around epochs 9 and 18) while EfficientNetB0 and VGG16 are more stable, but their peak accuracies are lower (~0.8–0.9).

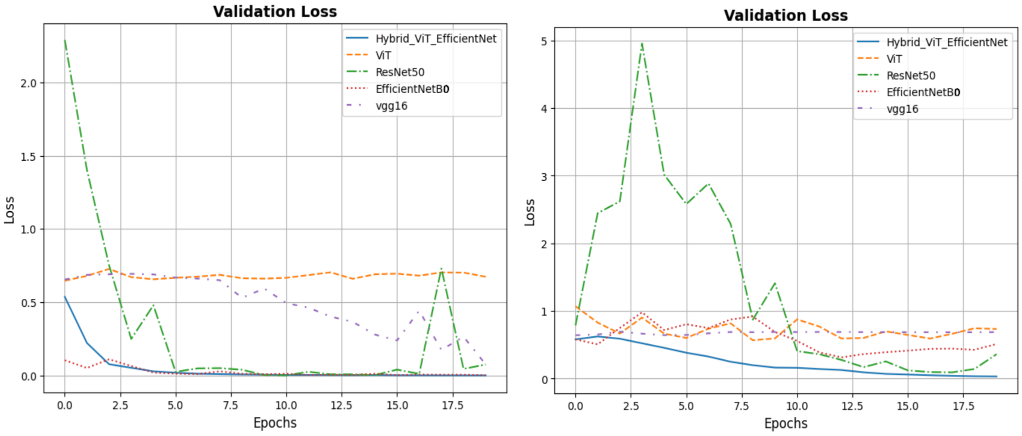


Figure 6: Validation Loss curve of BUS and UDIAT

In Figure 6 the validation loss curve of Baheya (BUS) dataset is plotted depicting that our model VISNET achieved the lowest losses (close to zero), very quickly and stay there along with EfficientNetB0. ViT's validation loss remains quite high (~0.65) and doesn't really improve. ResNet50 initially has a very high loss (>2.0!) but then quickly drops down to almost zero, although with some spikes while, VGG16 steadily decreases but not as low as the top models.
The validation loss curve on the left shows the proposed model VISNET, again has the best performance as the loss smoothly decreases over time and stays very low (~0.1 or less). ResNet50 this time has a wild spike — loss shoots up to ~5.0 around epoch 3 before finally stabilizing much later. ViT, EfficientNetB0, and vgg16 hover around a validation loss of ~0.7–0.9 without much major improvement.

5. Conclusion and Future Work

This research demonstrates the effectiveness of integrating CNNs and Transformers for breast tumour malignancy prediction in ultrasound images. The proposed model sets a new benchmark for accuracy and interpretability, paving the way for advanced AI-driven diagnostic tools in oncology. Our hybrid model VISNET, successfully captures the advantage of both the CNN and Vit by incorporating and infusing both local and global features making it a robust model to variations in ultrasound tumour morphology. Being light-weighted it can run on low resources with less GPU support. Thus, the results are delivered much faster as compared to other heavy accurate models. VISNET, tested on Baheya Hospital Dataset (BUS) and UDIAT dataset clearly outperforms other state-of-the-art models discussed in the paper. The plots ROC curve, Accuracy curve and Validation Loss curve together really confirms the claims of our hybrid model VISNET with the best accuracy (97.6%), Precision (97.51%), Recall (97.73%), F1-Score (97.71%), AUC (99.82), and lowest loss in performance over other state of the art models considered in our research across all the metrices. The claims of being light-weighted is indicated with no. of parameters 5.4 million with a low GPU support on which the model was trained and tested and validation confirmed. We've seen that in comparison with other modelled vision transformers like Visformer **[23]**, the no. of parameters were quite high, Visformer-S having 40.2(M) params and VisformerV2-S with 23.6 (M) parameters. However, the computational complexity of Transformers necessitates hardware resources that may limit deployment in resource-constrained settings. However, in future it may be resolved with further enhancements in the model.

Future Work

Our Future research will explore to make further enhancements in our proposed model to make it more light-weighted and explore other transformer variants. The enhanced model can be incorporated to work with multi-modal imaging datasets. Furthermore, the same model can be trained for other ultrasound images of tumours found in different sites of lungs, thyroid, liver and more.

References

[1] Wang L (2017) Early diagnosis of breast cancer. Sensors 17(7):1572

[2] Duffy SW, Taba´r L, Chen H-H, Holmqvist M, Yen M-F, Abdsalah S, Epstein B, Frodis E, jungberg E, Hedborg-Melander C et al (2002) The impact of organized mammography service screening on breast carcinoma mortality in seven Swedish counties: a collaborative evaluation. Cancer Interdiscip Int J Am Cancer Soc 95(3):458–469

[3] Asiedu MN, Benjamin AR, Singh VK, Wang S, Wu K, Samir AE, Kumar VS (2022) A generative adversarial network for ultrasound signal enhancement by transforming low-voltage beamformed radio frequency data to high-voltage data. In: Medical Imaging 2022: Ultrasonic Imaging and Tomography, vol12038, pp 246–254. SPIE

[4] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sa´nchez CI (2017) A survey on deep learning in medical image analysis. MedImage Anal 42:60–88

[5] Shamshad F, Khan S, Zamir SW, Khan MH. Hayat M, Khan FS, Fu H (2022) Transformers in medical imaging: a survey. arXiv preprint arXiv:2201.09873

[6] Gheflati B, Rivaz H (2021) Vision transformer for classification of breast ultrasound images. arXiv preprint arXiv:2110.14731

[7] Ge S, Ye Q, Xie W, Sun D, Zhang H, Zhou X, Yuan K (2021) AI assisted method for efficiently generating breast ultrasound screening reports. arXiv preprint arXiv:2107.13431

[8] Amin MS, Ahn H (2023) FabNet: A features agglomeration-based convolutional neural network for multiscale breast cancer histopathology images classification. Cancers 15:1013

[9] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520. https://doi.org/10.48550/arXiv.1801.04381

[10] Abunasser BS, Al-Hiealy MRJ, Zaqout IS, Abu-Naser SS (2023) Con- volution neural network for breast cancer detection and classification using deep learning. Asian Pac J Cancer Preven: APJCP 24:531

[11] Tiara Lailatul Nikmah, Risma Moulidya Syafeina Devi Nurul Anisa "Inception ResNet v2 for Early Detection of Breast Cancer in Ultrasound Images" in Journal of Information System Exploration and Research (JOISER)Vol. 2, No. 2, July 2024, pp. 93-102, p-ISSN 2964-1160, e-ISSN 2963-6361

[12] Kalafi EY, Jodeiri A, Setarehdan SK, Lin NW, Rahmat K, Taib NA, Ganggayah MD, Dhillon SK (2021) Classification of breast cancer lesions in ultrasound images by using attention layer and loss ensemble in deep convolutional neural networks. Diagnostics 11(10):1859

[13] Luo Y, Huang Q, Li X (2022) Segmentation information with attention integration for classification of breast tumour in ultra- sound image. Pattern Recogn 124:108427

[14] Byra M (2021) Breast mass classification with transfer learning based on scaling of deep representations. Biomed Signal Process Control 69:102828. ttps://doi.org/10.1016/j.bspc.2021.102828

[15] Du R, Chen Y, Li T, Shi L, Fei Z, Li Y (2022) Discrimination of breast cancer based on ultrasound images and convolutional neural network. J Oncol 2022:7733583

[16] Moon WK, Lee Y-W, Ke H-H, Lee SH, Huang C-S, Chang R-F (2020) Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput Methods Programs Biomed 190:105361

[17] Tummala S, Kim J, Kadry S (2022) Breast-net: multi-class classification of breast cancer from histopathological images using ensemble of swin transformers. Mathematics 10:4109

[18] PannelKelei He , Chen Gan , Zhuoyuan Li Islem Rekik , Zihao Yin , Wen Ji Yang Gao, Qian Wang, Junfe ng Zhang, Dinggang Shen (2023) Transformers in medical image analysis. Intell Med 3:59–7819.

[19] Chaddad A, Peng J, Xu J, Bouridane A (2023) Survey of explainable ai techniques in healthcare. Sensors 23:634

[20] Tim Hulsen (2023) Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare *AI* 2023, *4*(3), 652-666; **https://doi.org/10.3390/ai4030034**

[21] Juan Gutierrez-Cardenas Carrera de Ingeniería de Sistemas, Universidad de Lima, Lima-Perú (2014) Breast Cancer Classification Through Transfer Learning with Vision Transformer, PCA, and Machine Learning Models (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 15, No. 4, 2024

[22] 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

[23] image recognition at scale. arXiv preprint arXiv:2010.11929

[24] Chen Z, Xie L, Niu J, Liu X, Wei L, Tian Q (2021) Visformer: the vision-friendly transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 589–598

[25] Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A (2020) Dataset of breast ultrasound images. Data Brief 28:104863{Baheya hospital dataset

[26] 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

[27] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on "computer vision and pattern recognition", pp 770–778

[28] Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556

[29] Fawcett T (2006) An introduction to roc analysis. Pattern Recogn Lett 27(8):861–874

[30] Jiang, Y., et al. "Breast Ultrasound Image Classification Using Deep Convolutional Neural Networks." Journal of Medical Imaging, 2020.

[31] Vaswani, A., et al. "Attention is All You Need." Advances in Neural Information Processing Systems, 2017.

[32] **31**. Chigozie Enyinna Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall.

[33] "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning" arXiv:1811.03378v1 [cs.LG] 8 Nov 2018

[34] 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

[35] Jiang, Y., et al. "Breast Ultrasound Image Classification Using Deep Convolutional Neural Networks." Journal of Medical Imaging, 2020

[36] Touvron, H., et al. "Training data-efficient image transformers & distillation through attention." Proceedings of the International Conference on Machine Learning, 2021.

[37] 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

[38] Vivek Kumar Singh, Ehab Mahmoud Mohamed, Mohamed Abdel-Nasser "Aggregating efficient transformer and CNN networks using learnable fuzzy measure for breast tumour malignancy prediction in ultrasound images" Neural Computing and Applications (2024) 36:5889–5905 https://doi.org/10.1007/s00521-023-09363-6

[39] 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

[40] Sohns C, Angic BC, Sossalla S, Konietschke F, Obenauer S (2010) "Cad in full-field digital mammography—influence of reader experience and application of cad on interpretation of time." Clin Imaging 34:418–424

[41] Aggarwal R et al (2021) "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ digital medicine 4:65

[42] 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 2023