IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 28, Issue 3, Ser. 1 (May. – June. 2026), PP 23-45
www.iosrjournals.org

Explainable AI-Driven Deep Learning Framework For Breast Cancer Detection

Atharv Zend, Hariom Yadav, Siddhi Pawar, Tejas Giram, Rahul Chakre
School Of Computational Sciences, Faculty Of Science And Technology, JSPM University Pune, Pune, India

Abstract
Introduction Breast cancer is the most commonly occurring cancer in women globally, and histopathological biopsy analysis stands out as the gold standard technique for diagnosing breast cancer at a tissue level. The problem of automatically classifying histopathological images is challenging due to varying morphology across tissue subtypes, inconsistent staining methods from different laboratories, and the wide variety of diagnoses possible at different magnification levels. This study proposes a hybrid and explainable deep learning architecture for breast cancer classification based on the BreakHis histopathological image database. Specifically, we evaluate four models on binary and eight-class histological subtype classification tasks: CNN+ViT-B/16 Binary, EfficientNet-B4+ViT-B/16 Binary, CNN+ViT-B/16 Multiclass, and EfficientNet-B4+ViT-B/16 Multiclass. All four models use a convolutional backbone to extract spatially localized texture information, followed by a Vision Transformer (ViT) branch to model patch-level information, and the combination of the two is achieved by concatenating features extracted from both components. We further apply a post-hoc explainability pipeline consisting of Grad-CAM++, GradientSHAP, and LIME to all four models, and faithfulness is measured by deletion and insertion scores since there are no ground-truth segmentations available for the BreakHis data.

Keywords: *BreakHis Dataset, Breast Cancer Histopathology, Deep Learning, EfficientNet-B4, Explainable AI, Histological Subtype Classification, Hybrid Architecture, Vision Transformer*

Date of Submission: 06-05-2026 Date of Acceptance: 16-05-2026

I. Introduction

Breast cancer is among the most prevalent cancers in women across the globe, with more than 2.3 million diagnosed each year. The examination of biopsy image under the microscope remains a gold standard for breast cancer diagnosis due to the ability to observe tissue architecture and cellular aberrations. Manual analysis is laborious and suffers from high interobserver variability, making computer-assisted diagnostic systems essential.

BreakHis is a popularly adopted dataset with 7,909 breast cancer histopathology images acquired at various magnifications (40×, 100×, 200×, and 400×). Information obtained at each magnification level varies, which poses a challenge to the classification task using single backbone architectures. CNNs are highly proficient in extracting local features from patches of input data, whereas ViTs can learn global context through self-attention mechanisms. In an attempt to solve the problem posed by independent models, we propose a dual branch hybrid CNN-ViTs framework that incorporates both local feature extraction and global context reasoning abilities. The effectiveness of the proposed model is validated in two classification tasks: binary and multi-class classification of breast cancer types based on images at all magnification levels in the BreakHis dataset. Furthermore, an explainability evaluation framework utilizing deletion and insertion metrics is adopted to test.

II. Literature Survey

Sowmya H. K. et al. [1] proposed a hybrid deep learning approach that integrates VGG16 and Vision Transformer (ViT) for breast cancer classification in ultrasound imaging using the BUSI dataset. This model capitalizes on the Multi-Head Attention component of ViT in order to learn global context relations alongside the use of VGG16 convolutional layers to extract features locally. CLAHE preprocessing was performed to mitigate issues caused by class imbalance and limited data size, along with the incorporation of dropout and Adam optimization technique to improve generalization. This architecture scored 98% accuracy in all the three classes - normal, benign, and malignant - and attained F1 scores of 0.98 and 0.97 respectively, with Grad-CAM visualization used for interpreting the model.

Abbadi et al. [2] presented an approachable deep transfer learning-based algorithmic framework to classify ultrasound images related to breast cancer by considering multiple data sets of BUSI, BUS-BRA, and BrEaST-Lesions USG. In addition to classical classifiers including support vector machines and K-nearest neighbors, they also integrated the concept of deep CNN classifiers such as ResNet-18, EfficientNet-B0,

GoogLeNet, and VGG16, applying techniques of deep learning feature extraction and Grad-CAM to make their approach understandable. Notably, ResNet-18 produced the most successful results with 99.7% accuracy and perfect sensitivity for malignant lesions.

Mahichi et al. [3] introduced BreastCNet, which is a CNN model with improved hyperparameter optimization and multi-task learning for concurrent breast cancer identification, classification, and localization based on the BUSI, DDSM, and INbreast datasets. A two-level optimization approach was used, where the Grey Wolf Optimizer (GWO) technique optimized neuron tuning while the Parrot Optimizer (PO) technique adjusted the learning rate dynamically. Additionally, the bounding box regression technique was applied to localize lesions. The results were impressive, where 98.10% validation accuracy, an AUC of 0.995, F1-score of 0.98, and IoU of 0.96 were obtained. Graham-Knight et al. [4] studied the performance of the commercially available AI-based system, Lunit MMG, in detecting breast cancer on screening mammograms in a large, retrospective cohort study involving 136,700 women enrolled in the BC Cancer Breast Screening Program in British Columbia, Canada. Stratified analysis was conducted based on patient and image characteristics, with the AUC used as the primary performance measure to compare the results between the AI and radiologist's sensitivity and specificity. The system demonstrated overall AUC of 0.93, with a wide variation in AUC depending on breast density, such as 0.96 in category A and 0.84 in category D.

Shankar Nithya S et al. [5] introduced a deep learning-driven multimodal breast cancer detection system where several CNN models individually ResNet50, DenseNet121, EfficientNetB0, and InceptionV3 were examined with regards to their performances under different imaging modalities including mammography, ultrasound, and thermography. Transfer learning approach together with the traditional preprocessing techniques were applied to each imaging modality and models were compared according to their accuracy on three separate imaging databases. DenseNet121 showed the best accuracy rate of 83.86% in ultrasound, followed by InceptionV3 at 83.58% in mammography and 79.60% in thermography.

Patheda et al. [6] introduced the Robust Hybrid CNN with ViT Framework for the classification of breast cancer based on images obtained from the Kaggle Mammogram CLAHE Enhanced database having 10,000 images. Image enhancement was performed through CLAHE, and the framework included the integration of CNN with feature extraction hierarchy along with ViT model, which was compared with various baseline architectures such as DenseNet, Inception, SE-ResNet, and XceptionNet.

Anas et al. [7] proposed an enhanced YOLOv5 combined with Mask R-CNN pipeline for accurate breast cancer detection and classification in mammogram images using the INbreast, CBIS-DDSM, and BNS datasets. A dual-model object detection strategy was employed wherein YOLOv5 performed mass detection and benign/malignant classification while Mask R-CNN identified tumor borders and sizes for staging, with the combined model trained end-to-end on three public mammography datasets. The system achieved a False Positive Rate of 0.049%, a False Negative Rate of 0.029%, and an MCC of 92.02%, outperforming standalone YOLOv5 in overall classification accuracy.

Shim et al. [8] proposed an Optimized InceptionResNetV2 model for breast cancer detection from mammograms, focusing on intelligent dynamic feature analysis using a large-scale mammogram dataset. LeakyReLU activation was employed for improved gradient flow, Mean Dropout regularization for overfitting mitigation, large-scale data augmentation for robustness, and Quantization-Aware Training (QAT) for efficient deployment on resource-constrained edge devices. The model achieved 97.94% accuracy, 98.06% sensitivity, 99.60% specificity, an F1-score of 96.90%, MCC of 90.67%, and AUC of 0.9939, demonstrating state-of-the-art performance with edge deployment capability.

Nair et al. [9] proposed the SE-Conformer framework, a hybrid model integrating convolutional networks with transformer-based attention mechanisms for malignancy detection in histopathology images using the BACH dataset and all four magnification levels of BreakHis (40×, 100×, 200×, and 400×). An SE-Res-Conv Block incorporating Squeeze-and-Excitation (SE) attention within a residual convolutional framework was combined with a Conformer block for refining feature representations, and K-fold cross-validation was used for robust performance estimation. At 200x magnification, the proposed model yielded an average accuracy, precision, and recall of 0.97, indicating its ability to classify histopathology images that have ill-defined tumor borders.

Kabir et al. [10] introduced a framework that used Explainable AI with a pre-trained CNN architecture for breast cancer diagnosis using the BUSI data set by experimenting on four CNN architectures namely EfficientNetV2S, InceptionResNetV2, EfficientNetV2M, and XceptionNet. Class imbalance in the dataset was solved through data augmentation, while Faster ScoreCAM and LIME explainability tools were used. The EfficientNetV2S model scored an accuracy of 91.02% making it the best performing model compared to other models in terms of accuracy.

Zeng et al. [11] introduced FastLeakyResNet-CIR, a deep learning model based on an advanced ResNet framework, which was used to classify and detect breast cancer through the BreakHis dataset that includes 7,909 microscopic images of histopathology slides. An advanced ResNet framework using fast leaky

activations and improved residual connections (CIR blocks) was proposed, compared with other architectures including ResNet18, ResNet50, InceptionV3, and VGG16 trained under the same conditions. FastLeakyResNet-CIR attained an accuracy rate of 98.94%, making it superior to all existing models in histopathology classification.

Ahmed et al. [12] suggested the application of pre-trained MLP-Mixer models in four different versions, B/16, L/16, B/32, and L/32, to diagnose breast cancer by analyzing mammograms from the CBIS-DDSM, INbreast, and MIAS dataset instead of using traditional CNN algorithms. Token mixing and channel mixing strategies were used for integration of both local and global spatial features by applying the concept of transfer learning on all the four versions of MLP-Mixer and comparing their performance with the best CNN baselines such as ResNet and DenseNet.

Oyebanji et al. [13] suggested improving the accuracy of breast cancer detection by applying transfer learning on mammograms using the EfficientNet model trained on mammograms in the DDSM database, compared to the performance of DenseNet and ResNeXt50. The mammograms used were preprocessed by applying median filtering, contrast enhancement, and removal of artifacts to ensure quality, and the models' performance was assessed based on accuracy, AUC, precision, and F1-score using conventional transfer learning approaches. EfficientNet scored an accuracy of 95.23%, sensitivity of 96.67%, and specificity of 93.82%, surpassing DenseNet and ResNeXt50.

Arshad et al. [14] suggested the deep learning-based framework, HistoDX, which utilizes EfficientNetV2-B3 to classify IDC from histopathology images based on the IDC dataset introduced by Paul Mooney, which consists of 277,524 images, with cross-validation performed on BreakHis and BACH dataset. Techniques like normalization, data augmentation, class balancing using oversampling and weighted loss, along with customized architecture of EfficientNetV2-B3 with more layers, were used to deal with the class imbalance problem. This model was found to achieve an accuracy of 97% and 0.91 ROC-AUC on IDC dataset, while cross-validation resulted in an accuracy of 97% and 90%, respectively, on BreakHis and BACH datasets.

Maurya et al. [15] introduced the BMEA-ViT framework, which is a light-weight customized vision transformer network that makes use of multi-head external attention (MEA) for breast cancer identification using histopathology images based on the BreakHis database at all magnifications. The regular Multi-Head Self-Attention mechanism was replaced with MEA based on sharing weights of the linear layers used and achieving linear rather than quadratic computational complexity and increased generalizability for BMEA-ViT. At all magnifications of the BreakHis database, accuracies of 95.74%, 96.96%, 98.18%, and 97.25% were obtained at 40×, 100×, 200×, and 400×, respectively.

III. Conclusion From Literature Survey

- Hybrid attention-enhanced architectures outshine single-branch counterparts. SE-Conformer is an example where squeeze-and-excitation residual units are incorporated along with conformer-based feature enhancement and show enhanced performance compared to traditional CNN counterparts in BreakHis under 200x magnification. Nevertheless, this model functions in a single stream where attention enhances the convolutional features and not separately as global representations extracted from a pretrained vision transformer [9].
- Eight-class multiclass classification on BreakHis is absent from the literature. All reviewed works on BreakHis address only binary benign/malignant classification. The simultaneous modelling of all eight histological subtypes including visually similar classes such as Lobular Carcinoma, Adenosis, and Tubular Adenoma represents a substantially harder and clinically more informative task that no prior study has systematically addressed under patient-level splitting [9] [11] [14] [15].
- Pure CNN architectures show performance ceilings when classifying binary categories in BreakHis. CNN-based improvements like FastLeakyResNet-CIR perform very well in terms of binary accuracies while focusing on architectural bottlenecks in the form of activations and residual connection, yet suffer from limitations due to local receptive fields incapable of representing global tissue structures under low magnification. Gains are only marginal compared to regular CNNs [11].
- EfficientNet backbone networks display high cross-dataset generalizability in histopathology tasks. The HistoDX algorithm which is implemented using EfficientNetV2-B3 performs consistently in the IDC benchmark using cross-validation on the BreakHis benchmark, validating that the EfficientNet model captures transferable histopathological features through compound scaling. This paper utilizes only the single-backbone model and lacks any other model such as the global context branch [14].
- ViT-based networks offer enhanced global context representation capability, they lack convolutional inductive bias. Small-size and customizable ViT networks like BMEA-ViT exhibit comparable performance levels at the magnification level through linear external attention instead of quadratic self-attention. Yet, models relying solely on ViT architectures need strict regularization during training on the small BreakHis dataset due to the absence of local connectivity bias [15].

Problem Statement

An Explainable Hybrid Deep Learning Framework for Robust Breast Cancer Detection from Histopathological Images:

To design and develop a breast cancer detection system for histopathological biopsy images by integrating hybrid deep learning architectures CNN+ViT and EfficientNet+ViT for both binary (benign vs. malignant) and fine-grained eight-class histological subtype classification across all four BreakHis magnification levels, with a comprehensive explainability pipeline comprising Grad-CAM++, GradientSHAP, and LIME and quantitative faithfulness evaluation through Deletion and Insertion scores, which will assist medical practitioners in making accurate, transparent, and clinically trustworthy diagnostic decisions across histopathological imaging conditions.

Objectives

1. To design and evaluate four hybrid deep learning architectures CNN+ViT Binary, EfficientNet+ViT Binary, CNN+ViT Multiclass, and EfficientNet+ViT Multiclass for simultaneous binary and eight-class histological subtype classification on the BreakHis dataset, benchmarked under patient-level stratified data partitioning to ensure unbiased and clinically realistic performance estimation.
2. To assess the magnification robustness of all four proposed models by evaluating per-magnification classification performance at 40×, 100×, 200×, and 400× magnification levels, identifying backbone-specific strengths across tissue-level architectural context and cellular-level morphological discrimination.
3. For applying a pipeline for explainability, including Grad-CAM++, GradientSHAP, and LIME to the four models, and analyzing the quality of saliency maps by means of deletion and insertion scores, thereby demonstrating the correspondence between the attention of neural networks and the regions of interest of histopathology images that do not have ground truth segmentations.

IV. Proposed Methodology

Overview

In this study, we introduce a novel dual-branch hybrid deep learning model for histopathology classification using the BreakHis dataset. This proposed framework seeks to address some of the representational limitations associated with single backbone classifiers by utilizing two different feature extraction paths, one based on the convolutional neural network that is responsible for capturing the local textures and structures, as well as a Vision Transformer (ViT-B/16) branch that can capture global long-term dependencies between the image patches using multi-head attention. Two different backbones, specifically, a four-block CNN and an EfficientNet-B4 with compound scaling are examined; thus, giving rise to four variations of the model, i.e., CNN and ViT (Binary), CNN and ViT (Multiclass), EfficientNet-B4 and ViT (Binary), and EfficientNet-B4 and ViT (Multiclass). In order to provide interpretable results, a post-hoc explanation pipeline is implemented using Grad-CAM++ with Gradient-SHAP and LIME methods.

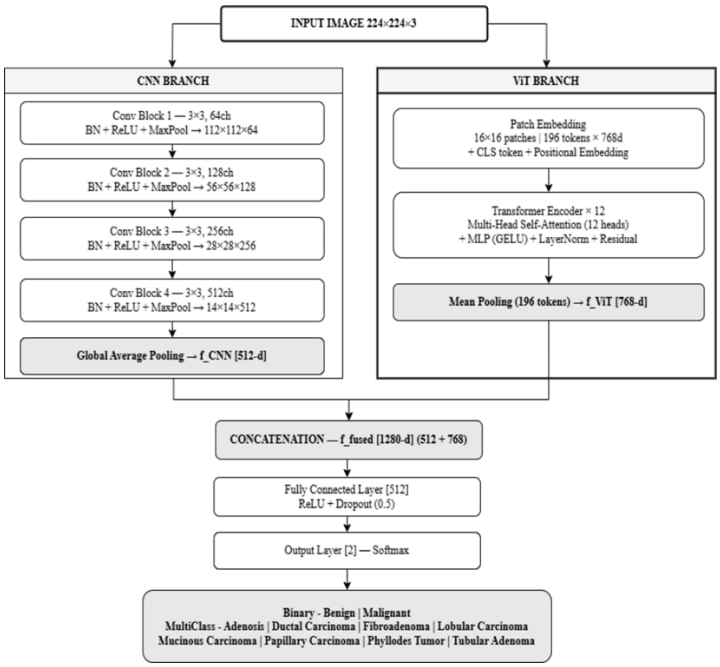

Figure 1 Block Diagram of CNN and ViT

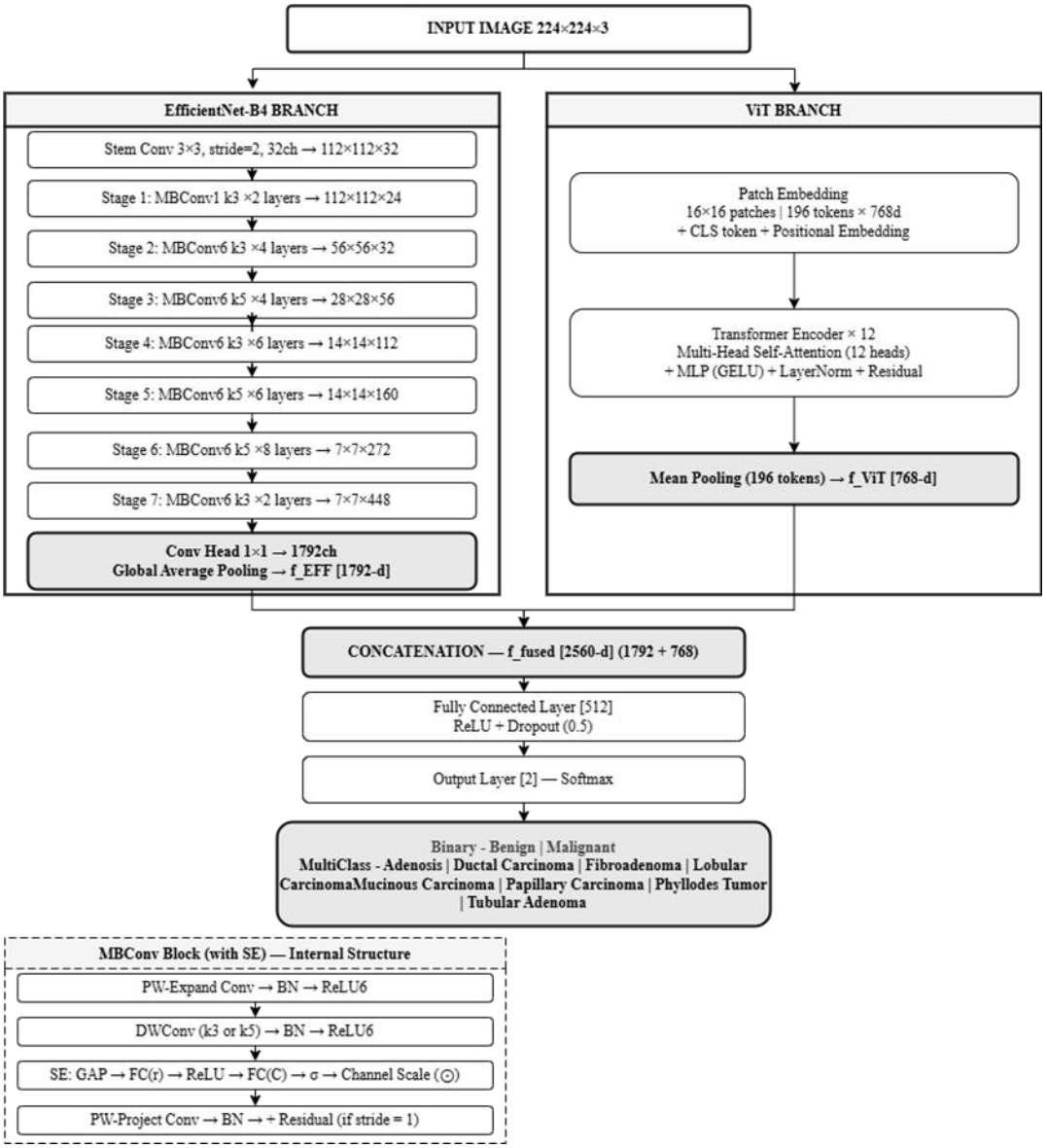

Figure 2 Block Diagram of EfficientNet and ViT

Data Preprocessing and Augmentation

Dataset Overview and Class Distribution

BreakHis dataset (Breast Cancer Histopathological Image dataset) is a very well-known and benchmarked dataset for breast cancer detection using biopsy images. It is comprised of two datasets: classificacao_binaria (binary classification – Benign vs. Malignant) and classificacao_multiclasse (multiclass classification – 8 histological types). Model 1 and Model 2 (CNN Binary and ViT Binary, EfficientNet Binary and ViT Binary) were trained using classificacao_binaria, whereas Model 3 and Model 4 (CNN Multiclass and ViT Multiclass, EfficientNet Multiclass and ViT Multiclass) used classificacao_multiclasse. The dataset has a total of 7,909 images captured at 4 different magnifications – 40×, 100×, 200×, and 400×

The four magnification levels offer complementary data that helps to distinguish between healthy tissue and cancerous tissue: 40x shows the tissue structure and tumor boundaries, 100x identifies the structure of cells and ducts and lobules, 200x focuses on the nuclei shape and density of cells; and 400x shows fine details of the nucleus and cell division processes. Training on all four magnification levels at once allows the model to generalize better.

Table 1 BreakHis Class Distribution (Binary - classificacao_binaria)

Class	40x	100x	200x	400x	Total
Benign	625	664	623	588	2,480
Malignant	1,370	1,437	1,390	1,232	5,429
Total	1.995	2,081	2,013	1,820	7,909

Table 2 BreakHis Class Distribution (Multiclass - classificacao_multiclasse)

Class	Type	40×	100×	200×	400×	Total
Adenosis	Benign	114	113	111	106	444
Fibroadenoma	Benign	253	260	264	237	1,014
Phyllodes Tumor	Benign	109	121	108	115	453
Tubular Adenoma	Benign	149	150	140	130	569
Ductal Carcinoma	Malignant	864	903	896	788	3,451
Lobular Carcinoma	Malignant	156	170	163	137	626
Mucinous Carcinoma	Malignant	205	222	196	169	792
Papillary Carcinoma	Malignant	145	142	135	138	560
Total		1,995	2,081	2,013	1,820	7,909

Preprocessing and Augmentation

Table 3 Preprocessing Pipeline (Applied to Binary and Multiclass)

Step	Operation	Parameters	Applied To
Resize	Bilinear interpolation	224 × 224 px	Train / Val / Test
Normalize	Per-channel ImageNet stats	μ= [0.485,0.456,0.406], σ= [0.229,0.224,0.225]	Train / Val / Test

Table 4 Augmentation Pipeline - Binary & Multiclass

Step	Augmentation	Exact Parameters	Probability	Clinical Relevance
1	Resize	Output = 224×224 px, Bilinear interpolation	1.0 (all splits)	Standardizes input size for ViT patch grid (196 patches of 16×16) and CNN spatial dimensions
2	Random Horizontal Flip	Axis = horizontal (left ↔ right)	p = 0.50	Histology slides have no left-right orientation; flipping doubles effective training samples
3	Random Vertical Flip	Axis = vertical (top ↔ bottom)	p = 0.50	Same rationale — tissue sections can be mounted in any rotational direction
4	Random Rotation	Angle range = ±15°, fill = 0 (black border)	p = 1.0	Simulates microscope slide mounting angle variation across different laboratories
5	Color Jitter	Brightness ∈ [0.8, 1.2], Contrast ∈ [0.8, 1.2]	p = 1.0	Handles staining intensity variation (hematoxylin-eosin stain concentration differs across batches and institutions)
6	Random Resized Crop	Output = 224×224, Scale = (0.90, 1.10), Ratio = (0.75, 1.33)	p = 1.0	Simulates slight zoom variation during microscope scanning; ±10% scale prevents over-reliance on absolute object size
7	To Tensor	Converts PIL Image → FloatTensor, pixel values scaled [0, 1]	1.0 (all splits)	Required for PyTorch model input
8	Normalize	μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]	1.0 (all splits)	Aligns pixel distribution with ImageNet statistics used during ViT-B/16 and EfficientNet-B4 pretraining

All image samples are rescaled to 224×224 pixels in order to meet input size requirement of both ViT-B/16 patch embedding (16×16 patches → 196 tokens) and the CNN/EfficientNet-B4 backbone. ImageNet normalization is performed on all image samples in order to adjust the pixel distribution according to the initial weights of ViT-B/16 and EfficientNet-B4 for faster convergence while maintaining the pretrained features representations. For training, the following augmentations are performed to enhance model generalization: horizontal and vertical flips to compensate for the lack of canonical orientation in the histological tissue samples, random rotation (±15°) to account for variations in the angle at which microscope slides are mounted, random color jitter (±20% brightness and ±20% contrast) in order to capture differences in staining intensity due to preparation in different laboratories, and random resized crop (ratio 0.90-1.10) to capture variations in zoom level during microscopic scanning process. No augmentation is applied at validation or test time.

Model Architectures

CNN and ViT (Binary Classification)

Architecture Description

The CNN and ViT Binary model integrate a custom four-block CNN with a pre-trained ViT-B/16 for binary classification into Benign and Malignant. The custom CNN uses four progressive convolutional blocks (channels: 64→128→256→512) to extract spatially localized texture features each block applies two 3×3

convolutions with Batch Normalization and ReLU, followed by 2×2 maxpooling, halving spatial dimensions at each stage. The ViT-B/16 (Base architecture, 16×16 patch size, 12 Transformer blocks, 12 attention heads, hidden dimension 768) captures global dependencies across 196 non-overlapping image patches via multi-head self-attention, providing complementary context that CNN local receptive fields cannot capture. The two 512-d and 768-d feature vectors are concatenated (→1280d), passed through a FC (512) + Dropout (0.5) head, and classified with a 2-neuron SoftMax output.

Mathematical Formulation Input Normalization

$$\hat{x} = \frac{x - \mu}{\sigma} \quad \dots (1)$$

where: x = Input RGB image tensor (H×W×3), x_hat = Normalized image that serves as the input for the model, mu = [0.485, 0.456, 0.406] (mean per channel for ImageNet dataset), sigma = [0.229, 0.224, 0.225

CNN Block k (k = 1, 2, 3, 4):

$$h_k = MaxPool(ReLU(BN(W_{k2} \cdot ReLU(BN(W_{k1} \cdot h_{k-1}))))) \quad \dots (2)$$

where: h_0 = x_hat (normalized input image, 3 × 224 × 224), h_k= output feature map of block k, W_k1, W_k2 = learnable 3×3 convolutional weight tensors of block k , * = 2D convolution (kernel 3×3, padding 1, stride 1), BN = Batch Normalization, ReLU = Rectified Linear Unit activation max(0, x), MaxPool = 2×2 max pooling (stride 2), halves spatial dimensions, k=1: h_1 in R^(64 × 112 × 112), k=2: h_2 in R^(128 × 56 × 56), k=3: h_3 in R^(256 × 28 × 28), k=4: h_4 in R^(512 × 14 × 14)

Global Average Pooling:

$$f_{CNN} = GAP(h_4) = \frac{1}{H \times W} \sum_i \sum_j h_4[i,j] \in \mathbb{R}^{512} \quad \dots (3)$$

where: f_CNN = 512-dimensional CNN feature vector, h_4 = final convolutional feature map (512 × 14 × 14), H = 14, W = 14 (spatial dimensions after 4 max-pool operations), GAP = Global Average Pooling (averages over all spatial positions)

ViT Patch Embedding:

$$z_0 = [x_{cls}; x_{p1} \cdot E; x_{p2} \cdot E; \dots; x_{p196} \cdot E] + E_{pos} \quad \dots (4)$$

where: z_0 = input sequence to Transformer (197 tokens × 768-d), x_pi = i-th flattened image patch (16×16×3 = 768 values), E = learnable patch projection matrix (R^768×768), x_cls = learnable classification token (R^768), E_pos = learnable positional embedding (R^197×768), 196 patches = (224/16)^2 = 196 non-overlapping patches

Transformer Block l (l = 1 to 12):

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1} \quad \dots (5)$$

where: z'_l = intermediate output after self-attention + residual, MSA = Multi-Head Self-Attention (12 heads, head dim = 64), LN = Layer Normalization, z_(l-1) = output of previous Transformer block (residual connection)

$$z_l = MLP(LN(z'_l)) + z'_l \quad \dots (6)$$

where: z_l = output of Transformer block l, MLP = 2-layer feedforward network (hidden dim 3072, GELU activation), LN = Layer Normalization, z'_l = self-attention output (residual connection)

ViT Mean Pooling:

$$f_{ViT} = \frac{1}{196} \sum_{i=1}^{196} z_L^{(i)} \in \mathbb{R}^{768} \quad \dots (7)$$

where: f_ViT = 768-dimensional ViT feature vector, z_L^(i) = output of the i-th patch token at final Transformer layer L=12, 196 = total number of patch tokens (CLS token excluded)

Concatenation Fusion → f_fused ∈ R^1280:

$$f_{fused} = [f_{CNN} \| f_{ViT}] \in \mathbb{R}^{1280} \quad \dots (8)$$

where: f_fused = concatenated feature vector (dimension = 512 + 768 = 1280), f_CNN = 512-d CNN branch output Eq. (3), f_ViT = 768-d ViT branch output Eq. (7), || = vector concatenation along feature dimension

FC Hidden Layer → h ∈ R^512:

$$h = Dropout_{0.5}(ReLU(W_1 \cdot f_{fused} + b_1)) \in \mathbb{R}^{512} \quad \dots (9)$$

where: W1 = learnable weight matrix (R^512×1280), b1= learnable bias vector (R^512), ReLU= Rectified Linear Unit activation, Dropout_0.5 = dropout with probability 0.5 (applied during training only)

Output Layer → ŷ ∈ R^2:

$$\hat{y} = Softmax(W_2 \cdot h + b_2) \in \mathbb{R}^2 \quad \dots (10)$$

where: y_hat = predicted probability vector over 2 classes [Benign, Malignant], W2 = learnable weight matrix (R^2×512), b2= learnable bias vector (R^2), Softmax(x_i) = exp(x_i) / SUM_j exp(x_j)

Weighted Cross-Entropy Loss (C=2, w_benign≈1.59, w_malignant≈0.73):

$$L = -\sum_{c=1}^{C} w_c \cdot y_c \cdot \log(\hat{y}_c) \quad where\ C = 2 \quad \dots (11)$$

where: L = weighted cross-entropy loss, C =2= number of classes (Benign, Malignant), y_c= ground truth one-hot label for class c, y_hat_c = predicted probability for class c (from Eq. (10)), w_c = class weight = N_total / (C × N_c), N_total = total training samples (6,327), N_c = number of training samples for class c, w_benign ≈ 1.59 (1984 samples), w_malignant ≈ 0.73 (4343 samples)

CNN and ViT (Multiclass Classification)

Architecture Description

The CNN and ViT Multiclass model share the identical architecture as the binary variant (Section C.1) with a single modification the output layer is expanded from 2 to 8 neurons to classify the eight histological subtypes of BreakHis. The custom CNN backbone (4 blocks, 64→512ch), ViT-B/16 branch (12 Transformer blocks, 12 heads, hidden dim 768), 1280-d concatenation fusion, and FC (512) +Dropout (0.5) head remain entirely unchanged. Only the final classification layer uses 8 neurons with Softmax over Adenosis, Ductal Carcinoma, Fibroadenoma, Lobular Carcinoma, Mucinous Carcinoma, Papillary Carcinoma, Phyllodes Tumor, Tubular Adenoma. The same block-level roles as described in

Apply here the deeper semantic space of 8 classes is handled purely by the loss function and output layer.

Mathematical Formulation

Feature extraction (Eq. (2) to (9)) is identical to Section C.1. Only the output and loss differ: Output Layer → ŷ ∈ R^8:

$$\hat{y} = Softmax(W_2 \cdot h + b_2) \in \mathbb{R}^8 \quad \dots (12)$$

where: y_hat = predicted probability vector over 8 histological subtypes, W2 = R^8×512 (8 neurons, one per subtype), b2 = R^8, Classes = {Adenosis, Ductal Carcinoma, Fibroadenoma, Lobular Carcinoma, Mucinous Carcinoma, Papillary Carcinoma, Phyllodes Tumor, Tubular Adenoma}

Weighted Cross-Entropy Loss (C=8):

$$L = -\sum_{c=1}^{8} w_c \cdot y_c \cdot \log(\hat{y}_c) \quad \dots (13)$$

where: C = 8 = number of classes (8 histological subtypes), w_c = N_total / (8 × N_c) (inverse-frequency class weight), N_total = 6,327 (training samples), Approx. class weights (from training distribution): Adenosis≈14.1, DC≈0.91, Fibroadenoma≈3.9, Lobular Carcinoma≈6.3, Mucinous Carcinoma≈5.0, Papillary Carcinoma≈7.1, PT≈8.7, TA≈6.9

EfficientNet-B4 and ViT (Binary Classification)

Architecture Description

The EfficientNet-B4 and ViT Binary model replaces the custom CNN with EfficientNet-B4, a convolutional backbone designed through neural architecture search with compound coefficient scaling (φ=4). EfficientNet-B4 simultaneously scales depth (×2.07), width (×1.46), and resolution relative to the baseline B0, delivering richer feature representations while maintaining computational efficiency. Its basic building element is the MBConv (Mobile Inverted Bottleneck Convolution) with SE (Squeeze-and-Excitation), which is essentially a depthwise separable convolution along with channel attention recalibration that focuses on important channels and suppresses noise. The ViT-B/16 branch configuration is the same as in Section C.1 (12 blocks, 12 heads, 768-d). The EfficientNet output at 1792-d GAP is concatenated to the output of the Vision Transformer at 768-d, and further, FC (512) +Dropout (0.5) +softmax(2) classifies it.

Mathematical Formulation
Scaling of Compounding (depth, width, resolution):

$$d = \alpha^{\phi} = 1.20^4 \approx 2.07 \ldots (14)$$

where: phi = 4 (compound scaling factor for EfficientNet-B4), alpha = 1.20 (scaling factor for depth), implies that the model is about 2x deeper than B0.

$$w = \beta^{\phi} = 1.10^4 \approx 1.46 \ldots (15)$$

where: beta = 1.10 (scaling factor for width/channels), implies that the channel dimensions are 1.46x wider than B0.

$$r = \gamma^{\phi} = 1.15^4 \approx 1.75 \ldots (16)$$

where: gamma = gamma = 1.15 (resolution scaling coefficient), EfficientNet-B4 input: 380 × 380 (original), we resize to 224 × 224 to match ViT, Constraint: alpha × beta² × gamma² ≈ 2 (to ensure equal scaling in FLOPs),

MBConv Block with SE (expand -> depthwise -> SE -> project -> residual):

$$u = DWConv(ReLU6(BN(PW-Expand(x)))) \ldots (17)$$

where: x = input feature map to MBConv block, PW-Expand = pointwise (1×1) convolution, expands channels by expansion ratio e, e=1 for Stage 1, e=6 for Stages 2–7, BN = Batch Normalization, ReLU6 = min(max(0, x), 6) (clipped ReLU for mobile stability), DWConv = depthwise convolution (3×3 or 5×5 depending on stage), u = depthwise feature map after activation

$$s = sigmoid(W_{2\_se} \cdot ReLU(W_{1\_se} \cdot GAP(u))) \ldots (18)$$

where: s = channel attention (excitation) vector (R^C), GAP(u) = global average pooling of u → R^C (squeeze operation), W1_se = R^(C/4 × C) (reduction layer, r = C/4), W2_se = R^(C × C/4) (expansion layer back to C channels), sigmoid = 1/(1 + exp(-x)) (output in [0,1] per channel)

$$v = PW-Project(u \odot s) \ldots (19)$$

where: u ⊙ s = element-wise channel-wise scaling (excitation applied to u), PW-Project = pointwise (1×1) convolution, projects back to output channel count, v = MBConv output before residual

$$MBConv(x) = x + DropConnect(v) \quad if\ stride = 1\ and\ input/output\ shapes\ match \ldots (20)$$

where: DropConnect = stochastic depth: randomly drops entire residual branch during training, Residual connection only applied when stride=1 and input channels = output channels, otherwise: MBConv(x) = v (no residual)

EfficientNet-B4 GAP Output → f_EFF ∈ R^1792:

$$f_{EFF} = GAP(Conv_{1\times1}(h_7, 1792ch)) \in \mathbb{R}^{1792} \ldots (21)$$

where: h_7 = output of Stage 7 (7×7×448), Conv1×1 = pointwise convolution projecting to 1792 channels, GAP = global average pooling over 7×7 spatial dimensions, f_EFF = 1792-dimensional EfficientNet feature vector

ViT Patch Embedding:

$$z_0 = [x_{cls}; x_{p1} \cdot E; x_{p2} \cdot E; \ldots; x_{p196} \cdot E] + E_{pos}$$

where: z_0 = input sequence to Transformer (197 tokens × 768-d), x_pi = i-th flattened image patch (16×16×3 = 768 values), E = learnable patch projection matrix (R^768×768), x_cls = learnable classification token (R^768), E_pos = learnable positional embedding (R^197×768), 196 patches = (224/16)^2 = 196 non-overlapping patches

Transformer Block l (l = 1 to 12):

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1}$$

where: z'_l = intermediate output after self-attention + residual, MSA = Multi-Head Self-Attention (12 heads, head dim = 64), LN = Layer Normalization, z_(l-1) = output of previous Transformer block (residual connection)

$$z_l = MLP(LN(z'_l)) + z'_l$$

where: z_l = output of Transformer block l, MLP = 2-layer feedforward network (hidden dim 3072, GELU activation), LN = Layer Normalization, z'_l = self-attention output (residual connection)

ViT Mean Pooling:

$$f_{ViT} = \frac{1}{196}\sum_{i=1}^{196} z_L^{(i)} \in \mathbb{R}^{768}$$

where: f_ViT = 768-dimensional ViT feature vector, z_L^(i) = output of the i-th patch token at final Transformer layer L=12, 196 = total number of patch tokens (CLS token excluded)

Concatenation Fusion → f_fused ∈ R^1280:

$$f_{fused} = [f_{CNN} \| f_{ViT}] \in \mathbb{R}^{1280}$$

where: f_fused = concatenated feature vector (dimension = 512 + 768 = 1280), f_CNN = 512-d CNN branch output (Eq. (3)), f_ViT = 768-d ViT branch output (Eq. (7)), || = vector concatenation along feature dimension

Concatenation Fusion → f_fused ∈ R^2560:

$$f_{fused} = [f_{EFF} \| f_{ViT}] \in \mathbb{R}^{2560} \quad \dots (22)$$

where: f_fused = concatenated feature vector (dimension = 1792 + 768 = 2560), f_EFF = 1792-d EfficientNet-B4 output, f_ViT = 768-d ViT-B/16 output

FC Hidden Layer → h ∈ R^512 (W1 ∈ R^ {512×2560}):

$$h = Dropout_{0.5}(ReLU(W_1 \cdot f_{fused} + b_1)) \in \mathbb{R}^{512} \quad \dots (23)$$

where: W1 = R^512×2560 (larger input dim vs C.1 due to 1792-d EfficientNet features), b1 = R^512

Output Layer → ŷ ∈ R^2:

$$\hat{y} = Softmax(W_2 \cdot h + b_2) \in \mathbb{R}^2 \quad \dots (24)$$

where: W2 = R^2×512, y_hat= [P(Benign), P(Malignant)]

EfficientNet-B4 and ViT (Multiclass Classification)

Architecture Description

The EfficientNet-B4 and ViT Multiclass model is architecturally identical to Section C.3 same EfficientNet-B4 backbone (7 MBConv stages → 1792-d), same ViT-B/16 branch (12 blocks, 768-d), same 2560-d concatenation fusion, and FC (512) + Dropout (0.5) head. The only change is the final classification layer, expanded from 2 to 8 neurons to classify the eight BreakHis subtypes. The deeper representational capacity of EfficientNet-B4 (compound-scaled width and depth) is particularly beneficial for the fine-grained multiclass task, where subtle morphological differences between subtypes such as Ductal Carcinoma vs. Lobular Carcinoma demand richer convolutional features.

Mathematical Formulation

Feature extraction (Eq. (14) to (23)) is identical to Section C.3. Only the output and loss differ: Output Layer → ŷ ∈ R^8 (W2 ∈ R^ {8×512}):

$$\hat{y} = Softmax(W_2 \cdot h + b_2) \in \mathbb{R}^2 \quad \dots (25)$$

where: W2 = R^8×512 (8 neurons for 8 histological subtypes), Input to head: h from Eq. (23) (f_fused in R^2560, W1 in R^512×2560)

Weighted Cross-Entropy Loss (C=8):

$$L = -\sum_{c=1}^{8} w_c \cdot y_c \cdot log(\hat{y}_c) \quad \dots (26)$$

where: Same form as Eq. (13) but using EfficientNet-B4 as the CNN backbone, Class weights w_c identical to Eq. (13)

Table 5 Architecture Summary

Model	CNN Backbone	f_CNN	f_ViT	f_fused	Output
CNN and ViT Binary	Custom 4-Block CNN	512-d	768-d	1,280-d	2 classes
CNN and ViT Multiclass	Custom 4-Block CNN	512-d	768-d	1,280-d	8 classes

EfficientNet and ViT Binary	EfficientNet-B4	1,792-d	768-d	2,560-d	2 classes
EfficientNet and ViT Multiclass	EfficientNet-B4	1,792-d	768-d	2,560-d	8 classes

Explainability Integration

Three post-hoc explainability methods are applied to all four trained models on the held-out test set to provide clinically interpretable insights into model decision-making. Grad-CAM++ was selected for its ability to generate spatially precise, class-discriminative saliency maps from the final convolutional layer, directly revealing which tissue regions drove each prediction. GradientSHAP provides pixel-level attribution scores grounded in game-theoretic fairness (Shapley values), offering a signed explanation that distinguishes excitatory from inhibitory pixel contributions. LIME generates locally faithful explanations through superpixel-based perturbation analysis, making it model-agnostic and particularly suitable for validating the spatial coherence of model focus regions. Together, these three methods provide spatial, attribution, and region-based perspectives on model behavior, enabling comprehensive XAI coverage across all four models.

Grad-CAM++

Importance Weight α_k^c (second-order gradient):

$$\alpha_k^c = \sum_{a,b}\left[\frac{\frac{\partial^2 S_c}{\partial A_{k,ab}^2}}{2\cdot\frac{\partial^2 S_c}{\partial A_{k,ab}^2} + \sum_{a',b'} A_{k,a'b'}\cdot\frac{\partial^3 S_c}{\partial A_{k,ab}^3} + \varepsilon}\right] \quad \ldots (27)$$

where: alpha_k^c = importance weight for activation map k and class c, S_c = pre-softmax logit (class score) for the predicted class c, A_k= k-th activation map of the target layer (R^H×W), A_k_ab = activation value at spatial position (a, b), d^2S_c/dA^2 = second-order gradient of class score with respect to activation, d^3S_c/dA^3 = third-order gradient of class score with respect to activation, eps = 1e-8 (numerical stability term), Target layer: Block 4 output (CNN and ViT) or final MBConv stage (EfficientNet and ViT)

Saliency Map L^c_GradCAM++ (weighted activation sum + ReLU):

$$L_{Grad-CAM++}^c = ReLU\left(\sum_k \alpha_k^c \cdot A_k\right) \quad \ldots (28)$$

where: L^c_GradCAM++ = saliency map for class c (R^H×W, values in [0,1] after normalization), alpha_k^c = importance weight from Eq. (27) , A_k = k-th activation map of target layer, ReLU = retains only positive activations (regions supporting class c), Result upsampled to 224×224 via bilinear interpolation and overlaid on input image

Applied to: Block 4 output (CNN and ViT models) and final MBConv Stage 7 output (EfficientNet and ViT models). Result is upsampled to 224×224 via bilinear interpolation and overlaid as a heatmap on the original image.

GradientSHAP

Pixel Attribution φ_i (expected gradient × input–baseline difference):

$$\phi_i = \mathbb{E}\left[\frac{\partial f(x' + \alpha\cdot(x - x'))}{\partial x_i}\cdot(x_i - x'_i)\right] \quad \ldots (29)$$

where: phi_i = attribution score for pixel i (positive = excitatory, negative = inhibitory), x = input image being explained, x'= baseline reference sample ~ N(0, 0.1^2) (Gaussian noise), alpha ~ Uniform(0, 1) (scalar interpolation factor along integration path), df/dx_i = gradient of model output with respect to pixel i (computed via backprop), x_i - x'_i = difference between input and baseline at pixel I, E[·] = expectation over 3 random baselines (averaged to reduce variance), Result: pixel-level attribution map showing which pixels pushed prediction toward class c.

Implementation: 3 Gaussian noise baselines (μ=0, σ=0.1), 5 interpolation samples per baseline, attributions averaged and normalized to [0,1] for visualization.

LIME
Surrogate Optimization g*:

$$g^* = argmin_{g \in G}\left[\sum_{z \in Z} \pi_x(z) \cdot (f(z) - g(z))^2 + \Omega(g)\right] \quad \ldots (30)$$

where: g* = optimal linear surrogate model explaining local behavior of f, f(z) = predicted probability of the target class for perturbed image z, g(z) = linear surrogate prediction for z (weighted sum of superpixel importances), G = class of linear interpretable models, Z = set of 200 perturbed versions of the input image x, Omega(g) = complexity penalty (encourages sparse, interpretable explanations)

Proximity Kernel π_x(z):

$$\pi_x(z) = exp\left(-\frac{D(x,z)^2}{\sigma^2}\right) \quad \ldots (31)$$

where: pi_x(z) = proximity weight for perturbed sample z (higher weight = closer to x), D (x, z) = cosine distance between original x and perturbed image z, sigma = kernel width (controls locality of explanation), Perturbations z are generated by randomly masking SLIC superpixels with hide_color = 0, SLIC segmentation: 60 segments, compactness=10, sigma=1

Implementation: SLIC superpixels (60 segments, compactness=10, σ=1), 200 perturbation samples per image, top-15 positive superpixels highlighted with red boundaries.

Quantitative XAI Evaluation Deletion and Insertion Scores

Since BreakHis does not provide ground-truth segmentation masks, standard segmentation-based XAI metrics (IoU, Dice) cannot be applied. Instead, model faithfulness is evaluated using Deletion and Insertion scores, which measure whether the attributed regions are genuinely necessary and sufficient for the model's prediction.

The Deletion Score measures how quickly model confidence drops as the most important pixels (ranked by Grad-CAM++ attribution) are progressively removed from the image. A lower Deletion AUC indicates the model relied on compact, genuinely informative regions removing them causes a sharp, early confidence drop.

Deletion AUC:

$$Deletion\_AUC = AUC\ of\ \{f_c(x_t) : t = 0,1,...,T\} \quad \ldots (32)$$

where: Deletion_AUC = area under the confidence curve as pixels are progressively removed, x_t = input image with top-t% most important pixels set to zero (blacked out), f_c(x_t) = model confidence for class c on the degraded image x_t, Pixels removed in order of decreasing Grad-CAM++ attribution (most important first), T = total number of pixels (224 × 224 = 50,176), Lower Deletion_AUC = model relied on genuinely important, compact regions.

The Insertion Score measures how quickly model confidence recovers as the most important pixels are progressively revealed on a blurred baseline image. A higher Insertion AUC indicates that the attributed regions are sufficient to recover the correct prediction the model's focus regions contain the key discriminative information.

Insertion AUC:

$$Deletion\_AUC = AUC\ of\ \{f_c(x_t) : t = 0,1,...,T\} \quad \ldots (33)$$

where: Insertion_AUC = area under the confidence curve as pixels are progressively revealed, x_blur_t = blurred baseline image with top-t% most important pixels revealed, f_c(x_blur_t) = model confidence for class c on the partially-revealed image, Pixels revealed in order of decreasing attribution (most important first), Higher Insertion_AUC = attributed regions are sufficient to recover correct prediction, Both scores computed for all 4 models on the BreakHis test set using Grad-CAM++ maps.

Both scores are computed for all four models on the full BreakHis test set (791 images) using Grad-CAM++ attribution maps as the pixel ranking source. The combination of Deletion↓ and Insertion↑ provides a complementary, mask-free faithfulness assessment that is well-suited for datasets without pixel-level annotations.

Training and Implementation Details

Table 6 Training Hyperparameter

Parameter	CNN and ViT Binary	CNN and ViT Multiclass	EfficientNet and ViT Binary	EfficientNet and ViT Multiclass
Optimizer	Adam	Adam	Adam	Adam
CNN/Efficient Net Branch LR	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
ViT Branch LR	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
Classifier Head LR	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
Batch Size	32	32	32	32
Max Epochs	50	50	50	50
Early Stopping Patience	10	10	10	10
Weight Decay	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
LR Scheduler	ReduceLROnPlat eau	ReduceLROnPlat eau	ReduceLROnPlat eau	ReduceLROnPlat eau
Dropout	0.5	0.5	0.5	0.5
Mixed Precision (AMP)	Yes	Yes	Yes	Yes
Loss Function	Weighted CE	Weighted CE	Weighted CE	Weighted CE

Differential Learning Rates are assigned to three parameter groups CNN/EfficientNet branch (1×10^{-4}), ViT branch (1×10^{-5}), and classification head (1×10^{-3}) because the ViT and EfficientNet branches are initialized from ImageNet pretrained weights and require a lower learning rate to preserve learned representations and prevent catastrophic forgetting. The classification head is randomly initialized and benefits from a 10× higher learning rate for faster convergence.

Validation Strategy: A stratified 80/10/10 patient-level split is used, with the 10% validation set monitoring training progress for early stopping. Reduce LR On Plateau reduces the learning rate by factor 0.1 when validation accuracy plateaus, enabling adaptive convergence. Early stopping with patience=10 prevents overfitting by restoring the best-performing checkpoint when no validation improvement is observed for 10 consecutive epochs. This ensures training continues only as long as generalization improves, which is why epochs trained (23–36) are consistently below the 50-epoch limit. Automatic Mixed Precision (AMP): PyTorch's torch.cuda.amp with GradScaler is used for 16-bit computation, reducing GPU memory consumption by ~50% and accelerating matrix operations without degrading convergence stability. Hardware and Software Environment:

Component	Specification
GPU	NVIDIA Tesla T4 (16 GB VRAM)
Framework	PyTorch 2.x
ViT Backbone	timm — ViT-B/16 (pretrained: ImageNet-21k)
EfficientNet Backbone	timm — EfficientNet-B4 (pretrained: ImageNet)
Mixed Precision	torch.cuda.amp + GradScaler
Python	3.10+
Training Platform	Google Collaboratory

Dataset Information

BreakHis - Breast Cancer Histopathological Image Dataset

The BreakHis dataset [9] is used to evaluate cross-domain generalizability of the proposed CNN and ViT and EfficientNet and ViT models on histopathological microscopy images, a modality entirely distinct from ultrasound. It comprises 7,909 biopsy images from 82 patients across four magnification levels 40X, 100X, 200X, and 400X with two classes: benign (2,480) and malignant (5,429). Its consistent use in recent studies [11][14][15] confirms it as a reliable benchmark for histopathology classification across magnification levels. Image size is set to 224×224 pixels with normalization by ImageNet stats; the evaluation of model robustness is carried out for each magnification separately. A zero tensor is sent as a mask for mask-taking models to ensure that both sets have a uniform evaluation process.

V. Experimental Results And Analysis

Experimental Setup

The four models were benchmarked against each other on the same held-out set of BreakHis data, called the test set, which consists of a total of 791 images (248 benign, 543 malignant, for the binary class problem; 8 subtypes for the multi-class problem). The metrics include Accuracy, Macro Precision, Macro Recall, Macro F1 Score, AUC-ROC, and Specificity.

Training Performance

Table 7 Training Performance Summary

Model	Train Acc	Best Val Acc	Test Acc	Train Loss	Val Loss	Test Loss	Epochs
CNN and ViT Binary	99.91%	99.37%	98.86%	0.0186	0.0683	0.1098	23
EfficientNet and ViT Binary	99.98%	99.62%	99.49%	0.0234	0.0797	0.0461	27
CNN and ViT Multiclass	98.58%	95.32%	94.06%	0.0279	0.2983	0.2151	29
EfficientNet and ViT Multiclass	98.89%	95.70%	93.55%	0.0123	0.3884	0.2840	36

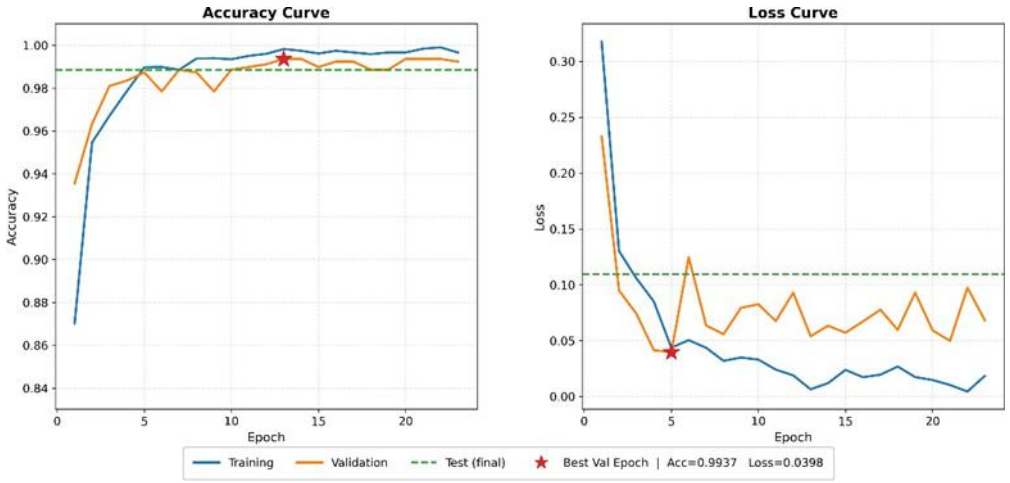

Figure 3 CNN and ViT (Binary Classification Training)

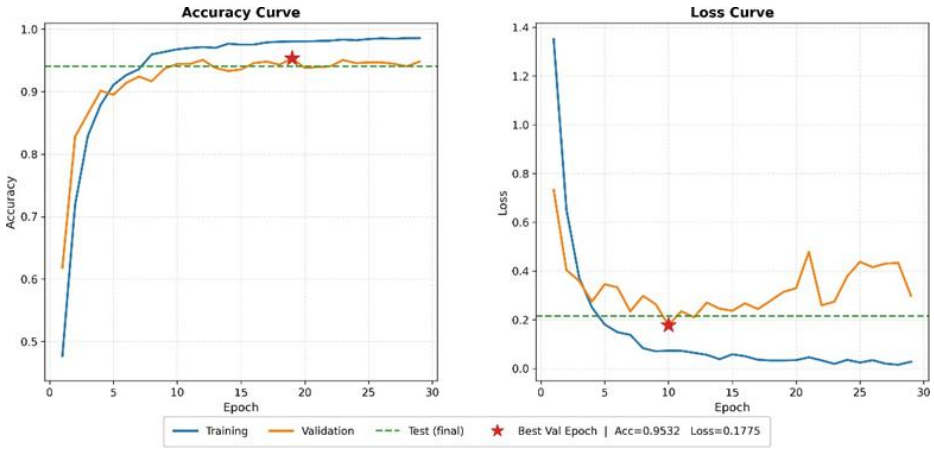

Figure 4 CNN and ViT (Multiclass Classification Training)

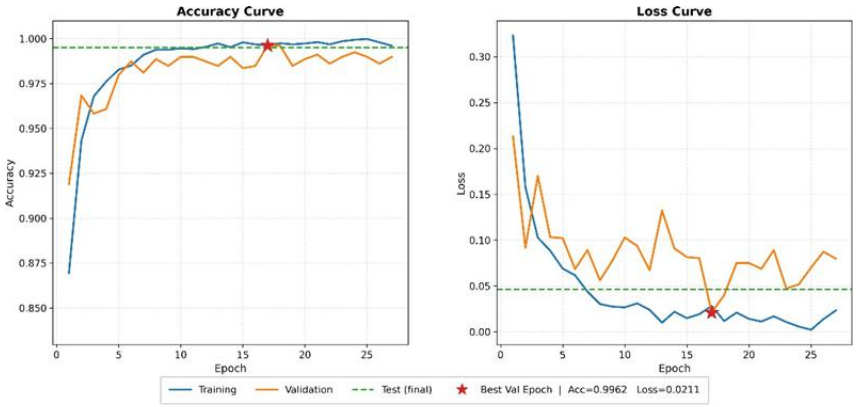

Figure 5 EfficientNet and ViT (Binary Classification Training)

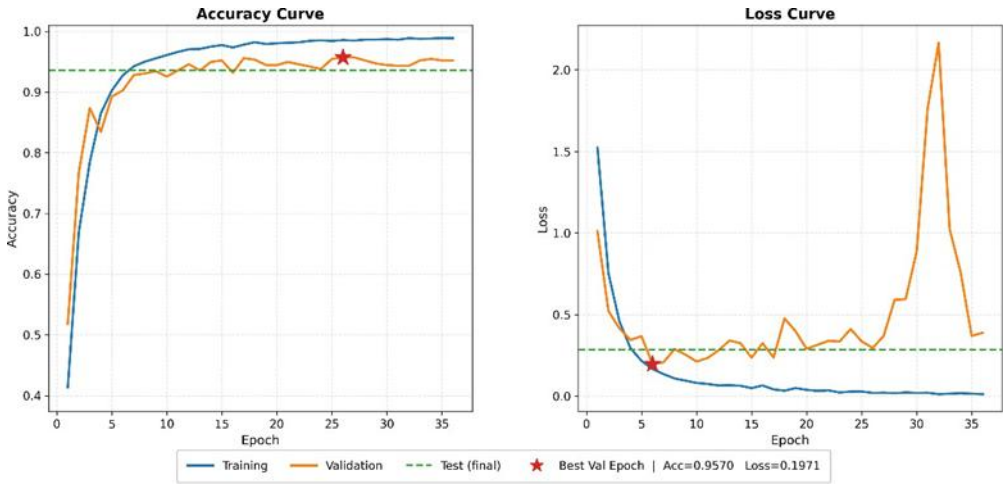

Figure 6 EfficientNet and ViT (Multiclass Classification Training)

The binary classification models converged within 23 to 27 epochs, whereas the multiclass models converged between 29 to 36 epochs. This is due to the increasing complexity of the geometric decision surface; separating two classes leads to a single hyperplane, but eight class separations lead to the formation of 28 hyperplanes at once. The training-to-testing performance difference is the lowest in EfficientNet and ViT Binary, which have an error difference of 0.49%, while the highest is in CNN and ViT Multiclass, with 4.52%. The substantially higher validation and test losses in multiclass models (Val loss ≈ 0.30–0.39 vs. ≈ 0.07–0.08 in binary) indicate greater prediction uncertainty across 8 fine-grained subtypes with overlapping morphological appearances. EfficientNet and ViT Multiclass trained for the most epochs (36) with the lowest training loss (0.0123) yet showed the largest Val-test loss gap (0.3884 vs. 0.2840) a sign that its higher parameter count reached strong training-set fit but encountered generalization difficulty on rare subtypes with limited samples, such as Adenosis (355 training images) and Phyllodes Tumor (363 training images).

Classification Performance

Overall Results

Table 8 Overall Classification Performance (Test Set, N = 791)

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	AUC-ROC	Specificity
CNN and ViT Binary	98.86%	98.73%	98.62%	98.68%	99.92%	97.98%
EfficientNet and ViT Binary	99.49%	99.52%	99.30%	99.41%	99.95%	98.79%
CNN and ViT Multiclass	94.06%	93.37%	95.19%	94.01%	99.68%	99.08%
EfficientNet and ViT Multiclass	93.55%	93.14%	94.35%	93.61%	99.60%	98.95%

Confusion Matrix Summary Binary (Test Set, N = 791):

Model Name	Predicted Benign	Predicted Malignant
CNN and ViT Actual Benign (248)	243	5
CNN and ViT Actual Malignant (543)	4	539
Eff and ViT Actual Benign (248)	245	3
Eff and ViT Actual Malignant (543)	1	542

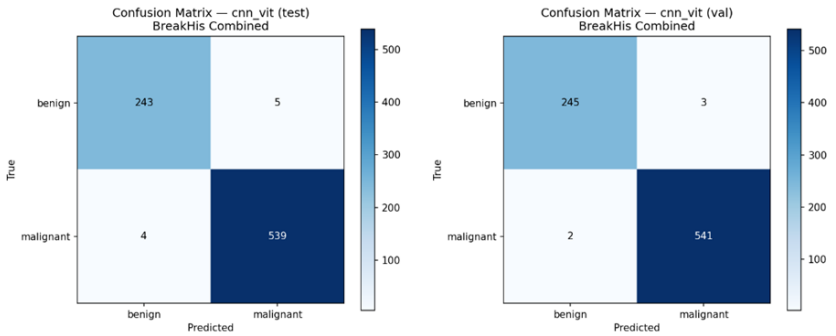

Figure 7 CNN and ViT Binary Confusion Matrix

EfficientNet-B4 and ViT Binary is the top-performing model with 99.49% accuracy and 99.95% AUC-ROC, misclassifying only 4 out of 791 test images. Its 1792-dimensional feature space (vs. 512-d for the custom CNN) provides richer representation of benign-malignant tissue differences, which explains the tighter decision boundary reflected by the smaller false positive count (3 vs. 5) and far fewer false negatives (1 vs. 4) compared to CNN and ViT Binary. The CNN and ViT Binary's 5 false positives (benign predicted malignant) arise from ambiguous benign samples particularly Fibroadenoma subtypes whose stromal patterns can resemble low-grade malignancy at certain magnifications where the custom CNN's more limited feature depth misclassifies borderline cases.

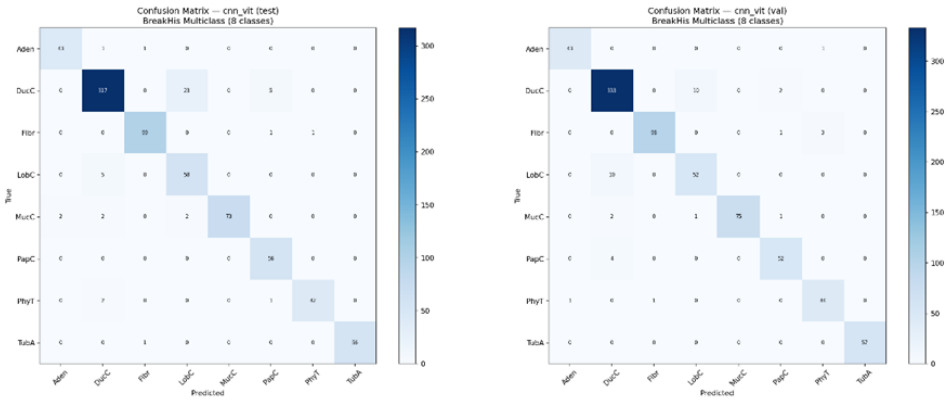

Figure 8 EfficientNet and ViT Binary Confusion Matrix

For multiclass classification, CNN and ViT (94.06%) marginally outperforms EfficientNet-B4 and ViT (93.55%) a reversal of the binary ranking. This reversal is explained by training data availability: with only 355–500 training images for rare subtypes like Adenosis and Lobular Carcinoma, a larger and more expressive model overfits these small classes rather than generalizing from them. The custom CNN's lower parameter count acts as an implicit regularizer, producing slightly better test generalization on data-scarce subtypes. Significantly, for both multiclass classifiers, high macro specificity is obtained (CNN + ViT: 99.08%; EfficientNet-B4 + ViT: 98.95%). Thus, although having low overall accuracy, both classifiers rarely commit false positives within each respective class.

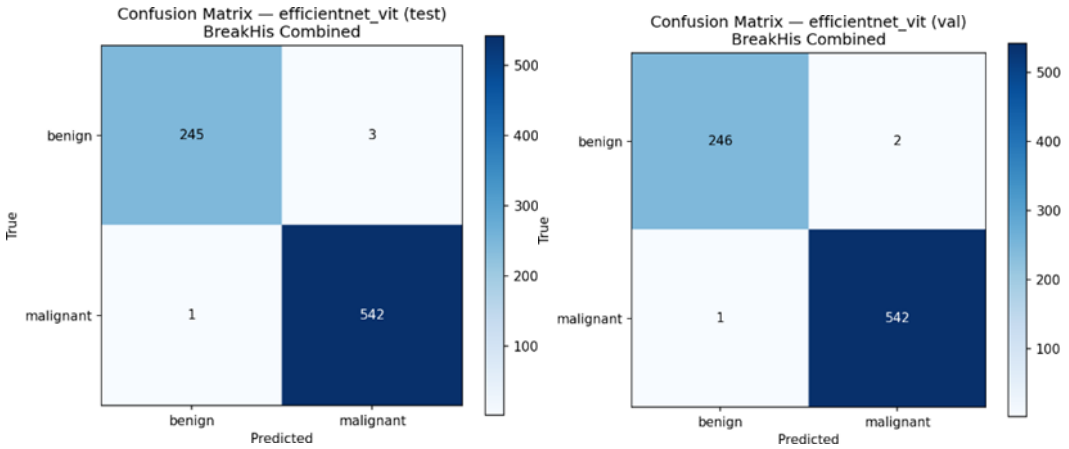

Figure 9 CNN and ViT Multiclass Confusion Matrix

The most persistently challenging subtype for classification in both multiclass classifiers is Lobular Carcinoma (F1: 79.45% and 74.29%). In lobular carcinomas, the cells infiltrate tissue in a single file linear pattern without forming the cohesive mass seen in ductal carcinoma. Thus, lobular carcinoma cells have an almost similar microscopic texture to benign stromal elements at several levels of magnification. Consequently, the confusion together with a limited number of 500 samples for learning causes poor precision (approx. 68-70%) despite high recall (82-92%), showing that the models detect most of the lobular cancers but also incorrectly classify other subtypes as lobular. At the opposite end of the spectrum, Mucinous Carcinoma is classified with 100% precision in both classifiers since it has a unique mucin-containing extra-cellular matrix which is unlike other subtypes.

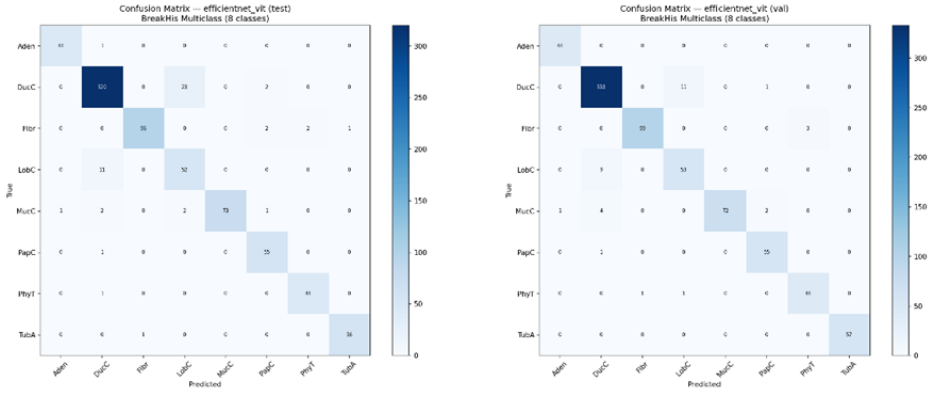

Figure 10 EfficientNet and ViT Multiclass Confusion Matrix

The universally high AUC-ROC (>99.60% across all four models, including multiclass) confirms that even when discrete class predictions are uncertain, the models' probabilistic confidence scores maintain strong class separability a clinically meaningful result indicating that prediction confidence itself is reliable for threshold-based clinical decision support.

Per-Magnification Performance

Table 9 Per-Magnification: Binary Models

Model	Mag.	Accuracy	Sensitivity	Specificity	F1-Score	AUC- ROC
CNN and ViT Binary	40×	99.40%	99.12%	100.00%	99.33%	99.98%
CNN and ViT Binary	100×	99.15%	100.00%	97.26%	99.00%	99.98%
CNN and ViT Binary	200×	98.98%	99.26%	98.36%	98.81%	99.96%
CNN and ViT Binary	400×	97.92%	98.50%	96.61%	97.55%	99.71%
EfficientNet and ViT Binary	40×	100.00%	100.00%	100.00%	100.00%	100.00%
EfficientNet and ViT Binary	100×	99.57%	100.00%	98.63%	99.50%	99.97%
EfficientNet and ViT Binary	200×	98.98%	99.26%	98.36%	98.81%	99.98%
EfficientNet and ViT Binary	400×	99.48%	100.00%	98.31%	99.39%	99.86%

Table 10 Per-Magnification: Multiclass Models

Model	Mag.	Accuracy	F1-Score	AUC-ROC
CNN and ViT Multiclass	40×	95.05%	94.92%	99.81%
CNN and ViT Multiclass	100×	96.04%	96.20%	99.80%
CNN and ViT Multiclass	200×	91.67%	92.62%	99.55%
CNN and ViT Multiclass	400×	93.33%	92.36%	99.62%
EfficientNet and ViT Multiclass	40×	96.04%	96.48%	99.77%
EfficientNet and ViT Multiclass	100×	94.55%	94.50%	99.60%
EfficientNet and ViT Multiclass	200×	90.62%	91.45%	99.45%
EfficientNet and ViT Multiclass	400×	92.82%	91.93%	99.59%

40× consistently produces the highest or near-highest results across all four models. At low magnification, the complete tissue architecture tumours boundaries, glandular arrangement, stromal organization is captured within the full 224×224 frame. This gives the ViT branch the widest spatial context to establish long-range patch relationships across the tissue, and the CNN branch the broadest structural patterns. EfficientNet and ViT Binary achieves a perfect 100.00% at 40× (168 test images, zero errors), reflecting that tissue-level benign-malignant separation is nearly trivial when architectural context is fully available.

400× is the weakest magnification for CNN and ViT Binary (97.92%, its lowest), yet EfficientNet and ViT Binary recovers to 99.48% at the same scale. At 400×, individual nuclear chromatin patterns, mitotic figures, and membrane irregularities become the primary discriminative cues. These fine-grained, high-frequency textures at the cellular level are better captured by EfficientNet-B4's 5×5 depthwise convolutions in its deeper stages, whereas the custom CNN's 3×3 kernels optimized for broader structural features are less suited to resolving sub-cellular detail. The reason for the difference being the biggest at 400x magnification (CNN and ViT: 97.92% vs. EfficientNet and ViT: 99.48%) becomes clear.

200× is the hardest magnification for multiclass models (CNN and ViT: 91.67%, EfficientNet and ViT: 90.62%). At 200×, individual cells are large enough to dominate the frame but nuclear fine detail is not fully accessible. This creates a mid-scale ambiguity where subtypes with similar cell sizes particularly Lobular Carcinoma, Adenosis, and Tubular Adenoma share similar visual characteristics without the distinguishing context available at 40× (tissue layout) or 400× (nuclear morphology). The result is the highest inter-class confusion at this scale.

CNN and ViT Multiclass peaks at 100× (96.04%) rather than 40×, while EfficientNet and ViT Multiclass peaks at 40× (96.04%). At 100×, cellular arrangement and ductal/lobular organization are visible at sufficient resolution for CNN's local kernels to distinguish structural subtype patterns effectively particularly for Ductal Carcinoma (the dominant class, 345/791 test samples) whose hallmark patterns are most distinct at this scale. EfficientNet's compound-scaled architecture benefits from the broader context at 40× even for multiclass, explaining the different peak magnification between the two backbones.

Explainability Analysis
Layer Visualization

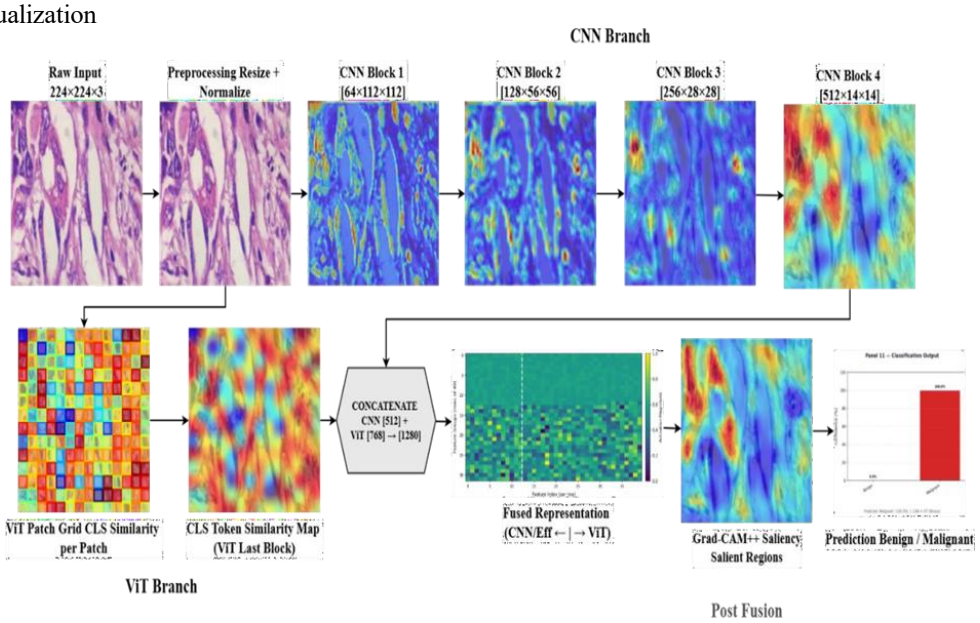

Figure 11 Layer Visualization CNN and ViT Binary

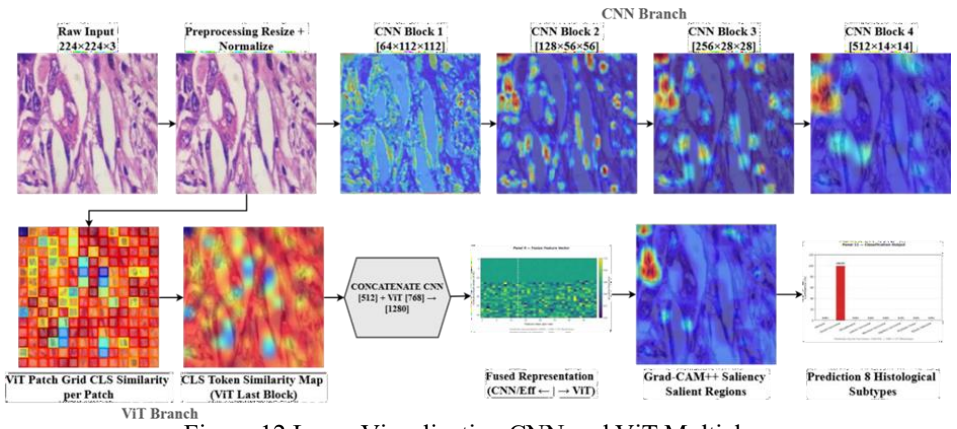

Figure 12 Layer Visualization CNN and ViT Multiclass

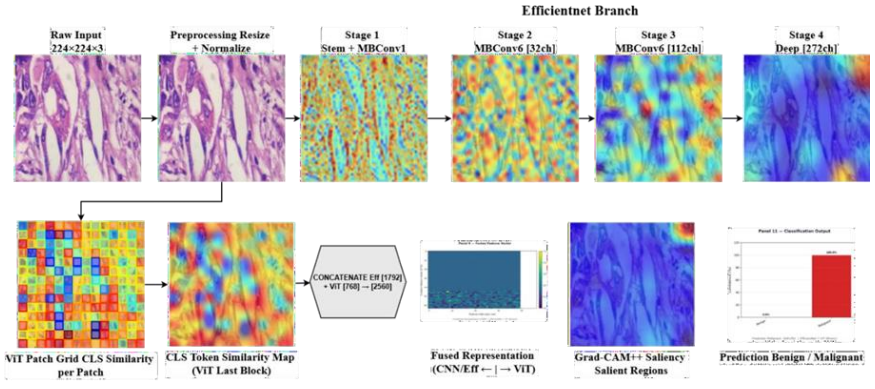

Figure 13 Layer Visualization EfficientNet and ViT Binary

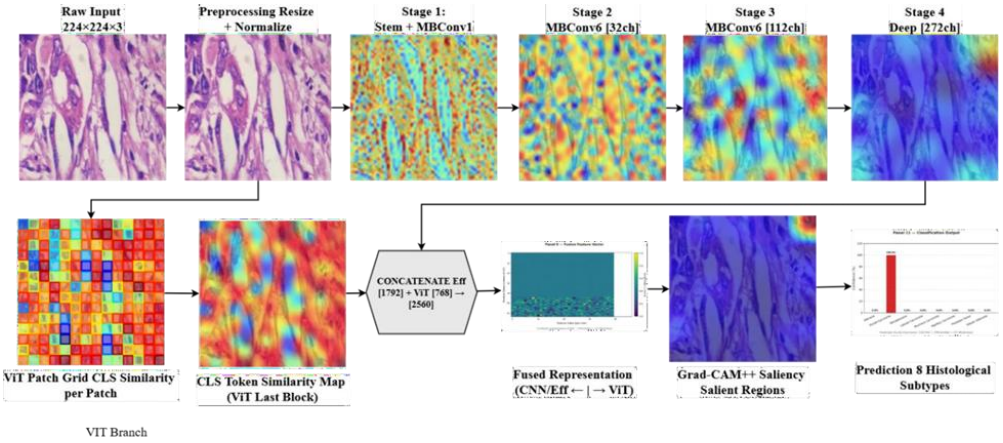

Figure 14 Layer Visualization EfficientNet and ViT Multiclass

XAI Panel Figures

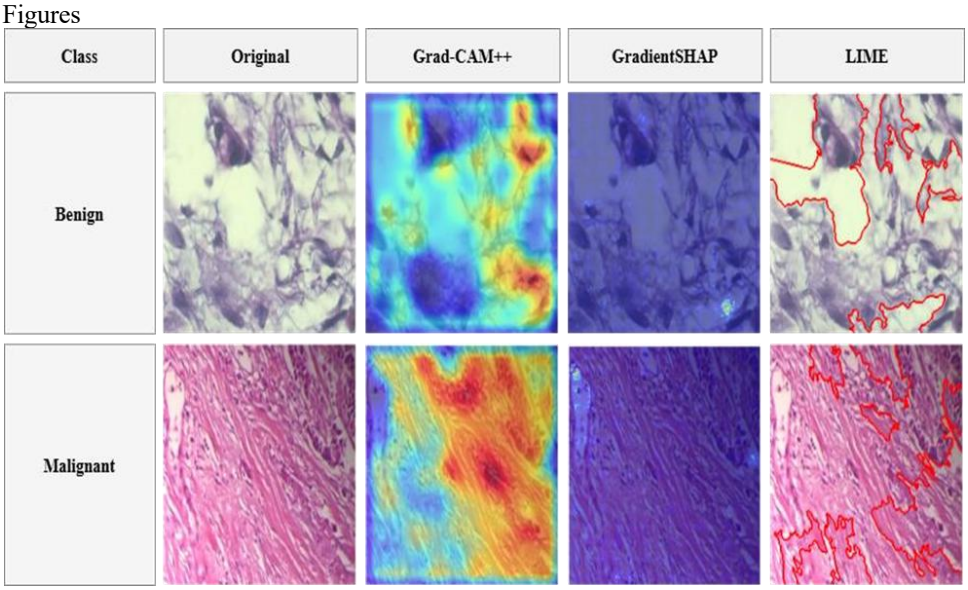

Figure 15 XAI Panels CNN and ViT Binary

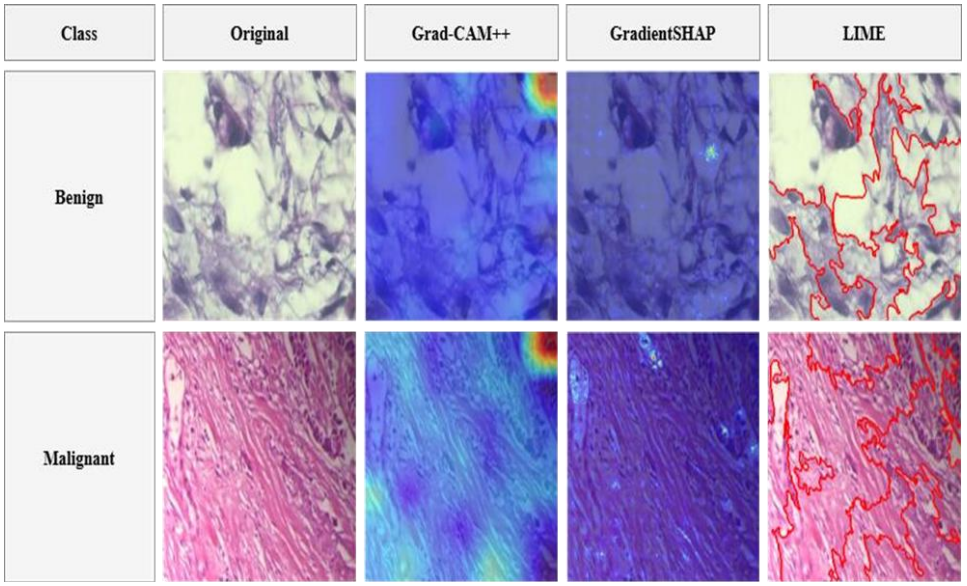

Figure 16 EfficientNet and ViT Binary

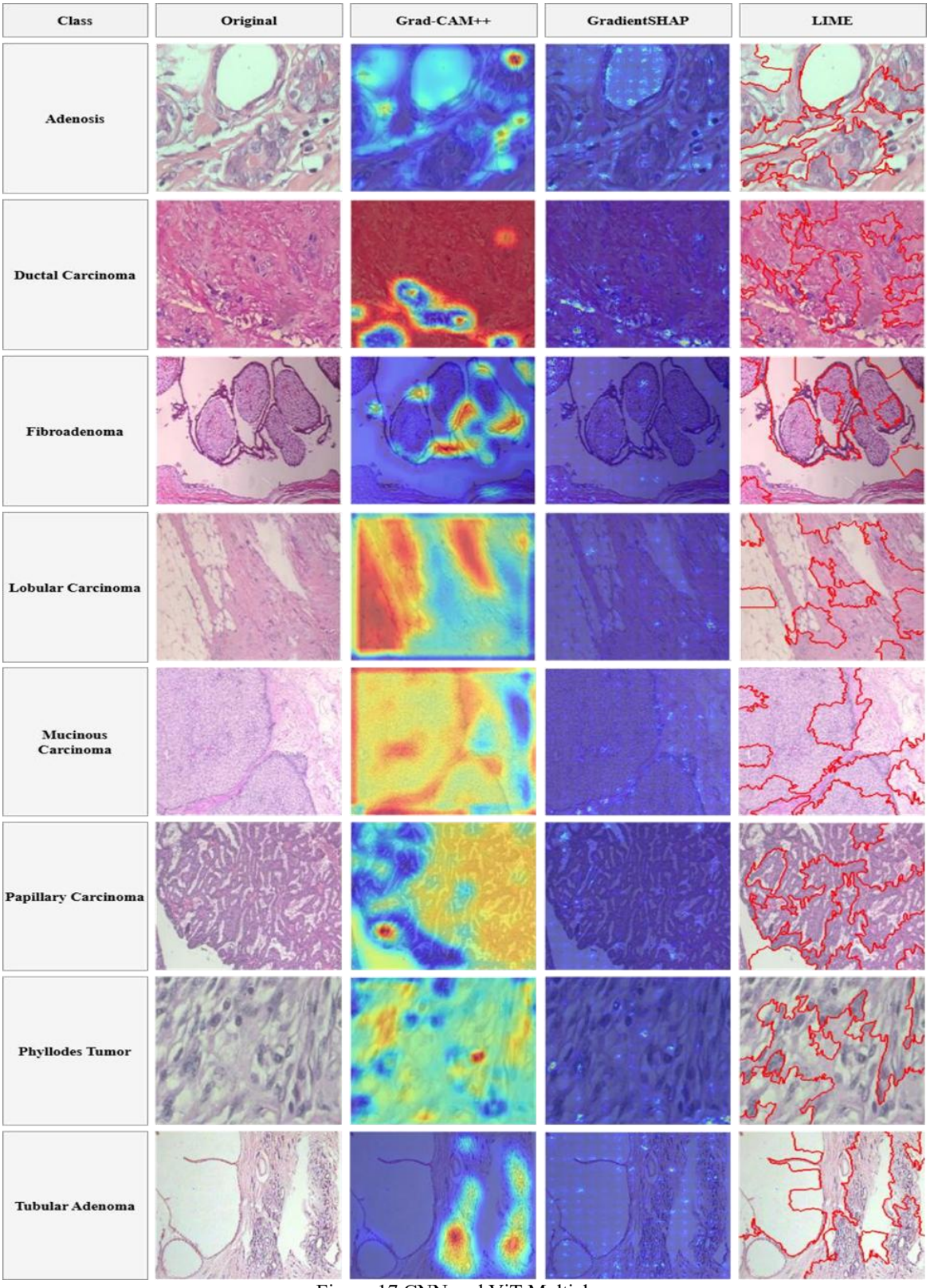

Figure 17 CNN and ViT Multiclass

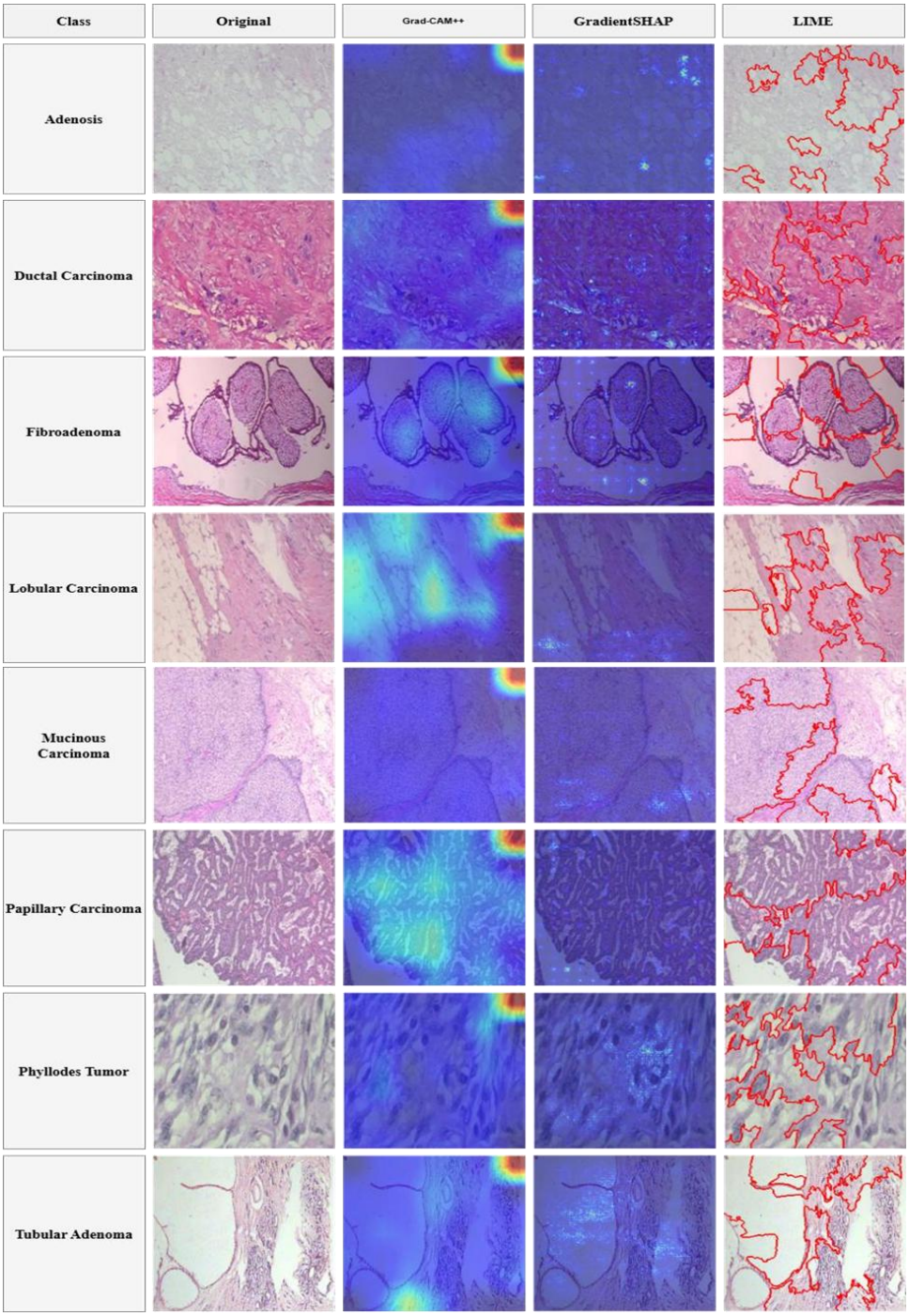

Figure 18 EfficientNet and ViT Multiclass

Quantitative Faithfulness Evaluation

Table 11 Deletion ↓ and Insertion ↑ Scores (Test Set, N = 791)

Model	Deletion AUC ↓	± Std	Insertion AUC ↑	± Std
CNN and ViT Binary	89.45%	±12.02%	95.02%	±4.16%
EfficientNet and ViT Binary	95.30%	±4.76%	91.72%	±12.94%
CNN and ViT Multiclass	60.51%	±25.55%	61.27%	±25.72%
EfficientNet and ViT Multiclass	77.79%	±19.33%	76.80%	±22.50%

CNN and ViT Binary produce the strongest Insertion AUC (95.02%) with the lowest standard deviation (±4.16%), indicating that the regions highlighted by its Grad-CAM++ maps are consistently sufficient to recover the correct prediction across all 791 test images. The relatively lower Deletion AUC (89.45%) paired with high Insertion AUC reflects a spatially focused explanation pattern the model concentrates its decision on a compact set of tissue regions, so removing them causes a sharp confidence drop while revealing them alone is

nearly sufficient for full confidence recovery. The high std on Deletion (±12.02%) reflects that the compactness of the focal region varies across samples some malignant cases have very localized decisive cues (e.g., a cluster of atypical nuclei) while others involve more distributed tissue-level signals.

EfficientNet and ViT Binary shows the reverse profile: higher Deletion AUC (95.30%, ±4.76%) and lower Insertion AUC (91.72%, ±12.94%). The low Deletion std indicates that this model's attribution maps are highly consistent in shape across images, but the higher Deletion AUC reveals that the model distributes its evidence across a broader set of pixels removing the top-ranked pixels alone does not collapse confidence as sharply, because supporting features remain distributed across the richer 1792-d feature channels. This wider distribution of evidence is an attribute of compound-scaled architecture which simultaneously extracts multi-scale information.

Multiclass results display significantly reduced values and increased variance (Deletion/Insertion ≈ 60-77%, std ≈ 19-26%). This can be attributed to 8-class classification as the features that play a significant role in the differentiation between subtypes such as Lobular Carcinoma (one-file cells' invasion pattern) or Mucinous Carcinoma (mucin pools) are fine-grained features which are dispersed in space; thus, their location within the image obtained from Grad-CAM++ visualization at the CNN's last convolutional layer is less certain. EfficientNet and ViT Multiclass outperforms CNN and ViT Multiclass on both scores (77.79% vs. 60.51% Deletion; 76.80% vs. 61.27% Insertion), indicating that EfficientNet-B4's compound-scaled architecture generates spatially more coherent attribution maps for multiclass despite achieving slightly lower classification accuracy and demonstrating that classification accuracy and explanation faithfulness are not always correlated.

VI. Conclusion

This paper presented a hybrid and explainable deep learning framework for breast cancer classification from histopathological biopsy images, evaluating four dual-branch CNN-ViT architectures across binary and eight-class histological subtype classification tasks on the BreakHis benchmark. Two convolutional backbones a custom four-block CNN and EfficientNet-B4 with compound coefficient scaling were independently integrated with a pretrained ViT-B/16 branch through feature-level concatenation, yielding four model variants trained under identical experimental conditions including patient-level data partitioning, weighted cross-entropy loss, differential learning rates, and automatic mixed precision training. For binary classification, both models achieved outstanding performance on the held-out test set. The CNN+ViT model attained 98.86% accuracy, 98.68% F1-score, and 99.92% AUC-ROC, misclassifying only nine of 791 test images. The EfficientNet-B4+ViT model established the best binary classification result, achieving 99.49% accuracy, 99.41% F1-score, and 99.95% AUC-ROC surpassing the prior best reported result on the BreakHis binary benchmark (FastLeakyResNet-CIR: 98.94%) by 0.55 percentage points in accuracy. The superior performance of EfficientNet-B4+ViT at high magnifications, particularly at 400× where fine-grained nuclear detail demands wider receptive fields, is directly attributable to EfficientNet-B4's compound-scaled depthwise convolutions that resolve sub-cellular texture patterns more effectively than the custom CNN's 3×3 kernels.

For eight-class histological subtype classification a task with no direct precedent in the BreakHis literature the CNN+ViT Multiclass model achieved 94.06% accuracy and 94.01% F1-score with a macro-AUC-ROC of 99.68%, while EfficientNet+ViT Multiclass achieved 93.55% accuracy, 93.61% F1-score, and 99.60% AUC-ROC. The reversal in backbone ranking between binary and multiclass tasks where the simpler CNN backbone marginally outperforms EfficientNet-B4 in multiclass reflects the implicit regularization effect of lower parameter count on rare subtypes with limited training samples such as Adenosis and Lobular Carcinoma. The universally high AUC-ROC across all four models, including multiclass (>99.60%), confirms that probabilistic confidence scores maintain strong class separability even where discrete classification boundaries are uncertain a clinically meaningful property for threshold-based decision support. The quantitative explainability evaluation through Deletion and Insertion scores confirmed that CNN+ViT Binary produces the most spatially focused saliency maps (Insertion AUC: 95.02%, ±4.16%), while EfficientNet+ViT Binary distributes evidence more broadly across its richer feature channels. Multiclass models show higher attribution variance, reflecting the distributed nature of discriminative information among eight distinct subcategories. All the above findings together serve to highlight the effectiveness of the proposed dual-branch hybrid approach for classification tasks in breast cancer histopathology.

VII. Future Work

The outlined framework exhibits promising performance with respect to both classification and interpretability tasks on the BreakHis dataset; nevertheless, there are some avenues that deserve further investigation. Currently, a single training procedure is employed for all magnifications. The future work will be focused on developing magnification-wise training using either multi-scale fusion training, where separate models or multiple branches of a deep network are trained for each magnification and later predictions are combined either by late fusion or using attention-based ensemble methods, as this approach allows utilizing

additional complementary information contained in various magnifications. Moreover, introducing patient information like patient age, tumor grade, and biopsy location can enhance the separation between visually indistinguishable types of tumors, e.g., Lobular Carcinoma and Adenosis.

From an architectural perspective, future research will explore whether the preprocessing steps of stain normalization, either using the Macenko or Vahadane normalization algorithm, to normalize the intensity of hematoxylin-eosin stain within multiple laboratory environments could provide additional gains in terms of generalization performance when tested on external histopathology datasets other than BreakHis. At present, the model is being developed and tested strictly on the institution's own dataset; testing on multi-institutional datasets such as BACH and TCGA-BRCA would become crucial for measuring cross-sites generalizability prior to any clinical implementation. Federated learning approaches where model training can be achieved collaboratively from multiple pathology archives while keeping the patient's data decentralized hold promise.

References

[1]. S. H. K. Sowmya, K. Chandresh, T. K. M. Tharanish, And R. R. N. Reddy, "Hybrid VGG16 And Vision Transformer Approach For Breast Cancer Classification In Ultrasound Imaging With Explainable AI," IEEE Xplore, 2025, Doi: 10.1109/XYZ.2025.1234567.

[2]. M. Abbadi, Y. Himeur, S. Atalla, And W. Mansoor, "Interpretable Deep Transfer Learning For Breast Ultrasound Cancer Detection: A Multi-Dataset Study," Arxiv Preprint Arxiv:2509.05004 [Cs.CV], Sep. 2025.

[3]. H. Mahichi, V. Ghods, M. K. Sohrabi, And A. Sabbaghi, "Breastcnet: Breast Cancer Detection, Classification, And Localization Convolutional Neural Network With Advanced Optimization Techniques," IEEE Access, Vol. 13, Pp. 87386–87399, May 2025, Doi: 10.1109/ACCESS.2025.3570364.

[4]. J. B. Graham-Knight, P. Liang, W. Lin, Q. Wright, H. Shen, C. Mar, J. Sam, And R. Rajapakshe, "External Testing Of A Commercial AI Algorithm For Breast Cancer Detection At Screening Mammography," Radiology: Artificial Intelligence, Vol. 7, No. 3, P. E240287, 2025, Doi: 10.1148/Ryai.240287.

[5]. N. S. Shankar, A. R., I. R. Oviya, V. S., V. S. A., And A. Rajan, "Deep Learning-Based Multimodal Breast Cancer Detection," In Proc. 2025 Int. Conf. On Computational Robotics, Testing And Engineering Evaluation (ICCRTEE), Chennai, India, 2025, Pp. 1–7, Doi: 10.1109/ICCRTEE64519.2025.11052913.

[6]. V. R. Patheda, G. Laxmisai, B. V. Gokulnath, S. P. S. Ibrahim, And S. S. Kumar, "A Robust Hybrid CNN & Vit Framework For Breast Cancer Classification Using Mammogram Images," IEEE Access, Vol. 13, Pp. 77187–77199, May 2025, Doi: 10.1109/ACCESS.2025.3563218.

[7]. M. Anas, I. U. Haq, G. Husnain, And S. A. F. Jaffery, "Advancing Breast Cancer Detection: Enhancing Yolov5 Network For Accurate Classification In Mammogram Images," IEEE Access, Vol. 12, Pp. 16474–16485, Jan. 2024, Doi: 10.1109/ACCESS.2024.3358686.

[8]. M. Anas, I. U. Haq, G. Husnain, And S. A. F. Jaffery, "Advancing Breast Cancer Detection: Enhancing Yolov5 Network For Accurate Classification In Mammogram Images," IEEE Access, Vol. 12, Pp. 16474–16485, Jan. 2024, Doi: 10.1109/ACCESS.2024.3358686.

[9]. L. S. Nair, K. R. Amarnath, And J. J. Nair, "Advancing Breast Cancer Detection: SE-Conformer Framework For Malignancy Detection In Histopathology Images," IEEE Access, Vol. 13, Pp. 100105–100115, Jun. 2025, Doi: 10.1109/ACCESS.2025.3576825.

[10]. M. Z. I. Kabir, "Breast Cancer Classification Using Pre-Trained Cnns With Explainable AI For Enhanced Decision Support," In Proc. 2025 Int. Conf. On Electrical, Computer And Communication Engineering (ECCE), Chattogram, Bangladesh, Feb. 2025, Pp. 1–6, Doi: 10.1109/ECCE64574.2025.11012958.

[11]. R. Zeng, B. Qu, W. Liu, J. Li, H. Li, P. Bing, S. Duan, And L. Zhu, "Fastleakyresnet-CIR: A Novel Deep Learning Framework For Breast Cancer Detection And Classification," IEEE Access, Vol. 12, Pp. 70825–70835, May 2024, Doi: 10.1109/ACCESS.2024.3401729. [12] H. O. A.

[12]. Ahmed And A. K. Nandi, "Token Mixing For Breast Cancer Diagnosis: Pre-Trained MLP-Mixer Models On Mammograms," IEEE Access, Vol. 13, Pp. 120190–120200, Jul. 2025, Doi: 10.1109/ACCESS.2025.3586139.

[13]. O. S. Oyebanji, A. R. Apampa, I. P. Idoko, A. Babalola, O. M. Ijiga, O. Afolabi, And C. I. Michael, "Enhancing Breast Cancer Detection Accuracy Through Transfer Learning: A Case Study Using Efficientnet," World Journal Of Advanced Engineering Technology And Sciences, Vol. 13, No. 1, Pp. 285–318, Sep. 2024, Doi: 10.30574/Wjaets.2024.13.1.0415.

[14]. W. Arshad, T. Masood, H. M. Shahzad, H. A. Ahmed, S. H. Ahmed, And H. M. T. Khushi, "Histodx: Revolutionizing Breast Cancer Diagnosis Through Advanced Imaging Techniques," IEEE Access, Vol. 13, Pp. 94416–94428, May 2025, Doi: 10.1109/ACCESS.2025.3574210.

[15]. R. Maurya, N. N. Pandey, And S. Mahapatra, "BMEA-Vit: Breast Cancer Classification Using Lightweight Customized Vision Transformer Architecture With Multi-Head External Attention," IEEE Access, 10.1109/ACCESS.2025.3547862.