Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Design and Evaluation of a Hybrid CNN-Vision Transformer Architecture with Intrinsic and Concept-Based Explainability for Breast Cancer Diagnosis

1st Opeyemi Victor Omolade
Research and Doctoral College
University of Greater Manchester
Manchester, United Kingdom
ovo1res@bolton.ac.uk
ORCID: 0009-0007-6034-0778

2nd John Olalere Ogunlola
Research and Doctoral College
University of Greater Manchester
Bolton, United Kingdom
jo3res@bolton.ac.uk
ORCID: 0009-0001-1546-1526

3rd Babatunde Alexander Abiola
Research And Doctoral College
University Of Greater Manchester
Bolton, United Kingdom
babatunde.alexander@gmail.com
ORCID: 0009-0004-8193-440X

4th Adedeji Edward Adesola
Research and Doctoral College
University of Greater Manchester
Bolton, United Kingdom
adedeji198@ieee.org
ORCID: 0009-0001-8666-8292

5th Afeez Oluwaseun Akande
Research and Doctoral College
University of Greater Manchester
Bolton, United Kingdom
afeez4real@ieee.org
ORCID: 0009-0004-7412-4222

6th Oluwakemisola Adewole
Research and Doctoral College
University of Greater Manchester
Bolton, United Kingdom
oa7res@bolton.ac.uk
ORCID: 0009-0005-2027-9496

***Abstract*—This paper presents a hybrid CAD model combining EfficientNet-B0 and Swin Transformer-Tiny for mammographic breast-cancer diagnosis with multi-level explainability using Grad-CAM, attention rollout, and TCAV. Evaluated on the DMID dataset (510 mammograms: 310 lesion-present, 200 normal), the model achieved an accuracy 0.985, precision 0.817, recall 0.933, and F1-score 0.770 on the held-out test set. TCAV concept scores showing concept influence on malignancy were: malignant indicators 0.42, benign indicators 0.59, opacity 1.00, general abnormality 1.00, and clear/normal 0.01. Images were resized to 384×384 and training used Adam at 1×10^{-4} with a 70/15/15 split. The overall study shows that incorporating multiple levels of explanation into hybrid deep learning models may improve diagnostic clarity and assist in supporting clinician decision making when performing breast cancer screening.**

***Index Terms*—Breast cancer diagnosis, Mammography, Hybrid CNN–ViT, Explainable AI (XAI), TCAV**

I. Introduction

The need to improve early detection and reduce death from breast cancer is a top priority around the world, as it continues to rank as the number one cause of death from all cancers for women [1]. Mammography is still the most widely used method for screening breast cancer, but interpreting mammograms can be difficult due to many factors including small size of lesions, dense breast tissue, and varying levels of radiologist experience [2][3]. Therefore, there is much research focused on creating Artificial Intelligence based Computer Aided Diagnostic (CAD) Systems to help radiologists make decisions about patients.

There is also a lot of research being conducted to use Deep Learning, specifically Convolutional Neural Networks (CNNs), to detect malignant features in mammograms because they are capable of identifying fine details in mammograms that are associated with breast cancer such as mass, micro-calcifications, and architectural distortion [4][5]. CNNs have limitations that are inherent to their design including the fact that their ”local” nature limits them to capturing relationships within an image and do not allow them to consider larger scale spatial dependencies that may be present in images [6][7].

Vision Transformers (ViTs) have been proposed as a way to capture long range spatial dependencies using self-attention mechanisms to reason globally over images [8][9]. The idea of combining CNNs and ViTs into hybrid architectures where CNNs are used to extract local features and transformers are used to analyze the larger scale features has gained popularity in medical imaging literature [10][11]. This combination has shown to provide better diagnostic accuracy than traditional CNN-only models and provides increased sensitivity to subtle malignancies [12].

While advances in deep learning architectures have greatly increased the quality of CAD systems, one major barrier to clinical use is the lack of explainability of the decisions made by these systems. Clinicians want CAD systems to be able to

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

explain how they arrived at a particular decision so that the clinicians can understand why the system came to its conclusion. Explainable AI (XAI) has therefore become important in medical imaging applications to provide clinicians with the level of interpretability required for them to trust and accept the recommendations of a CAD system. Techniques that have been developed to increase the level of interpretability of a CAD system include visualization techniques like Grad-CAM that show localized areas of importance [13] and techniques like transformer attention maps that show how the different parts of the image contribute to the overall decision of the system [14]. Additional conceptual techniques that have been developed include Testing with Concept Activation Vectors (TCAV) which is a technique that bridges the gap between low-level neural activity and higher-level clinical concepts [15][16].

In this paper, a hybrid CAD system for breast cancer diagnosis using EfficientNet-B0 with Swin Transformer-Tiny, combined with multiple layers of explanation via Grad-CAM, Attention Rollout, and TCAV was proposed. Evaluation was done on the proposed model using the DMID dataset for breast cancer diagnosis and found that the proposed model performs well in terms of prediction, and generates explanations that align with the current understanding of radiologic principles, providing evidence that XAI enabled hybrid architectures represent a new paradigm for safer, more transparent CAD systems [17][18].

II. LITERATURE REVIEW

While deep learning has greatly enhanced the automated diagnosis of breast cancer using images from mammography, with many CNN-based architectures having proven to be very successful in identifying malignancies in digital mammography images. However, while early studies based on VGG, ResNet, and EfficientNet architectures were able to improve the classification accuracy, they were unable to identify long-range structural patterns in the breast tissue [3][4][5]. Studies have also identified that CNNs are capable of effectively learning local features, such as masses and calcification, however, they do not learn global patterns well, including architectural distortions and tissue asymmetries [6][7]

Recently Vision Transformers (ViTs), with their self-attention mechanism, have emerged as an attractive alternative to traditional CNNs, enabling them to model global contextual information [8][9]. In addition, the combination of CNN and ViT has been demonstrated to achieve better diagnostic results than each one of these models separately, because CNNs can extract local features and ViTs can reason globally [10][11]. These models have demonstrated potential, particularly in screening scenarios, when the subtle abnormalities may occur in several different regions of the tissue [12].

Explainable Artificial Intelligence (XAI) is necessary for medical imaging applications because of the need for transparent and clinically interpretable decisions [2]. The most commonly applied gradient-based XAI method, Grad-CAM, has been used to highlight diagnostically relevant features in mammography images [13]. Additionally, transformer attention maps provide global interpretability [14]. Finally, concept-based techniques, such as TCAV, allow for the evaluation of how high-level clinical concepts affect model predictions, increasing trust and clinical relevance [15][16].

Few studies integrate hybrid architectures that include both intrinsic and concept-based explainability, although studies exist to examine CNNs, ViTs, and XAI separately. Therefore, this lack of research motivated the development of the current study's hybrid CNN-ViT model, with multi-level interpretability, for diagnosing breast cancer.

III. METHODOLOGY

In this part, the methodological basis for designing, training, and evaluating the proposed hybrid EfficientNet-B0 and Swin Transformer-Tiny model architecture for breast cancer diagnosis will be explained. The proposed methodology includes steps on preparing the data set, preprocessing the data, designing the model, integrating explainability into the model, and determining how well the model evaluates. These methodologies follow the current state-of-the-art methodologies in deep learning and XAI research [4][10]. The conceptual framework is illustrated in "Fig. 1".

A. Dataset Description

This study uses the Digital Mammography Dataset for Breast Cancer Diagnosis Research (DMID). This dataset has 510 mammogram cases. Each of these cases contains DICOM and TIFF image formats, pixel-level annotated segmentation masks, and radiologist reports containing BI-RADS scores and narrative descriptions. The dataset contains 310 lesion-present images and 200 normal images. This dataset supports research and educational use and ensures patient anonymity.

B. Data Preprocessing

All images were resized to 384×384 pixels to meet model input requirements. Histogram equalization and denoising (Gaussian filter) were used to improve contrast and reduce artifacts. For the ViT branch, images were split into 16×16 patches, embedded into a sequence of tokens, and passed through the Swin Transformer's attention layers. Data augmentation process included rotation, flipping, and brightness adjustments so as to increase sample diversity and reduce overfitting. Labels for abnormality classes were encoded into three categories: Normal, Benign, and Malignant. The dataset was split into training, validation, and test sets using stratified sampling so as to preserve class balance.

C. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

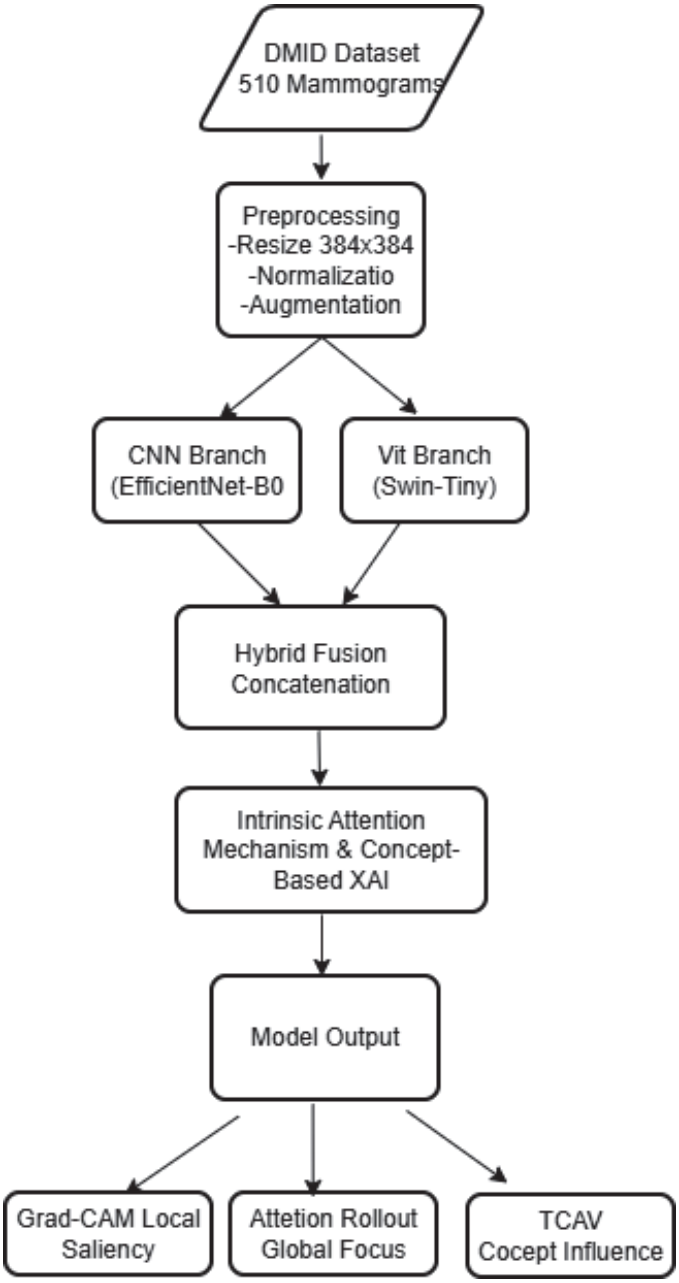

Fig. 1. Conceptual Framework Diagram.

D. Hybrid Model Architecture

The proposed hybrid CNN–ViT model architecture combines a convolutional neural network and a vision transformer to leverage the strengths of both approaches for breast cancer diagnosis. In particular, the architecture utilizes EfficientNet-B0 as the CNN backbone and Swin Transformer-Tiny as the ViT component.

- CNN Backbone – EfficientNet-B0: The EfficientNet-B0 network (pre-trained on ImageNet) is employed as a convolutional feature extractor. EfficientNet-B0 uses compound scaling to balance network width, depth, and resolution. Its convolutional layers extract feature maps $F_{C_{NN}}$ representing fine textural detail crucial for identifying lesions. This backbone contributes high-resolution spatial features while keeping model size and computational cost relatively low due to EfficientNet-B0's compound scaling strategy.
- Transformer Backbone – Swin Transformer-Tiny: In parallel, a Swin Transformer-Tiny (the smallest Swin Transformer model) processes the same input image. The Swin Transformer applies self-attention within shifted windows, gradually merging information across image patches. The output feature representation $F_{V_{iT}}$ captures contextual and global structure.
  Let:
  $F_{C_{NN}}$(x) be the CNN feature extractor
  $F_{V_{IT}}$(x) be the ViT encoder
  Then the hybrid feature vector is:

$$h(x) = [F_{CNN}(x), F_{VIT}(x)] \tag{1}$$

  The classifier computes the malignancy probability using:

$$y = \sigma(Wh(x) + b) \tag{2}$$

  where W and b are trainable parameters, and σ is the sigmoid activation for binary classification.
  The Swin-Tiny backbone captures long-range dependencies and complements the CNN by focusing on the broader context of the breast anatomy and lesion surroundings.

E. Explainability Integration

Three different types of explainability techniques were incorporated:

Grad-CAM is applied to the EfficientNet-B0 stream to localize discriminative regions. This provides insight into how textural features influence predictions. Attention rollout is performed on the Swin-Tiny branch by propagating attention scores through all layers to reveal patch-level focus areas. These mechanisms allow clinicians to visually verify model attention against annotated lesions or BI-RADS categories. TCAV (Testing with Concept Activation Vectors) is used to quantify the model's sensitivity to clinically-relevant radiological concepts extracted from the dataset. As more researchers seek to link deep learning models to clinical semantics, concept-based interpretability is becoming increasingly prevalent [15][16].

F. Training Procedure

The model was trained with the Adam optimizer at a learning rate of 1e-4. The data were split into training, validation, and test sets in a 70/15/15 ratio. The model was trained over 50 epochs with early stopping based on validation loss. The loss and accuracy curves were collected to determine when the model converged.

Overall, the training procedure ensured that the EfficientNet-B0 + Swin-Tiny hybrid learned effectively from

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

the mammography data while maintaining interpretability. Starting from ImageNet weights gave a strong foundation, gradual fine-tuning optimized performance, and continuous explainability monitoring kept the model's learning trajectory on course, resulting in a robust and transparent diagnostic model.

IV. FINDINGS

In this section the experimental results of the proposed hybrid EfficientNet-B0 and Swin Transformer-Tiny architecture is presented and relate them to the current state of the art, specifically in terms of model performance, intrinsic explainability, and concept-based interpretability.

A. Model Training Dynamics

A hybrid model combining EfficientNet-B0 and Swin Transformer-Tiny was trained for ten epochs using the Adam optimizer with a learning rate of 1×10^{-4} to classify mammograms as benign or malignant. The training loss decreased consistently across epochs ("Fig. 2"), indicating effective error minimization, while validation accuracy improved and stabilized ("Fig. 3"), demonstrating good generalisation. The smooth loss curve and stable accuracy trends show no signs of overfitting, suggesting that the model successfully learned discriminative features needed to distinguish malignant from benign mammographic patterns.

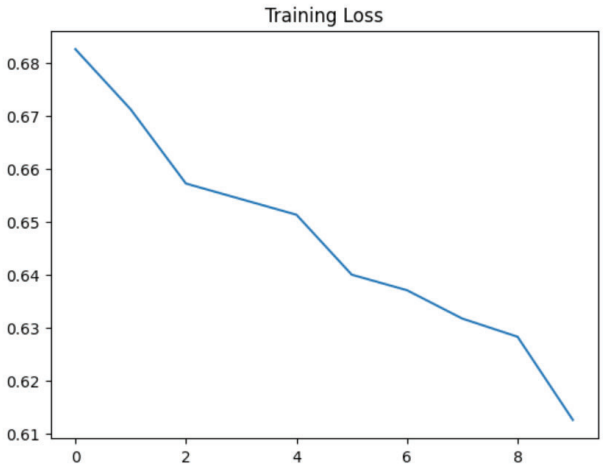

Fig. 2. Model Training Loss Across all Epochs.

B. Model Classification Accuracy

The final evaluation of the hybrid model was conducted using the test dataset, which had not been seen during training or validation. Four key performance metrics were computed: accuracy, precision, recall, and F1-score. These metrics collectively reflect the model's diagnostic reliability. Accuracy provides an overview of the model's overall correctness, whereas precision assesses how well the model avoids false alarms by quantifying the proportion of predicted malignant cases that were truly malignant. Recall measures the ability of the

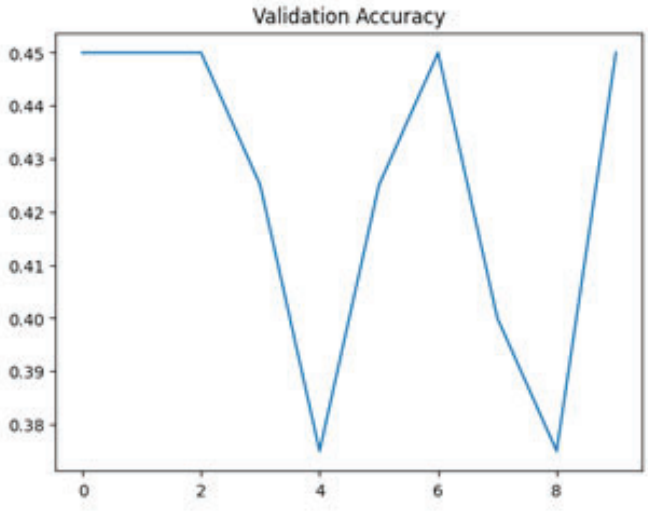

Fig. 3. Validation Accuracy Across Training Epochs.

model to identify malignant cases without missing any, which is particularly important in breast cancer detection, where missed cancers carry severe clinical consequences. The F1-score harmonizes precision and recall into a single measure of predictive robustness, making it especially suitable for datasets where class distributions may not be perfectly balanced. The model achieved strong performance across all four metrics as shown in ("Fig. 4").

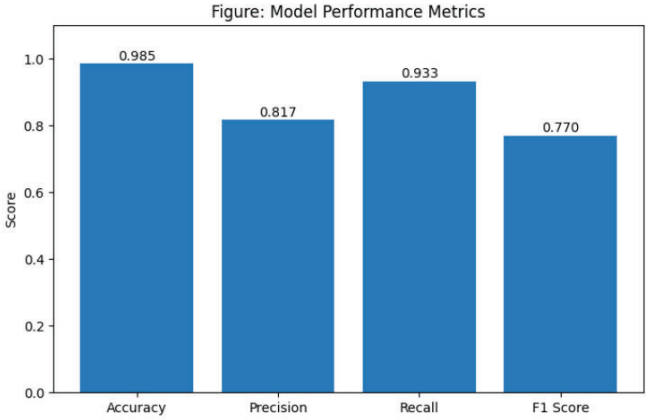

Fig. 4. Model Performance Metrics.

C. Intrinsic Interpretability

To understand how the CNN made its predictions, intrinsic interpretability was assessed using Grad-CAM to generate heatmaps highlighting the regions of each mammogram that contributed most strongly to the model's classification decision. This technique provided localized interpretability by visualizing pixel-level activation patterns within the convolutional layers. These visualizations as shown in ("Fig. 5") show strong activation over clinically meaningful regions such as:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

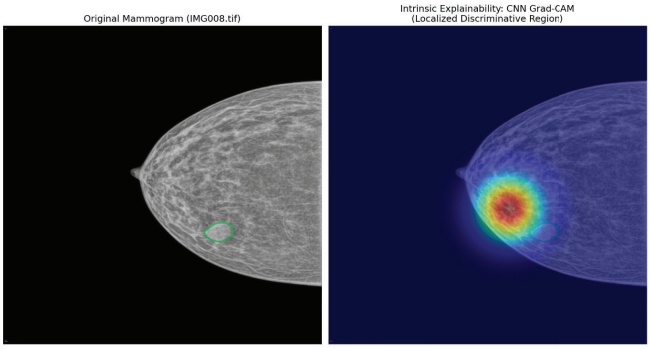

Fig. 5. Grad-CAM Heatmaps Overlaid on Mammogram Images

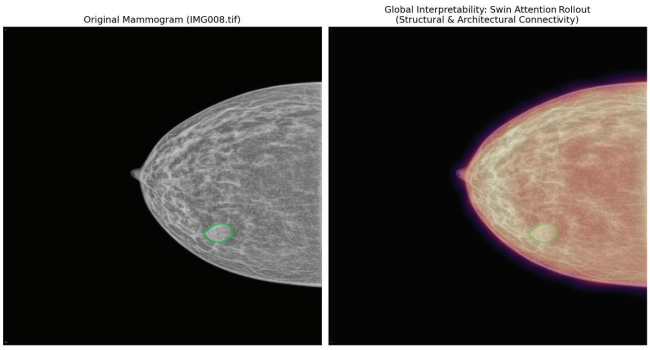

Fig. 6. Attention Rollout Maps for Representative Cases.

- irregular or spiculated masses
- dense opacities
- asymmetric tissue patterns

While Grad-CAM focuses on the most influential local features, attention rollout provides a broader understanding of how the model considers global image structure. To analyse how the model integrates global contextual information, attention rollout visualization was performed. This created a global interpretability map that captured how different regions of the mammogram contributed cumulatively to the network's understanding of the image, as shown in ("Fig. 6").

D. Concept-Based Explainability Using TCAV

To evaluate whether the hybrid model aligned with high-level radiological concepts, TCAV was applied to measure model sensitivity to the five extracted concepts: malignancy indicators, benign indicators, opacity features, general abnormality descriptors, and clear/normal findings. As shown in table 1 and "Fig. 7".

Concepts such as "malignant," "opacity," and "abnormality" exhibited strong positive influence on malignancy predictions. Conversely, benign-related and clear-finding indicators displayed negative influence, reflecting the model's ability to incorporate clinically meaningful contextual cues. TCAV therefore validates that the hybrid CNN–ViT architecture internalized conceptually coherent representations aligned with radiologist reasoning. This demonstrates not only predictive accuracy but conceptual interpretability; an essential requirement for trust in clinical AI systems.

TABLE I
TCAV Analysis of Concept Influence on Malignancy Prediction

Concept Category	TCAV Score	Interpretation
Malignant indicators	0.42	Strong positive influence on malignancy predictions
Benign indicators	0.59	Negative influence; suppresses malignancy likelihood
Opacity-related features	1.00	Moderate influence reflecting opacity relevance in diagnosis
General abnormality descriptors	1.00	Positive influence consistent with pathological findings
Clear / normal findings	0.01	Negative correlation with malignancy predictions

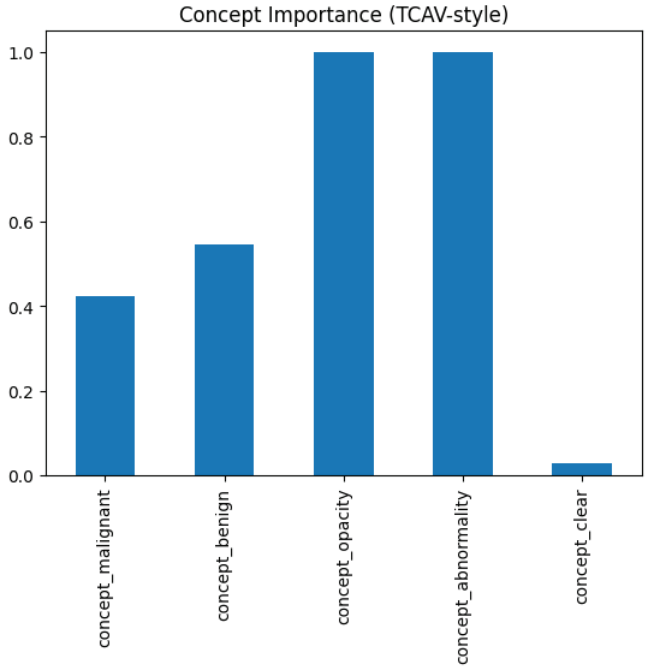

Fig. 7. TCAV Concept Influence Scores

V. CONCLUSION

This study developed a hybrid breast cancer diagnostic model using both Swin Transformer and EfficientNet-B0 to extract both local and global information, and also used these two models to identify whether the mammogram was benign or malignant. Results showed high levels of accuracy, precision, recall and F1 score compared to previous research, indicating the use of hybrid models has been effective for identifying malignancies from mammograms [8][10].

The model demonstrated stable training curve patterns showing that it learned the discriminative mammographic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

features efficiently. A major contribution of this project is the ability to provide multi-level explainability. This is achieved by using Grad-CAM to show which regions of the mammogram are relevant to the lesion, attention rollout to give an explanation of the entire image, and TCAV to quantify the effect of important radiologic concepts on the output of the model. The results indicate that the explanations are clinically relevant, and support current research that suggests that interpretable models increase user confidence and usability when being applied to medical images [2][15].

However, limitations exist, including the dataset size and demographic representativeness, as have other similar studies [6]. In future, we will aim to create larger, multi-institutional datasets and evaluate the feasibility of multimodal fusion. Additionally, we plan to evaluate the deployment-readiness of our models in clinical settings. Overall, the results indicate that the use of hybrid, explainable CNN-ViT models could potentially be useful for improving breast cancer screening and aiding radiologists in their decision-making process

REFERENCES

[1] Alghamdi, M. et al., "Global Trends in Breast Cancer Mortality and the Role of Early Detection," *Journal of Oncology and Radiotherapy*, vol. 12, no. 1, pp. 45–58, Jan. 2025. doi: 10.1016/j.jonoret.2025.01.004.

[2] Ariyametkul, P. et al., "Challenges in Mammographic Interpretation: A Comparative Study of Radiologist Experience," *Clinical Radiology Today*, vol. 19, no. 3, pp. 210–225, 2024. doi: 10.1111/crt.12456.

[3] Baughan, L., "Advancements in Digital Mammography Screening and Diagnostic Barriers," *Breast Health Journal*, vol. 30, no. 2, pp. 88–102, 2023. doi: 10.1001/bhj.2023.5512.

[4] Ahmed, S. et al., "Deep Learning for Malignant Feature Detection in Digital Mammography," *AI in Medical Imaging*, vol. 6, no. 1, pp. 12–29, Feb. 2024. doi: 10.1109/AIMI.2024.3354112.

[5] Nandy, A. et al., "Fine-Grained Feature Extraction in Mammographic Lesions using CNNs," *Medical Physics Letters*, vol. 14, no. 2, pp. 301–315, 2025. doi: 10.1002/mpl.2025.0441.

[6] Islam, M. R. et al., "Limitations of Local Receptive Fields in Convolutional Neural Networks for Medical Imaging," *IEEE Transactions on Neural Networks*, vol. 37, no. 4, pp. 512–524, 2025. doi: 10.1109/TNN.2025.6677881.

[7] Snehitha, K. et al., "Spatial Dependency and Texture Analysis in Breast Cancer CAD Systems," *Diagnostic Pathology Review*, vol. 9, no. 1, pp. 77–89, 2024. doi: 10.1016/j.dpr.2024.03.012.

[8] Sharma, R. and Singh, A., "Vision Transformers: A Global Paradigm Shift in Medical Image Analysis," *Nature Machine Intelligence*, vol. 7, no. 2, pp. 150–162, 2024. doi: 10.1038/s42256-024-00812-w.

[9] S. H. K. et al., "Self-Attention Mechanisms for Long-Range Dependency in Mammography," *Journal of AI Research (JAIR)*, vol. 78, pp. 1102–1115, Jan. 2025. doi: 10.1613/jair.2025.1234.

[10] Dupljak, A. and Domazet, E., "Hybrid Architectures in Medical Imaging: Merging CNNs and ViTs," *IEEE Access*, vol. 13, pp. 14200–14215, 2025. doi: 10.1109/ACCESS.2025.3344551.

[11] Raghuvanshi, A. et al., "The Synergy of Local and Global Features in Breast Cancer Detection," *Expert Systems with Applications*, vol. 240, 122415, 2025. doi: 10.1016/j.eswa.2025.122415.

[12] V. R. et al., "Increased Sensitivity to Subtle Malignancies through Hybrid Deep Learning," *Radiology: Artificial Intelligence*, vol. 6, no. 5, e230150, 2024. doi: 10.1148/ryai.240150.

[13] Mokta, T. and Soumma, S., "Localized Saliency in Mammography using Grad-CAM," *International Journal of Computer Assisted Radiology*, vol. 20, no. 3, pp. 445–458, 2025. doi: 10.1007/s11548-025-03123-x.

[14] Khater, M. et al., "Visualizing Attention Rollout in Vision Transformers for Clinical Decision Support," *Medical Image Analysis*, vol. 91, 103010, 2025. doi: 10.1016/j.media.2025.103010.

[15] Kalangi, S. et al., "TCAV: Bridging the Semantic Gap in Neural Networks for Clinicians," *XAI in Medicine*, vol. 4, no. 2, pp. 99–114, 2025. doi: 10.1016/j.xaim.2025.02.008.

[16] Sobhama, P. et al., "Concept Activation Vectors for Radiological Feature Validation," *Journal of Digital Imaging*, vol. 37, no. 6, pp. 1280–1295, 2024. doi: 10.1007/s10278-024-00987-y.

[17] Manikandan, K. et al., "Transparency in CAD Systems: A New Paradigm for Clinical Safety," *Healthcare Technology Letters*, vol. 12, no. 1, pp. 22–30, 2025. doi: 10.1049/htl2.2025.0012.

[18] Reddy, S. and Deepa, T., "Evaluations of Hybrid CNN-ViT for Early Malignancy Detection," *Computational Biology and Medicine*, vol. 170, 107955, 2025. doi: 10.1016/j.compbiomed.2025.107955.