Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Fusion-Based CNN-ViT Model for Breast Cancer Classification and Detection Using Digital Mammograms

Nelofar Bashir[1]*, S.P Singh[2]

[1,2]*Nims Institute of Computing, Artificial Intelligence and Machine Learning Nims University Rajasthan, Jaipur, India*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 12 Feb 2025

Accepted: 27 Feb 2025

ABSTRACT

Although it can affect men as well, breast cancer is currently the most common type of cancer that mostly affects women. It manifests when aberrant cells in breast tissue proliferate quickly and develop into tumors. Physicians use the mammogram procedure to examine the breast in order to diagnose early cancer. These mammograms fall into one of two cat- egories: benign or malignant.By merging Convolution Neural Networks (CNNs) with Vision Transformers (ViTs), this study attempts to improve classification accuracy and solve the variability and potential oversight in radiologists' manual mammography interpretations.. ViTs capture the global context but need a lot of data and processing power, while CNN is an effective image classification system that leverages hierarchical feature extraction. In this study, we trained a CNN+ViT model using CLAHE- enhanced mammography pictures from Kaggle. For comparative study, we have also employed a few pre-trained models, including DenseNet, In- ception, SE Resnet, and XceptionNet. With an accuracy of 90.1%, the CNN+ViT model demonstrated strong performance. XceptionNet's perfect accuracy could be a sign of overfitting

Keywords: Breast cancer, mammogram screening, CNN, ViT.

1.INTRODUCTION

Breast cancer is projected to be the leading cause of death among women world- wide by 2024. Breast cancer symptoms occur when abnormally growing breast cells create a lump or tumor. These cells grow faster than healthy ones. [7]. Breast cancer exhibits an increase in the percentage of these cells. According to the Breast Cancer Care (BCC) report, 42% of NHS trusts lack these ser- vices. A lack of experienced breast cancer specialists has led to a decrease in patient survival rates. [10].According to a World Health Organization (WHO) research, only 8% of affected patients are identified, 6% have died, and the re- mainder are undiagnosed, with females accounting for the majority [17, 2, 25]. Breast cancer has a mortality rate of 627,000 fatalities, with India accounting for 14.0% of total cases commencing in the early thirties [23, 27].Early detec- tion and diagnosis are crucial for improving the long-term survival incidence of breast tumor malignancy. Medical imaging techniques such as ultrasonography and mammography are commonly used to identify breast cancer[32].Relative to other mammography procedures, this technique is less unpleasant and non- invasive, capturing detailed x-ray pictures [21]. Mammograms detect abnor- malities, which are classed as benign or malignant[3].Benign tumors are often oval or spherical with smooth edges, but malignant tumors are cancerous and have uneven, wavy, or vague edges [7, 16].Despite its effectiveness in screen- ing, mammography interpretation remains challenging due to tissue density changes, making it difficult to discern between benign and malignant tumors in early stages.To address this issue, modern technology can help radiologists provide more accurate assessments, reduce errors in diagnosis, and improve self- exam practices for women [6]. Recent years have seen the emergence of exper- tized systems for breast cancer diagnosis using mammography image datasets. These systems can be classified into two types: conventional image processing[20] and deep learning-based diagnosis[13][31].Our related works apply several strategies to solve the mammography problem. The BC2EffINetNetRF frame- work combines EfficientNet-b0 with HRLG contrast enhancement, while other research use EfficientNet-B4 and Equilibrium-Jaya algorithms for feature selec- tion and transfer learning

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[10]. YOLO and U-Net models are commonly used for lesion segmentation [32], whereas CNN architectures like VGG-11, ResNet, and DenseNet are optimized using fuzzy ensemble techniques [26, 22].Research suggests that ViTs can improve image classification by capturing global con- text. Pre-training using big datasets [5, 24, 11] Despite their gains, there are limitations like as data diversity and scarcity, computational complexity, and overfitting when using CNN or ViTs independently [12].Our goal is to use the CNN+ViT model to develop an effective model that addresses some of the is- sues raised above.[1, 14].By combining the CNN and ViT models, we hope to increase classification and detection efficiency with a CNN+ViT model. so that we may take advantage of both models' capabilities, including ViT's capacity to capture context globally and CNN's ability to extract features locally[30].Using this CNN+ViT model, we want to create diagnostic tools that are more precise. To extract clarity features from the dataset, we employ CLACHE enhanced mammography images. In contrast to earlier research, our CNN+ViT model can extract balanced features and produce high classification accuracy without overfitting. Lastly, opening the door for a reliable and understandable model that can identify both benign and malignant tumors in order to diagnose breast cancer. To provide a concise overview of our findings, we divided this publi- cation into five sections, cutting off abstracts, keywords, and references. The problem is introduced in Section 1 along with some statistical analysis and back- ground information. A literature review and discussions of current technologies, including their successes and shortcomings, are covered in Section 2. The in- formation flow for the suggested solution is the main focus of Section 3. The manuscript is concluded in Sections 4 and 5, which provide visual confirmations of the findings.

2.RELATED WORKS

These articles demonstrate the most recent developments in the different deep learning models and methodologies for the identification and categorization of breast cancer from mammography pictures. To effectively categorize benign and malignant tumors, researchers created frameworks utilizing the EfficientNet ar- chitecture, Equilibrium-Jaya, Mask R-CNN, and U-Net. To maximize feature extraction, some of them combined the YOLO models with Vision Transform- ers and Eigen-CAM saliency maps. They work on fuzzy CNN ensembles, such as VGG-11 and Inception v3, to enhance the model's performance and effi- ciency.Hassan et al. developed an automated CAD system in the YOLO-Based CAD Framework with ViT Transformer for Breast Mass Detection and Classifi- cation in CESM and FFDM Images for the detection and classification of mass in both images. [7]This CAD framework has achieved noteworthy automa- tion in CESM mass detection and classification tasks by utilizing YOLOv4 for mass detection and a ViT transformer for classification. On both the INbreast and CESM datasets, the model achieved classification accuracy rates of 95.65% and 97.61%, respectively. Jabeen et al.'s BC2NetRF framework for classifying breast cancer from mammography images [10]obtained 95.4% and 99.7% accu- racy on the CBIS-DDSM and INbreast datasets, respectively. They employed HRLG Contrast Enhancement to improve image clarity and the EfficientNet-b0 model for feature extraction. To enhance performance, they incorporated the Equilibrium-Jaya Controlled Regula Falsi algorithm. In the future, they hope to deal with bigger and more varied clinical datasets.A model for classifying breast tumors using improved transfer learning and chaotic map-based optimization strategies for selection is presented by Chakravarthy et al. [2]. They used the EfficientNet-B4 architecture to classify breast cancer and achieved 98.4% and 96.1% accuracy on the INbreast and CBIS-DDSM datasets, respectively. They hope to use explainable AI in conjunction with multimodal data fusion in the future.With the primary goal of improving mammography accuracy for breast cancer screening, Wang [32]suggested a model for breast cancer detection that combines deep learning frameworks with mammograms. With an accuracy of 0.88%, their study demonstrates the use of Mask R-CNN and U-Net models for breast lesion segmentation to successfully classify benign and malignant can- cers. By maximizing computing efficiency, scientists hope to increase accuracy in their upcoming studies.Prinzi et al. [21]investigated the YOLO architecture and Eigen-CAM saliency maps for the diagnosis of breast cancer in mammog- raphy. Their strategy is centered on offering an accurately explained method that can assist doctors in making judgments regarding cancer screening pro- grams. They used a tiny YOLOv5 model and obtained a map of 0.621. They hope to work on bigger datasets with simpler models in the future.Mohammed and Ekmekci [13] provide a breast cancer diagnosis system that combines the Flattens threshold framework with YOLO-

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

based multiscale parallel CNN. They use specialized CNN components like PFES, DCB, and IB to improve feature extraction. They achieved better results with 98.7% accuracy and 91.1% map. In order to improve generalization, they wish to investigate domain adaptation and transfer learning.In order to diagnose breast cancer, Chakravarthy et al.[26] experimented using transfer learning in conjunction with a fuzzy orchestra approach in deep learning. They employed the fuzzy encapsulation technique, which ranks CNN models VGG-11, ResNet50, and Inception v3 using a modified Gompertz function. The accuracy it attained was 98.98%.They wish to enhance segmentation in the future to guarantee strong success.Vision Transformer (ViT) is introduced by Dosovitskiy [5] and Google researchers Ashish Vaswani and Noam Shazeer, who investigate the potential of transformers in picture clas- sifications without CNNs. ViT fine-tunes the tiny datasets after pre-training on large datasets to attain state-of-the-art accuracy across numerous models. They hope to collaborate with CNN on ViT in their upcoming studies.In order to develop a hybrid model for the detection and classification of breast cancer, Zarif et al. [33] focused on pretrained models. Reducing manula mistakes in image analysis is their goal. To address the issue of manula mistakes, they proposed a hybrid model that combines CNN and EfficientNetV2B3, achieving 96.3% accuracy. Tan et al. [29] present the RCM-YOLO network combined with residual asymmetric dilated convolution, adaptive selection mechanisms, and cross-layer attention to maximize detection performance and reach a map of 90.34% in order to address the issue of misdiagnosis rates in breast cancer screening. In order to produce better findings, they hope to develop the model and raise accuracy in the future to detect the entire mass's marginal area.In order to diagnose breast cancer, Nadkarni and Noronha [15]worked with con- volutional neural networks utilizing ensemble learning. To improve classifica- tion accuracy, their model incorporates the Hungarian optimization algorithm, which yielded a 95.7% accuracy rate.O'Shea [18]discussed An Introduction to CNN's Architecture in their study. Convolutional and pooling layers for fea- ture extraction in classification issues are used to explain CNN's image analysis capabilities. They point out that CNNs can assist in reducing model complex- ity without sacrificing performance and highlight problems like computational intensity. As a result, they recommend more research to strengthen CNN resis- tance in picture applications. We can infer from these results that future work will concentrate on combining CNN and ViTs to improve the feature extraction procedure. Therefore, we concentrated on hybrid models—a combination of CNNs and ViTs—to improve classification performance and interpretability in breast cancer diagnostics by utilizing both local and global feature extraction.

3.METHODOLOGY

A hybrid architecture method that supports CNN and ViT is part of our suggested solution.This method strikes a balance between the model's advan- tages—CNNs' ability to extract local features and ViTs' ability to capture global dependencies. Additionally, we have performed a comparison study using a few different architectures, such as CNN, Xception, DenseNet, InceptionV3, and SEResNet. This thoroughly assesses the models' performance using a range of feature extraction methodologies and architectures.

3.1 Dataset

The dataset used to classify breast cancer was extracted from the Mammo- gram CLAHE dataset on Kaggle. It comprises 10,000 annotated images that are equally split into two classes: benign and malignant, with 5,000 samples utilized in each class. Additionally, Figure 1depicts benign and malignant mam- mography pictures, providing a good indication of how the two types appear. Each image captures a tissue sample's appearance in tissue analysis and aids in distinguishing between benign and malignant cases.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

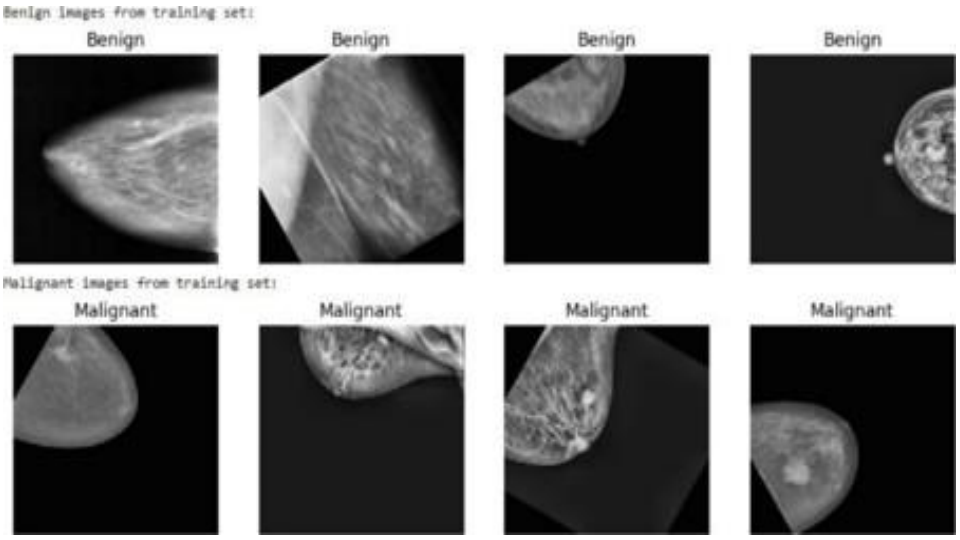

Figure 1: Benign and Malignant Mammogram Images

3.2 INTRODUCTION TO CNN AND ViT MODELS

CNN is the most widely used algorithm in deep learning. To improve perfor- mance, CNN can be mixed with other models to create a hybrid model.

In essence, CNNs tackle challenging image-based pattern recognition tasks, and because of their precise yet uncomplicated structure, they provide a more ap- proachable way to introduce ANNs to a newbie. These CNNs are a subset of deep learning models created especially for image-based applications. By ap- plying many layers of convolutional filters to the input, they enable the learning of significant features in images, including edges, textures, forms, and even in- tricate patterns. Put differently, CNNs make it possible to identify spatial local correlations in images, which makes them particularly helpful for tasks like ob- ject identification, image categorization, and analysis.CNNs can gradually learn higher-level features while reducing the spatial dimensions of data without los- ing crucial information. This is demonstrated in the training of the Custom CNN, which yielded an accuracy of 0.85 and a loss of 0.33. It is evident that the single CNN model overfitted and had really poor accuracy.Convolution Neural Networks (CNN) have been replaced by Vision Transformer (ViT). Originally intended for Natural Language Processing, Vision Transformer (ViT) uses trans- former architecture to identify images. ViT splits the image into small, fixed-size patches (16×16), flattens them, then processes them with a transformer to eliminate the reliance on convolutional layers for spatial feature extraction.To obtain the spatial relationships, positional embeddings are added after each patch in the image is handled as a token.The model then employs self-attention processes, which are more effective than CNN, across all patches, allowing us to capture global context and long-range dependencies. ViT is a more efficient model for image recognition tasks than CNN since it uses less computing power while retaining a greater level of accuracy.

3.3 PROPOSED WORK HYBRID MODEL

Custom Convolutional Neural Networks (CNN) and Vision Transformer (ViT- B16) are integrated in the suggested hybrid model. By utilizing the advantages of both designs, CNN and ViT produce effective performance in the image recognition domain. By combining the two models, ViT was able to extract the wide range relationships, while CNN was able to effectively extract the features in the early stages that encoded the fine-grained morphological patterns as shown in Figure 2. This combination improves computational efficiency and robustness without causing the model to overfit or underfit.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

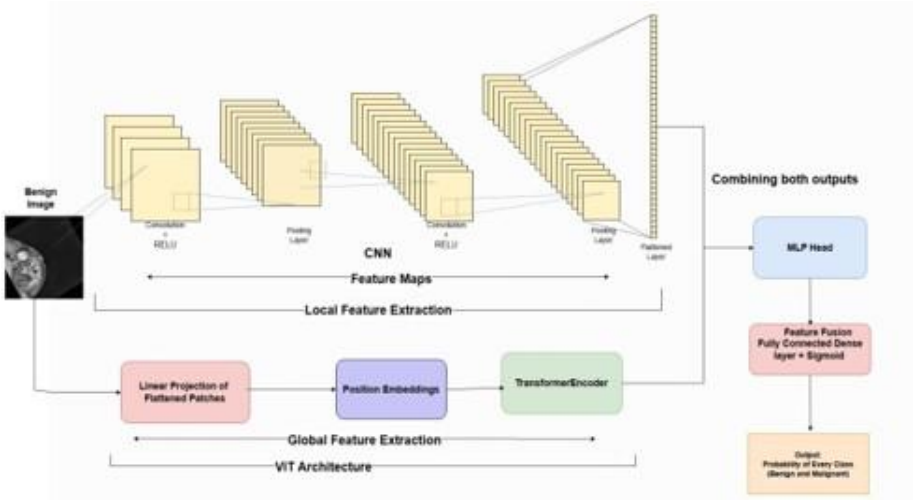

Figure 2: Proposed Hybrid Model

3.3.1 Cnn's Local Feature Extraction

The image is correlated by the convolutional layers, which also produce feature maps that highlight edges and minor morphological details. These patterns are essential for recognizing textures that could delineate benign or cancerous regions. Three convolutional layers make up the custom CNN's architecture, which is intended to extract low- and middle-level features. The layers consist:

Convolution Layer 1: 2 × 2 MaxPooling, 32 filters, kernel size (3 × 3), and ReLU activation.

Convolution Layer 2: 2 × 2 MaxPooling, 64 filters, kernel size (3 × 3), and ReLU activation.

Convolution Layer 3: 2 × 2 MaxPooling, kernel size (3 × 3), ReLU activation, and 128 filters. Additionally, the output is flattened to a 1 in the end.In order to provide a 1D vector of local features for the fusion that follows, the final output is flattened.

It moves locally over the image, creates a number of feature maps, and highlights different parts of the same image for each filter (e.g., edge detection, corner de- tection). Higher order and sophisticated characteristics, including clustering of deformed cells, are later learned from deeper convolutional layers.When search- ing for microcalcification or abnormalities in the breast tissue, these layers are helpful to have. In order for the model to understand both the linear and non-linear parts of the data, the ReLU activation function introduces a certain amount of non-linearity. Max takes a down sample of the feature maps. pooling layers to preserve any significant features while reducing the dimensionality of their spatial size. In ad- dition to improving computing efficiency, this introduces translation invariance, which allows an input image to be translated in the plane without changing the probability that the pattern of interest will show up in the picture. A pyramidal representation of features is created in pooling layers by progressively lowering the spatial information while concentrating on a small number of image re- gions.The convolutional and pooling layers provide feature maps at different degrees of abstraction, with the deeper levels identifying higher level features like forms, sectors of interest, etc., while the lower levels detect low level features like edges, fine structures, or textures. These feature maps undergo many convolution and pooling processes before being flattened and transformed into a single vector. Through this process, the 2D spatial data is unrolled into a format that may be used to link to the Vision Transformer's global feature maps. The local features that were taken from each area of the image were flattened, which made them linearly arranged to be concatenated to global features. This phase is a crucial link between the sug- gested hybrid architecture and the CNN-based local representation capability. CNN performs best on tasks that need a fine level of precision in small, local patterns that point to irregular ones.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.3.2 Global Extraction of Features Using Vision Trans- former (ViT)

By learning the input image's global relational characteristic, the Vision Trans- former enhances the CNN architecture. It considers the relationship between distant locations and processes the image all at once. The region of interest in the input image is divided into fixed-size patches, each of which represents a transformer token.The input image is patched to a predetermined size of non- overlapping patches of 16 × 16 pixels, or 196 patches per image, after being scaled to 224 × 224 pixels. The entire image is viewed in terms of sequential,comprehensible chunks thanks to these patches, which guarantee that no portion of the image is left unlabeled. This works especially effectively for medical imaging since anomalies may be found in tiny, obscure parts of a huge body. Po- sitional embeddings are concatenated to maintain spatial linkages, each patch is flattened into a 1D vector of length 768, and a linear layer is applied to the patch to produce dense vectors. This results in more pixel data with latent feature space representation and high dimensional feature vectors. As an embedding layer, the linear projection is helpful because it keeps each patch in the form of a high-dimensional vector that can communicate with other patches through the self-attention mechanism. To capture relationships between patches and extract long-range dependencies and global emergent features, the encoded patch sequence is then fed into a stack of 12 transformer layers, each of which has 12 self-attention heads and a feed-forward multilayer perceptron (MLP) with an intermediate dimension of 3072. The ViT now serves as the global feature extractor since the categoriza- tion head has been removed. With respect to the overall interaction with other regions, the decoder technique enables the model to focus on particular areas of the image. For instance, it can link a patch or irregularity in one area of the picture to another that is far away from the original one in the tissue map, giving a general notion of the tissue's architecture.

3.3.3 Categorization By Feature Fusion

In the final step, the CNN and ViT extracts are expertly combined to provide a compelling prognosis.A new composite feature vector is created by joining the CNN's feature maps, which we call local features, and the ViT's feature maps, which we call global features. Therefore, the integrated representation ensures that the model will consider both broad and large-scale trends as well as particular and small-scale aspects. The hybrid representation allows for more accurate model prediction and contains a lot of information. For example, although local features might identify a suspicious texture, global features might show that this texture is important to the tissue pattern's overall network.A dropout layer with a dropout probability of 0.5 is implemented to lessen the degree of over-fitting. In order to help the model learn from unseen data, this layer occasionally randomly turns off the relevant neurons during training. The final probabilities for each class—malignant or benign are then provided by a thick layer that is fully linked and has sigmoid activation. All of the retrieved features are coordinated into a single decision-making system by this layer. Another benefit is that the network's expectation clearly shows the confidence margin in its predictions, whereas the sigmoid activation function ensures that the output is a probability distribution.

3.4 Other Model Types

In order to test and assess the efficacy of various well-known deep learning archi- tectures on the breast cancer classification dataset, including DenseNet121, In- ceptionV3, XceptionNet, and SE-ResNet, we conducted our experiment. Highly complex architectures are renowned for their cutting-edge feature extraction methods and cutting-edge picture classification outcomes. They also reveal some of the advantages and disadvantages of various design philosophies. Our dataset of benign and malignant photos was used to test each of the models mentioned, each of which used a different approach to find underlying patterns and guarantee excellent classification performance. As seen in Figure 3 our ini- tial test using the bespoke CNN model yielded training accuracy of 0.878 and validation accuracy of 0.853. The training loss was significantly lower than that of the hybrid model, but the validation loss was higher.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

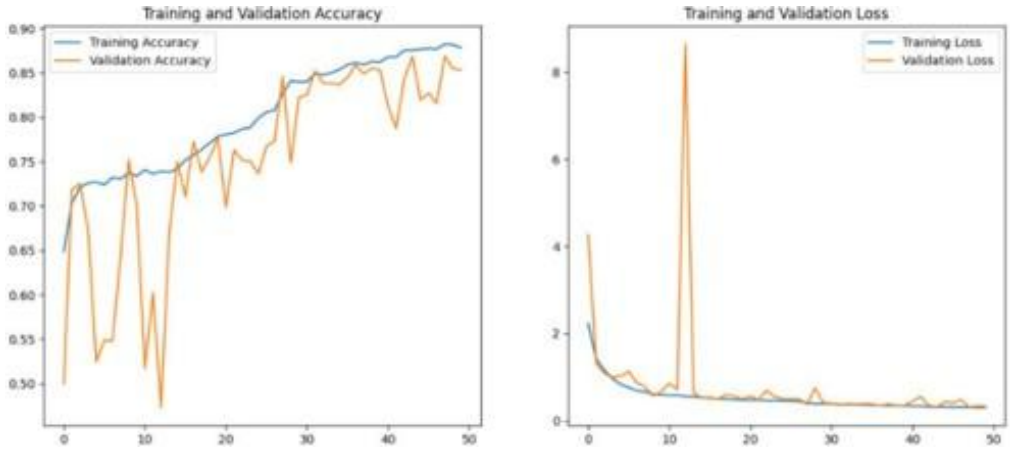

Figure 3: Accuracy And Loss Curves Of Custom CNN

A design called DenseNet121 [9]introduces dense connection, in which each layer is connected to every layer that came before it. This reduces redundancy and promotes feature reuse efficiency.The network ensures computational efficiency while effortlessly extracting rich features. DenseNet121 approached good validation metrics for our dataset, effectively capturing the subtle variations between benign and malignant samples. However, the model's high validation accuracy and low training loss raised concerns about the potential for overfitting, as il- lustrated in Figure 4 it may have been overly focused on training data and insufficiently focused on unseen data overall. The capacity to recognize a vari- ety of patterns in the input photos is introduced by InceptionV3 [28]. Making use of Inception modules. It uses a variety of filters with varying sizes to extract features at different scales. With factorized convolutions and auxiliary classi- fiers, the architecture is quite effective, although there is not much validation loss. When used in our dataset, it clearly demonstrated how important features could be extracted. It may not perform well on unseen datasets, indicating the need for more data augmentation or fine-tuning to enhance generalization. As seen in Figure 5 it was overfitted with an accuracy of 0.99, just like DenseNet. Squeeze-and-excitation blocks are added to the ResNet architecture's baseline by SE-ResNet[8], which recalibrates the feature maps to place greater emphasis on the pertinent patterns. It is evident that this approach enhances the net- work's representational capability and is appropriate for datasets with nuanced feature distinctions. With a fair balance between training and validation perfor- mance, SE-ResNet demonstrated strong results on our dataset. Unlike some of the other architectures, SE-ResNet exhibited a superior capability of generalization, hence exhibiting the ability to perform well over a wider range of clinical scenarios of medical imaging. Figure 6 displayed the SE-ResNet training and validation curves.

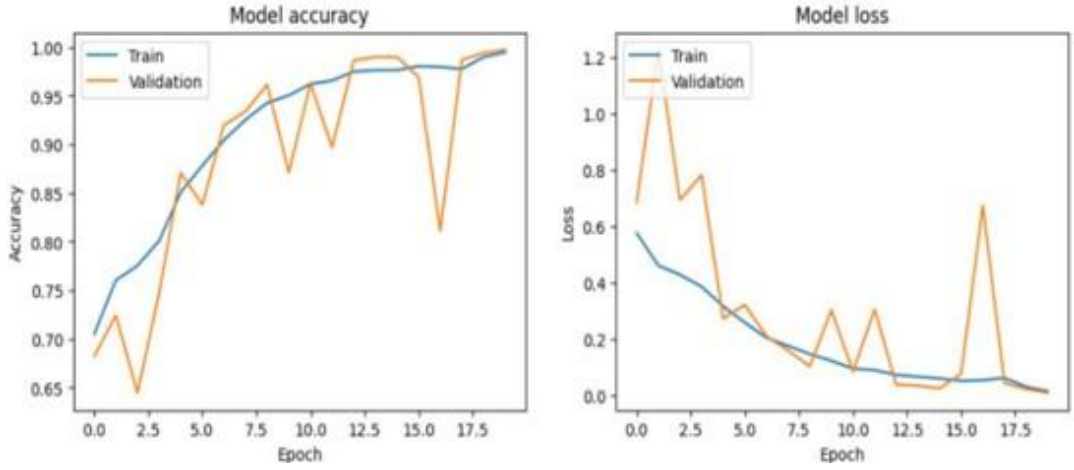

Figure 4: Accuracy and loss curves of DenseNet121

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

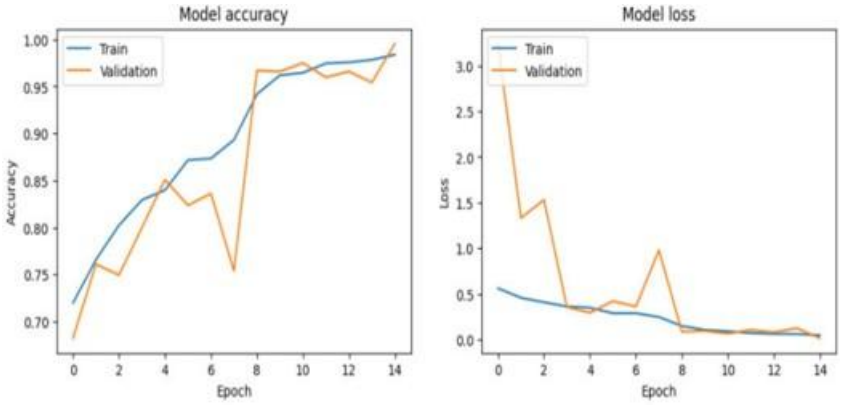

Figure 5: Accuracy and Loss Curves of InceptionV3

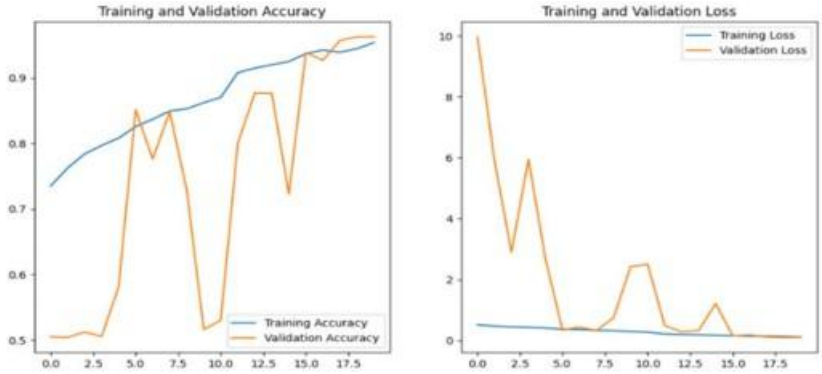

Figure 6: Accuracy and Loss curves if SE -Resnet

For a very lightweight architecture, XceptionNet [4] is an extension built on depth-wise separable convolutions that effectively separate spatial and channel- wise feature extraction. They are able to catch the image's finer details. As demonstrated in Figure 7, when applied to the dataset, XceptionNet provided nearly flawless accuracy, successfully differentiating benign tumor photos from malignant tumor images. The model may have a significant generalization issue in circumstances caused by real-world variability if its performance on other measures appears to suggest overfitting.

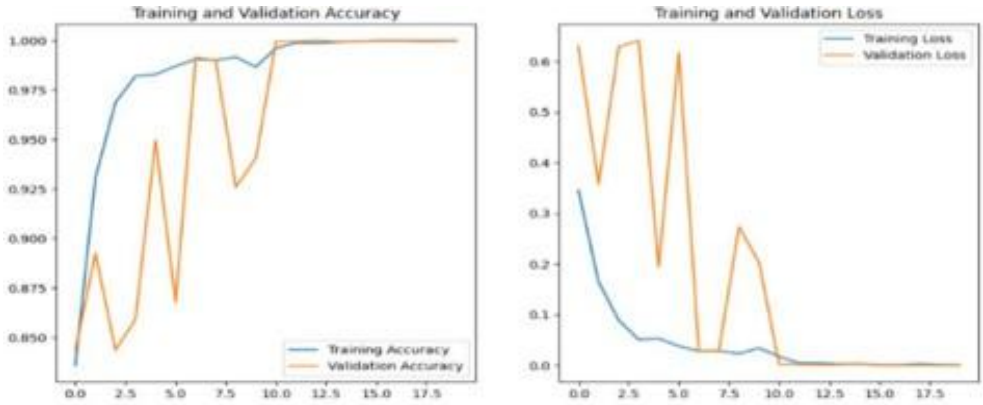

Figure 7: Accuracy and Loss Curves Of XceptionNet

4.RESULTS

The CNN-ViT model's experimental results demonstrate its efficacy in classi- fying breast cancer by striking a balance between generalization and accuracy. For 20 epochs, the training accuracy is 0.853 and the validation accuracy is 0.901, as seen in Figure 8.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

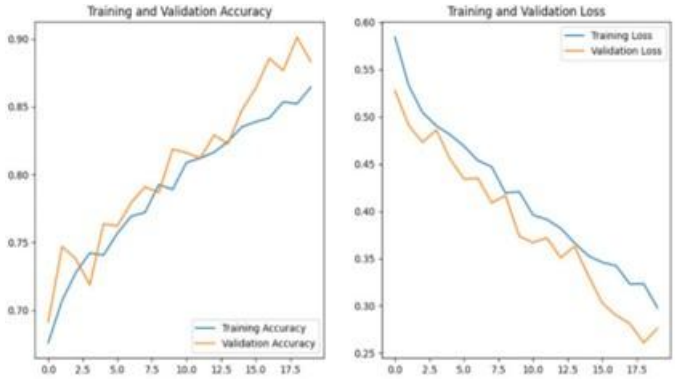

Figure 8: Training and validation graphs of CNN+ViT model

These findings show that the CNN+ViT method effectively leverages the com- plementary benefits and applications of ViTs and CNNs to provide a model that performs well when applied to fresh data. This one mostly alludes to the sig- nificance of robustness in medical screening. Our study's CNN+ViT confusion matrix is shown in Figure 9.

A comparatively little difference between training and validation measures shows that the model operates well with such accuracy and without overfitting. This precision is achieved by fusing the ability of ViT to capture the global dependencies throughout the image with CNN's ability to catch local features, such as edges and textures. For both benign and malignant instances, the categorization report table 1 offers comprehensive metrics such as precision, recall, and F1-score. A accuracy of 0.91 and recall of 0.89 were noted for benign patients. Precision was 0.89 and recall was 0.92 for malignant patients. The categorization has an overall accuracy of 90% and F1-scores of 0.90 for both categories. All indicators' macro and weighted averages produced a steady 0.90 as well, suggesting that performance was balanced across classes.

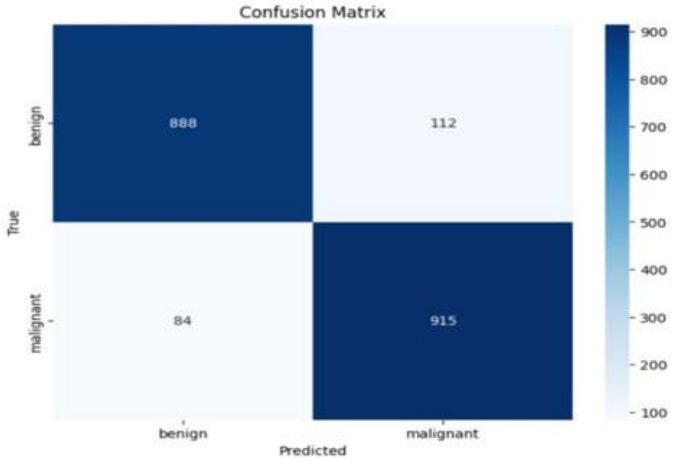

Figure 9: Confusion matrix of CNN+ViT model.

Table 1: Classification report showing precision, recall, and F1-score for benign and malignant classes in CNN+ViT model.

	Precision	Recall	F1-score	Support
Benign	0.91	0.89	0.90	1000
Malignant	0.89	0.92	0.90	999
Accuracy			0.90	1999
Macro avg	0.90	0.90	0.90	1999
Weighted avg	0.90	0.90	0.90	1999

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The model retains similarly optimal performances for benign and malignant cases, demonstrating great effectiveness through accurate classification. In med- ical settings, data analysis demands extra care because both false positive and negative results can lead to serious health problems.

5 DISCUSSION

Results from the additional pre-trained models, base models, and hybrid mod- els—classic CNN, Hybrid (CNN Combined with ViT), and pre-training models DenseNet121, Inception V3, XceptionNET, and SE-ResNet—are shown in the table 2 below[19]. Thus, when both models are employed together, CNN + ViT, the lowest loss (0.238) correlates to the best accuracy result. As seen in Figure 8, it has demonstrated training accuracy of 85.3%, validation accuracy of 90.1%, and validation loss of 0.260. We are confident that the hybrid model does not over-fit on the training data and can, therefore, be trusted more in the actual world because of the difference in aggressiveness between the training and val- idation sets.Conversely, pre-trained models show overfitting even though they get good metrics.For example, DenseNet121 has incredibly low losses (training loss: 0.015, validation loss: 0.009) and a near-perfect validation accuracy of 99.7%. Similarly, the XceptionNET model has a validation loss of 0 and the Inception V3 and XceptionNET models offer validation accuracy of 99.5% and 100%, respectively. Practically speaking, these outcomes might be striking, but they could also suggest that the models are only remembering training data rather than identifying powerful, broad patterns.Clinical metrics evaluation of models is shown in ”Table 3 using Precision, Recall, and F1-Score evaluation to offer robust metrics analysis that goes beyond accuracy measurements. With a 90% F1-score, the hybrid model outperforms the baseline CNN model by com- bining CNN with ViT. Since XceptionNet attains 100% precision, recall, and F1-score metrics, it performs the best out of all the models.

Table 2: Comparison table with all models.

Model	Training Acc	Validation ACC	Training Loss	Validation Loss
CNN	0.878	0.853	0.306	0.332
Hybrid	0.853	0.901	0.323	0.260
DenseNET121	0.997	0.915	0.015	0.009
InceptionV3	0.983	0.995	0.047	0.016
XceptionNet	1.000	1.000	0.000	0.000
SE-ResNet	0.954	0.962	0.114	0.106

Table 3: Comprehensive performance comparison of models using multiple met- rics.

Model	Precision (%)	Recall (%)	F1-Score (%)
CNN	51.00	51.00	50.00
Hybrid	90.00	90.00	90.00
SE-ResNet	96.00	96.00	96.00
XceptionNet	100.00	100.00	100.00

Note: DenseNet121 and InceptionV3 showed performance comparable to XceptionNet, with F1-scores close to 99%, and were excluded from the table for clarity.

6 CONCLUSION AND FUTURE SCOPE

Using the CNN+ViT architecture, which combines the integrating model CNN for the detailed local feature extraction and the ViT to capture the global contexts, the primary goal is to create a reliable breast cancer classification system. This methodology enhances classification accuracy while maintaining generalization capabilities. With balanced training and validation accuracy, the CNN+ViT model performs well and can

distinguish between benign and malig- nant cases with little overfitting. For upcoming use in clinical settings, model architecture optimization may further increase the suggested model's computa- tional effectiveness. Future studies will also compare the computational efficiency metrics of all the models examined in the study, such as inference time, memory usage, and errors. in order to assess their suitability for clinical application in real time. Ad- ditionally, we intend to investigate the XceptionNet model's unusually high validation accuracy, which may indicate overfitting or data leakage.We plan to employ interpretability techniques such as Grad-CAM and t-SNE to further un- derstand model behavior by visualizing feature relevance and confirming that the network is actually learning important diagnostic patterns rather than artifacts. Last but not least, expanding the dataset from a sample size of 5,000 to lakhs of samples can add several histopathological images from various tissue samples. With this feature, the established model's diagnostic capabilities and adapt- ability are further enhanced. Enhancing a cloud-based program or interactive interface facilitates real-time access by medical experts, enhances early detec- tion, and improves patient outcomes.

REFRENCES

[1] Meshrif Alruily, Alshimaa Abdelraof Mahmoud, Hisham Allahem, Ay- man Mohamed Mostafa, Hosameldeen Shabana, and Mohamed Ezz. En- hancing breast cancer detection in ultrasound images: An innovative ap- proach using progressive fine-tuning of vision transformer models. *Interna- tional Journal of Intelligent Systems*, 2024(1):6528752, 2024.

[2] Sannasi Chakravarthy, Bharanidharan Nagarajan, V Vinoth Kumar, TR Mahesh, R Sivakami, and Jonnakuti Rajkumar Annand. Breast tumor classification with enhanced transfer learning features and selection using chaotic map-based optimization. *International Journal of Computational Intelligence Systems*, 17(1):18, 2024.

[3] Xuxin Chen, Ke Zhang, Neman Abdoli, Patrik W Gilley, Ximin Wang, Hong Liu, Bin Zheng, and Yuchen Qiu. Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. *Diagnostics*, 12(7):1549, 2022.

[4] Franc¸ois Chollet. Xception: Deep learning with depthwise separable con- volutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Min- derer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Samuel Derbie Habtegiorgis, Daniel Shitu Getahun, Animut Takele Telayneh, Molla Yigzaw Birhanu, Tesfa Mengie Feleke, Alemu Basazn

[7] Mingude, and Lemma Getacher. Ethiopian women's breast cancer self- examination practices and associated factors. a systematic review and meta-analysis. *Cancer epidemiology*, 78:102128, 2022.

[8] Nada M Hassan, Safwat Hamad, and Khaled Mahar. Yolo-based cad frame- work with vit transformer for breast mass detection and classification in cesm and ffdm images. *Neural Computing and Applications*, 36(12):6467– 6496, 2024.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Pro- ceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Wein- berger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700– 4708, 2017.

[11] Kiran Jabeen, Muhammad Attique Khan, Jamel Balili, Majed Alhaisoni, Nouf Abdullah Almujally, Huda Alrashidi, Usman Tariq, and Jae-Hyuk Cha. Bc2netrf: breast cancer classification from mammogram images us- ing enhanced deep learning features and equilibrium-jaya controlled regula falsi-based features selection. *Diagnostics*, 13(7):1238, 2023.

[12] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fa- had Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[13] Ella Mahoro and Moulay A Akhloufi. Breast cancer classification on ther- mograms using deep cnn and transformers. *Quantitative InfraRed Ther- mography Journal*, 21(1):30–49, 2024.
[14] Ahmed Dhahi Mohammed and Dursun Ekmekci. Breast cancer diagno- sis using yolo-based multiscale parallel cnn and flattened threshold swish. *Applied Sciences*, 14(7):2680, 2024.
[15] Ehsan Monjezi, Gholamreza Akbarizadeh, and Karim Ansari-Asl. Ri-vit: A multi-scale hybrid method based on vision transformer for breast cancer detection in histopathological images. *IEEE Access*, 2024.
[16] Swati Nadkarni and Kevin Noronha. Breast cancer detection using an en- semble of convolutional neural networks. *International Journal of Electrical & Computer Engineering (2088-8708)*, 14(1), 2024.
[17] Maged Nasser and Umi Kalsom Yusof. Deep learning based methods for breast cancer diagnosis: a systematic review and future direction. *Diag- nostics*, 13(1):161, 2023.
[18] Marwa Obayya, Mashael S Maashi, Nadhem Nemri, Heba Mohsen, Ab- delwahed Motwakel, Azza Elneil Osman, Amani A Alneil, and Mo- hamed Ibrahim Alsaid. Hyperparameter optimizer with deep learning-based decision-support systems for histopathological breast cancer diag- nosis. *Cancers*, 15(3):885, 2023.
[19] Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.
[20] Vasudha Rani Patheda, Gunda Laxmisai, Gokulnath Bv, SP Siddique Ibrahim, and S Selva Kumar. A robust hybrid cnn+vit framework for breast cancer classification using mammogram images. *IEEE Access*, 2025.
[21] Danil Yu Pimenov, Leonardo RR da Silva, Ali Ercetin, O ğuzhan Der, Tadeusz Mikolajczyk, and Khaled Giasin. State-of-the-art review of ap- plications of image processing techniques for tool condition monitoring on conventional machining processes. *The International Journal of Advanced Manufacturing Technology*, 130(1):57–85, 2024.
[22] Francesco Prinzi, Marco Insalaco, Alessia Orlando, Salvatore Gaglio, and Salvatore Vitabile. A yolo-based model for breast cancer detection in mam- mograms. *Cognitive Computation*, 16(1):107–120, 2024.
[23] S Vishnu Priyan, K Rajalakshmi, J Parivendhan Inbakumar, and A Swami- nathan. Enhanced melanoma detection using a fuzzy ensemble approach integrating hybrid optimization algorithm. *Biomedical Signal Processing and Control*, 89:105924, 2024.
[24] Suman Rani, Minakshi Memoria, Mamta Rani, and Rajiv Kumar. Breast cancer detection using mammographic images over convolutional neural network. In *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 73–78. IEEE, 2023.
[25] Margo Sabry, Hossam Magdy Balaha, Khadiga M Ali, Tayseer Hassan A Soliman, Dibson Gondim, Mohammed Ghazal, Tania Tahtouh, and Ay- man El-Baz. A vision transformer approach for breast cancer classification in histopathology. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2024.
[26] SR Sannasi Chakravarthy, N Bharanidharan, and Harikumar Rajaguru. Multi-deep cnn based experimentations for early diagnosis of breast cancer. *IETE Journal of Research*, 69(10):7326–7341, 2023.
[27] SR Sannasi Chakravarthy, N Bharanidharan, V Vinoth Kumar, TR Ma- hesh, Mohammed S Alqahtani, and Suresh Guluwadi. Deep transfer learn- ing with fuzzy ensemble approach for the early detection of breast cancer. *BMC Medical Imaging*, 24(1):82, 2024.
[28] SR Shrivastava, P Saurabh Shrivastava, and Jegadeesh Ramasamy. Self breast examination: A tool for early diagnosis of breast cancer. *Am J Public Health Res*, 1(6):135–139, 2013.
[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbig- niew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recog- nition*, pages 2818–2826, 2016.
[30] Ling Tan, Ying Liang, Jingming Xia, Hui Wu, and Jining Zhu. Detection and diagnosis of small target breast masses based on convolutional neural networks. *Tsinghua Science and Technology*, 29(5):1524–1539, 2024.
[31] Sudhakar Tummala, Jungeun Kim, and Seifedine Kadry. Breast-net: Multi- class classification of breast cancer from histopathological images using ensemble of swin transformers. *Mathematics*, 10(21):4109, 2022.
[32] Lulu Wang. Deep learning techniques to diagnose lung cancer. *Cancers*, 14(22):5569, 2022.

Journal of Information Systems Engineering and Management
2026, 11(3s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[33] Lulu Wang. Mammography with deep learning for breast cancer detection.*Frontiers in oncology*, 14:1281922, 2024.
[34] Sameh Zarif, Hatem Abdulkader, Ibrahim Elaraby, Abdullah Alharbi, Wail S Elkilani, and Pawel- Plawiak. Using hybrid pre-trained models for breast cancer detection. *Plos one*, 19(1):e0296912, 2024.
[35] Bashir N, Bhosle N. Deep learning driven multi-modal breast cancer detection for early and accurate diagnosis. Journal of Information Systems Engineering and Management. 2026;11