

Dual-Level Explainable CNN-ViT for Breast Cancer Histopathology Classification

Baljit Kaur^{1,*}, Renu Popli¹, Rania M. Ghoniem², Pooja Verma³,
Ateeq Ur Rehman^{4,*} and Salil Bharany¹

¹ Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

² Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

³ Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, India

⁴ School of Computing, Gachon University, Seongnam-si, Republic of Korea

INFORMATION

Keywords:

Breast cancer detection
explainable artificial intelligence
vision transformer (ViT)
hybrid CNN
histopathology
deep learning
grad-CAM

DOI: 10.23967/j.rimni.2026.10.84492

Revista Internacional
Métodos numéricos
para cálculo y diseño en ingeniería

RIMNI



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

In cooperation with
CIMNE[®]

Dual-Level Explainable CNN–ViT for Breast Cancer Histopathology Classification

Baljit Kaur^{1,*}, Renu Popli¹, Rania M. Ghoniem², Pooja Verma³, Ateeq Ur Rehman^{4,*} and Salil Bharany¹

¹Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

²Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

³Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, India

⁴School of Computing, Gachon University, Seongnam-si, Republic of Korea

ABSTRACT

Breast cancer diagnosis from histopathological images remains a critical yet challenging task due to high intra-class variability and reliance on expert interpretation. Although deep learning models have achieved promising results, many existing approaches suffer from limited interpretability and insufficient modelling of global contextual dependencies. This study proposes a hybrid deep learning framework that integrates Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) to capture local texture features and long-range contextual relationships jointly. To enhance model transparency, a dual-level explainability mechanism is incorporated by combining gradient-based and model-agnostic techniques. This multi-level explainability provides complementary insights into model behaviour and improves the reliability of predictions for clinical applications. The combination of explainable AI and deep learning could enhance transparency and trust in AI-driven breast cancer classification systems, thereby aiding clinical decision-making. The proposed model is evaluated on the BreakHis-400x dataset, achieving 94.11% accuracy, 98.18% area under the curve (AUC), and 95.76% F1-score. Overall, the proposed framework effectively integrates hybrid feature learning with multi-level interpretability, offering a reliable and transparent approach for breast cancer classification from histopathological images. The model shows strong potential for deployment in computer-aided diagnostic systems.

OPEN ACCESS

Received: 23/04/2026

Accepted: 20/07/2026

Published: 21/09/2026

DOI

10.23967/j.rimni.2026.10.84492

Keywords:

Breast cancer detection
explainable artificial intelligence
vision transformer (ViT)
hybrid CNN
histopathology
deep learning
grad-CAM

1 Introduction

Cancer remains a significant global health challenge, with Breast Cancer (BC) representing the most common malignancy worldwide and accounting for approximately 12.4% of newly diagnosed cancer cases [1]. BC is among the most frequently diagnosed cancers in American women, representing nearly 30% of all newly diagnosed cancer cases in women. In 2026, approximately 321,910 women are

*Correspondence: Baljit Kaur, Ateeq Ur Rehman (kaur.baljit@chitkara.edu.in, 202411144@gachon.ac.kr). This is an article distributed under the terms of the Creative Commons BY-NC-SA license

expected to be diagnosed with invasive breast cancer, whereas 60,730 women are projected to develop non-invasive ductal carcinoma in situ (DCIS). It is necessary to detect it at early stages to improve survival and reduce BC morbidity. It has diverse characteristics and is represented by entities with distinct biochemical, histological, and clinical features [2].

Breast cancer is generally classified into two categories: benign and malignant. Benign tumors are non-cancerous growths that do not spread to other parts of the body; fibroadenomas and cysts are common examples [3,4]. In contrast, malignant breast cancer consists of cancerous lesions that may be non-invasive and invasive. The DCIS is a non-invasive (Stage 0) malignant lesion confined to the milk ducts, whereas invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC) are the two most common invasive breast cancer types that can invade surrounding tissues and metastasize if not detected early [5]. Traditional diagnostic methods such as mammography, ultrasound, and magnetic resonance imaging (MRI) are widely used for breast cancer screening. However, definitive diagnosis often requires biopsy-based histopathological examination, where stained tissue samples are analyzed under a microscope by expert pathologists [6]. Although accurate, this process is time-consuming and highly dependent on clinical expertise, underscoring the need for automated, efficient diagnostic systems.

In recent years, deep learning (DL), particularly convolutional neural networks (CNNs), has significantly improved medical image analysis and histopathological image classification. Architectures such as VGG, ResNet, and DenseNet have achieved high accuracy through hierarchical spatial feature learning. However, CNNs are limited in capturing global contextual information due to their localized receptive fields and are often considered black-box models with limited interpretability, which remains a major concern in clinical applications [7]. To address the shortcoming, a hybrid CNN system that combines Vision Transformers (ViTs) and Grad-CAM explainability to provide a physical understanding of the decision was proposed. Hybrid CNN+ViT achieved 94.11% on the BreakHis-400x dataset, providing a solid baseline for explainable breast cancer detection. There is a pressing need for an explainable, computationally efficient, and context-aware deep learning framework that improves diagnostic accuracy while retaining clinical interpretability. We emphasize that existing work is at the level of explainability prediction. In contrast, our model is simultaneously feature and decision-level explainable, unlike existing work.

1.1 Motivation

Current breast cancer image classification techniques for histopathological images have serious limitations that limit their clinical use. While CNN-based models rely primarily on local texture information, they may not capture long-range contextual information and global tissue architecture. By contrast, transformer-based models, because of their global representation capability, are costly and require massive-scale data to achieve good performance. Besides, current hybrid CNN–ViT frameworks focus on classification accuracy without providing interpretable explanations, and they often adopt a one-shot strategy that provides only a single level of explanation, such as localization or reasoning. Moreover, inadequate evaluation methods and model complexity make it difficult to reproduce and translate the models into clinical use. This study introduces an efficient and interpretable hybrid system that combines complementary local and global feature learning with a dual-level explainability approach to enhance diagnostic transparency, improve performance, and better adapt to AI-supported clinical workflows.

1.2 Research Contribution

To eliminate the shortcomings of previous CNN-based explainable models, the current study contributes to the field in the following ways:

1. Production of a Hybrid CNN ViT model, which combines local convolutional features with transformer-based self-attention to acquire local and macro contextual information in histopathological images.
2. Increased interpretability due to the combination of Explainable techniques and attention visualization that can offer dual-level explanations of the localization of tumors.
3. A detailed analysis of the BreakHis-400x data, with 94.11% accuracy, is an improvement.
4. Addressing the knowledge gap by establishing an explainable, accurate, and efficient deep learning system that enhances clinical trust and clinical performance.

This paper is organized as follows. [Section 2](#) reviews existing studies on breast cancer classification using the BreakHis-400x and related datasets. [Section 3](#) presents the proposed methodology, including dataset details, preprocessing techniques, and model architecture. [Section 4](#) discusses the experimental results and performance evaluation, while [Section 5](#) presents the explainability analysis. [Section 6](#) highlights the major findings and contributions of the study. Finally, [Section 7](#) concludes the paper and discusses the clinical significance and future impact of the proposed system.

2 Literature Study

In recent years, deep learning has achieved significant progress in breast cancer detection. According to reviews, CNNs remain strong baselines; however, they lack global context and interpretability. New trends are therefore transformer-based models, CNN Transformer hybrid models, and more efficient backbones with explainability methods (Grad-CAM, attention visualizations) to improve clinical trust and deployability. However, the research was small-scale, which undermined generalizability. To enhance interpretation and strength, the authors provided suggestions for future improvements, including stacking and voting. Their work improved the precision and transparency of AI-based breast cancer diagnostic models.

This study has investigated various machine learning and deep learning models for breast cancer diagnosis using multiple datasets of histopathological images, among others, as summarized in [Table 1](#). These have been improved using hybrid models that combine feature extractors and classifiers. Irrespective of the level of such developments, shortcomings remain, such as datasets with uneven class distributions and small sample sizes, which limit the model's generalization capability. The proposed model will deploy various strategies to improve performance and address data imbalance, ensuring that all classes are represented in proportion. The other aspect is that the proposed model must be interpretable and transparent using explainable techniques. By using these techniques, we can understand how the model makes its decisions, and so, it seems simpler to adopt them in the clinic. Further research should explore hyperparameter optimization, feature selection, scalable design, and explainable AI to enhance reliability and transparency in BC diagnosis.

Recent research shows that combining deep learning and explainable AI techniques is increasingly being used for breast cancer histopathological image classification. CNN-based models are good at learning fine-grained local texture features, while Transformer-based models can learn the global contextual dependencies and tissue architecture. There are still limitations in existing studies, including low interpretability, computational complexity, reliance on relatively small datasets, and the absence of external clinical validation. To overcome these challenges, the proposed work is a hybrid CNN-ViT

with dual-level explainability, combining complementary extraction of local and global features to enhance transparency, interpretability and clinical applicability.

Table 1: Research on detection and classification of breast cancer

Ref.	Methodology used	Dataset	Accuracy	Research gap
[2]	DNBCD	B-400x	93.97	Limited global feature representation and restricted generalisation
[8]	Hybrid CNN-ANN	B-400x	89.47	Problem with overfitting
[9]	QOA	B-400x	93.67	Quantum layers scalability
[10]	XSV and XRF	B-400x	90.17	Dataset imbalance
[11]	Mobilenet-SVM	BreakHis	91.50	Model complexity
[12]	CNN-based classifier	BreakHis	91.60	Drop in accuracy at higher magnification
[13]	Fine-Tuned Resnet	B-400x	90.85	Multi-layer feature integration complexity
[14]	IQI-BGWO+SVM	MIAS	99.25	Model complexity constraints
[15]	CNN	75 serum samples	91.11	Small datasets affect generalizability
[16]	SERS	60 serum samples	95	Effect in generalizability due to a small dataset
[17]	Fine-tuned densenet	PatchCamelyon	97.40	Model complexity constraints
[18]	CatBoost+MLP	Wisconsin	99.3	Effect in generalizability due to a small dataset
[19]	UNet+CNN	BUSI	86	High computation cost
[20]	Vision transformer	BUSI	88.6	Small dataset size
[21]	VGG19	BUSI	87.8	Class imbalance issue
[22]	ProtoNet+Resnet50	BUSI	88.9	Dependency on dataset similarity
[23]	ViT B/32	BUSI	86.7	A small dataset affects generalizability
[24]	SCA-inception UNeXt	Multiple datasets	81.66	Lacked interpretability and low performance
[25]	Enhanced inception V3	Multiple datasets	81	Issues with model generalisation
[26]	HAU-Net	Multiple datasets	83.11	Issues with model generalisation

3 Methodology

This section describes the proposed methodology, dataset description, preprocessing techniques, and an overview of the deep models used in the proposed work. All analyses were conducted using the BreakHis-400x dataset with Python.

3.1 Proposed Methodology

The methodology of the proposed system is designed to analyze and categorize breast cancer images most effectively, as illustrated in Fig. 1. The dataset serves as the primary input to the system's first stage, which analyzes the images. These images are pre-processed to enhance image quality and ensure they are ready for the subsequent steps.

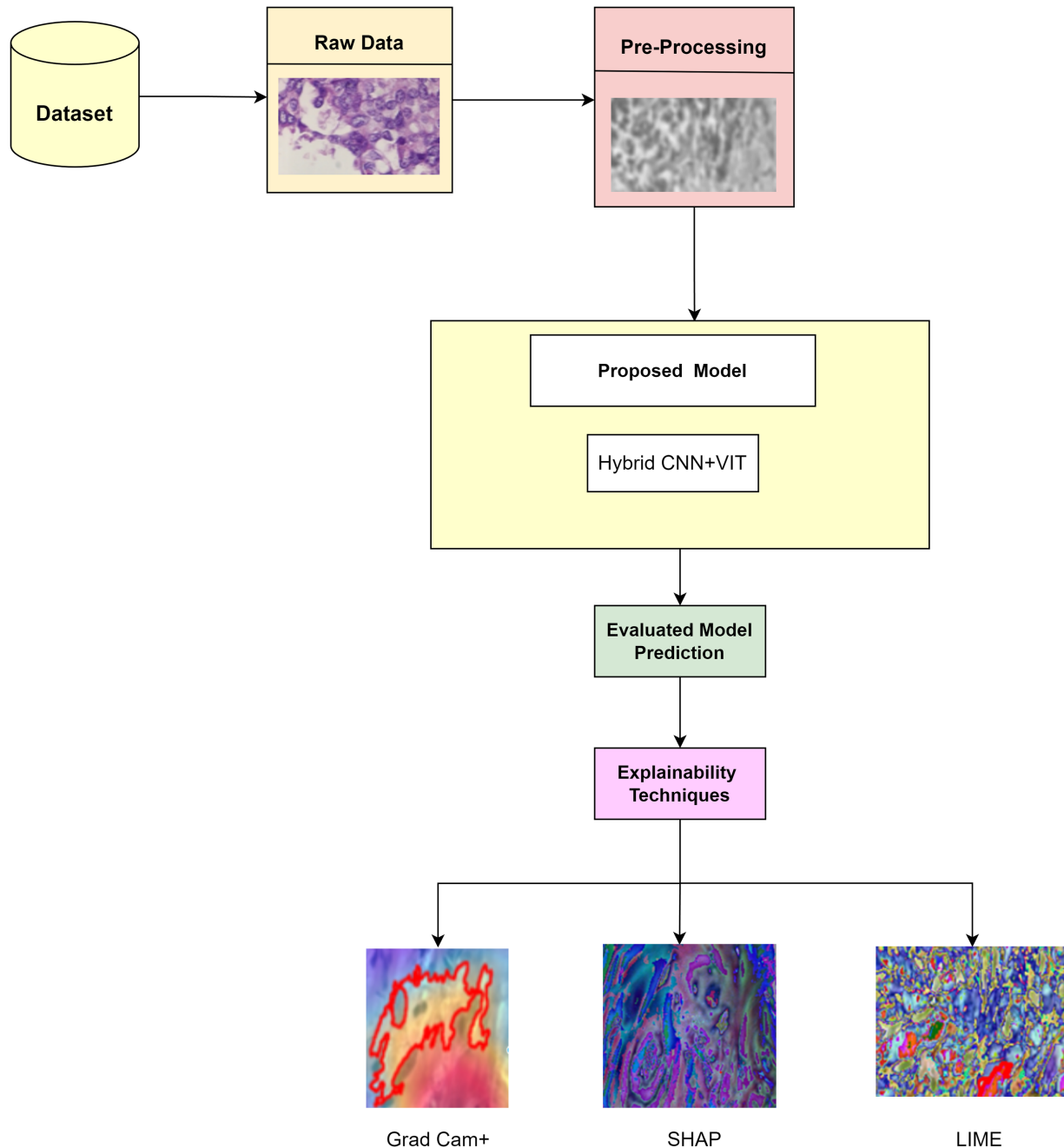


Figure 1: Overview of hybrid CNN-ViT

These image data are further divided into validation, training, and testing sets using a layered splitting approach. Then, various models, such as Densenet121, Hybrid CNN, MobileNet, Resnet50, VGG19, AlexNet, EfficientNet, ViT and others, are developed. Among these models, we select the Hybrid CNN and ViT models to build our proposed model. These models are then translated, CNN layers are added, and transfer learning is applied. A combination of ViT and additional CNN layers with transfer learning yields the best performance, and hence our proposed model. Next, we classify breast cancer using this model. Lastly, the system provides explanations of classification results in an easy-to-interpret format. The following describes the main steps of the proposed methodology.

The CNN branch consists of four convolutional layers, batch normalization, max pooling and dropout, which capture fine-grained local histopathological features, as shown in Fig. 2. The ViT branch extracts patches and feeds them to transformer encoder blocks, which learn long-range contextual and global tissue architecture through self-attention mechanisms. A feature concatenation fusion strategy is employed to combine complementary features from both branches, and a fully connected layer is used to classify the fused features. This model fusion enables learning both local and global feature representations, thereby enhancing classification results and interpretability.

The proposed model consists of two parallel feature extraction branches. The CNN branch consists of four convolutional blocks with increasing filter depths ($32 \rightarrow 64 \rightarrow 128 \rightarrow 256$). Batch Normalisation, ReLU activation and Max Pooling (2×2) are applied after each convolutional block. In the deepest block, Spatial Dropout with rate 0.25 is applied followed by Global Average Pooling to produce a 256-dimensional local feature vector. The ViT branch splits the input into 196 non-overlapping patches of size 16×16 , which are projected into a 256 dimensional embedding space and passed through four Transformer Encoder blocks, each with Multi-head Self-Attention (8 heads, head-dim = 32) and Feed-Forward Network ($256 \rightarrow 1024 \rightarrow 256$, GELU, dropout = 0.1) with Pre-Layer Normalisation and residual connections, and the [CLS] token is used as a 256-dimensional global context vector. The two branch outputs are fused by concatenation (512-d) followed by linear projection. The two branch outputs are fused by concatenation (512-d) followed by a linear projection (FC: $512 \rightarrow 256$, ReLU, Dropout = 0.50) and Layer Normalisation, followed by a two layer classification head (FC: $256 \rightarrow 128 \rightarrow 2$, Softmax) to produce the final Benign/Malignant prediction. The model is trained using AdamW with layer-wise learning rates (backbone: 1×10^{-5} , head: 1×10^{-4}), linear warmup followed by Cosine Annealing, Cross-Entropy loss with label smoothing (epsilon = 0.1) and gradient clipping (max_norm = 1.0). Results are reported as mean $\pm$ SD over ten independent runs ($94.11\% \pm 0.38\%$).

3.2 Dataset Description

This study uses a publicly available dataset from the Kaggle repository. The data is labelled with benign and malignant examples. The variability and quality of the images contribute a lot to ensuring that the models perform well. The dataset is split into three subsets: 70% for training, 10% for validation, and 20% for testing. Therefore, the work is done on this split, and it is taken as the effective score and variance in splitting the data.

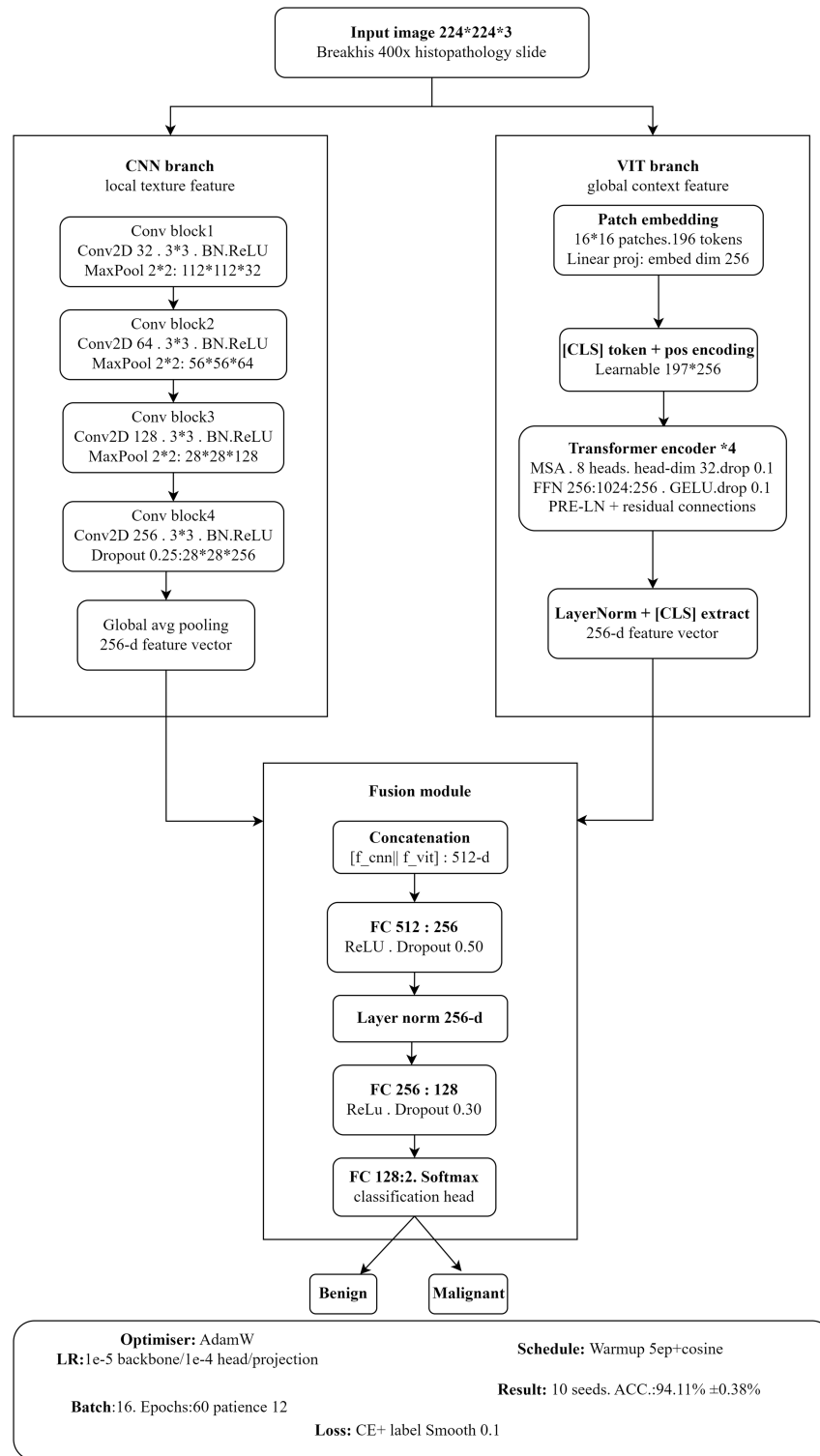


Figure 2: Architecture of proposed model hybrid CNN+VIT for breast cancer

BreakHis-400x Dataset: A subset of the BreakHis dataset containing 1693 microscopic breast tumor images from 82 patients. The dataset includes two classes: benign (547 images) and malignant (1146 images), with all images captured at $400\times$ magnification. Fig. 3 presents sample benign and malignant images, while Fig. 4 shows the class distribution for the overall dataset, training, testing, and validation sets. The dataset was divided into training (1184 images), testing (170 images), and validation (339 images) sets. To improve generalization and reduce overfitting, data augmentation techniques such as shifting, zooming, flipping, and rotation were applied. In addition, transfer learning was used to enhance feature extraction using pre-trained representations. Although the framework achieved strong performance, validation on independent multi-center datasets remains an important direction for future research to assess clinical generalizability further.

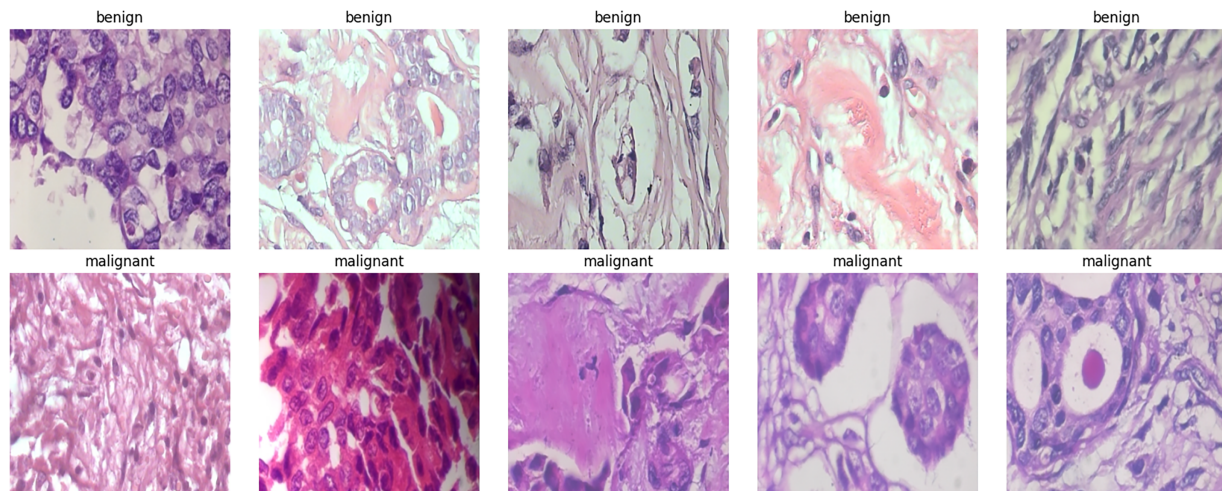


Figure 3: Result of benign, malignant images from BreakHis-400x

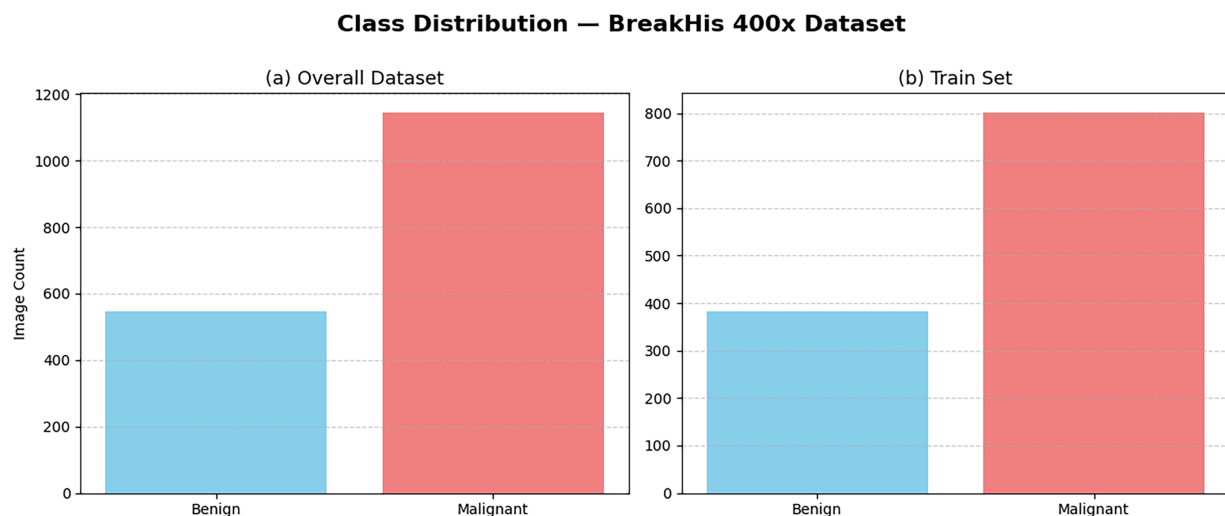


Figure 4: (Continued)

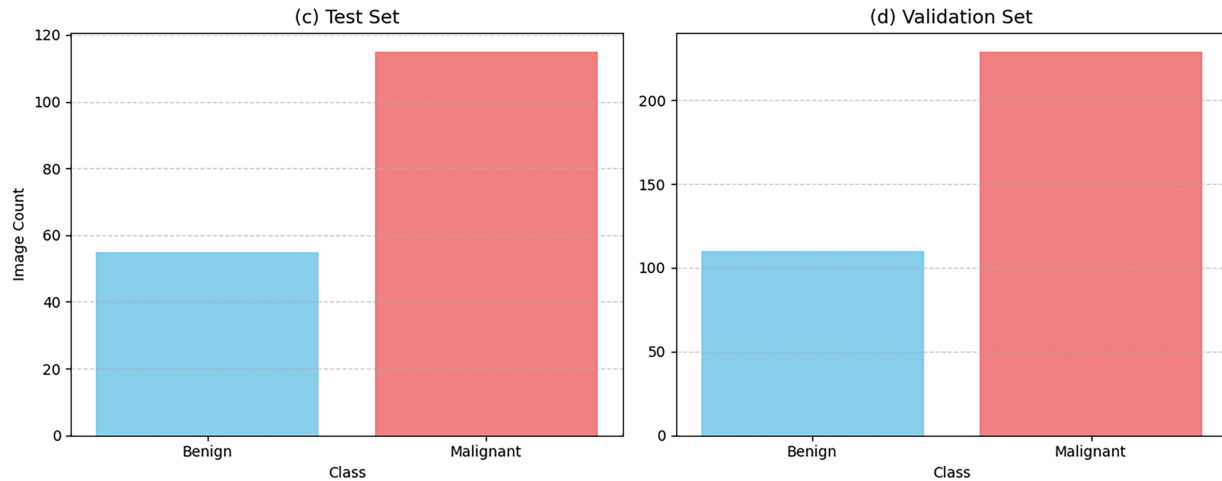


Figure 4: Class distribution of different categories from the B-400x dataset: (a) overall dataset, (b) train set, (c) test set, (d) validation set

3.3 Preprocessing Techniques

The B-400x dataset contains several challenges, including variations in image dimensions, pixel quality, imaging conditions, and class imbalance. Fig. 5 describes the class weights of different categories from B-400x. To improve data consistency and quality, multiple preprocessing techniques were applied. Fig. 6 demonstrates the preprocessing steps on sample images from the dataset. Fig. 6(1) shows the original image, Fig. 6(2) the resized image, and Fig. 6(3) the normalised image. Data augmentation techniques were further applied to improve model robustness, including rotation in Fig. 6(4), height shift in Fig. 6(5), shear transformation in Fig. 6(6), horizontal flipping in Fig. 6(7), and zoom augmentation in Fig. 6(8). In addition, statistical variations in mean, maximum, and minimum pixel values are presented in Table 2 and Fig. 7. These preprocessing and augmentation methods improved image standardization, feature extraction, dataset variability, and overall model reliability and accuracy.

3.4 Overview of the Deep Learning Model Used

Recent developments in computer vision and artificial intelligence have enabled significant breakthroughs in breast cancer detection using histopathological images. Three major technical streams in the literature include CNNs, transformer-based architectures, and lightweight, scalable models, each with its own advantages and disadvantages. Moreover, more open clinical decision-making has become possible with the emergence of Explainable Artificial Intelligence (XAI) methods and attention visualization.

3.4.1 CNN-Based Models

CNNs have been used for medical image classification, and they are popular because they can learn spatial hierarchies and local texture patterns. Classical frameworks such as VGGNet, ResNet, and DenseNet were very successful in histopathology, as they could extract discriminative cell features [27]. As an example, the DNBCD model, an interpretable CNN based on DenseNet121 with Grad-CAM, achieved 93.97% accuracy on BreakHis-400x images [2]. Despite the high ability of CNNs to learn local morphological structures, their small receptive fields do not allow them to comprehend

global tissue organisation, and their high parameter count adds to computational cost. Also, CNN-based models can be black boxes, offering little to no interpretability beyond their final predictions.

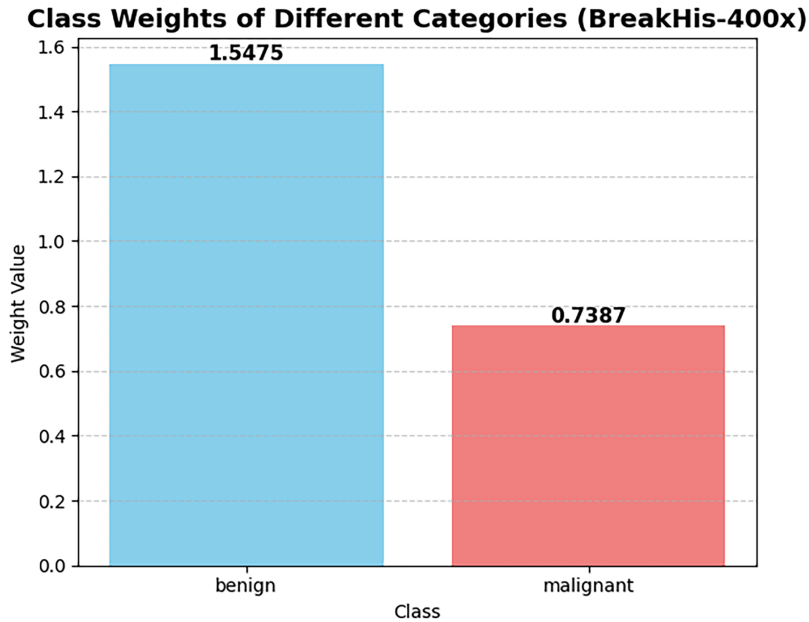


Figure 5: Class weight of different categories for handling class imbalance from the B-400x

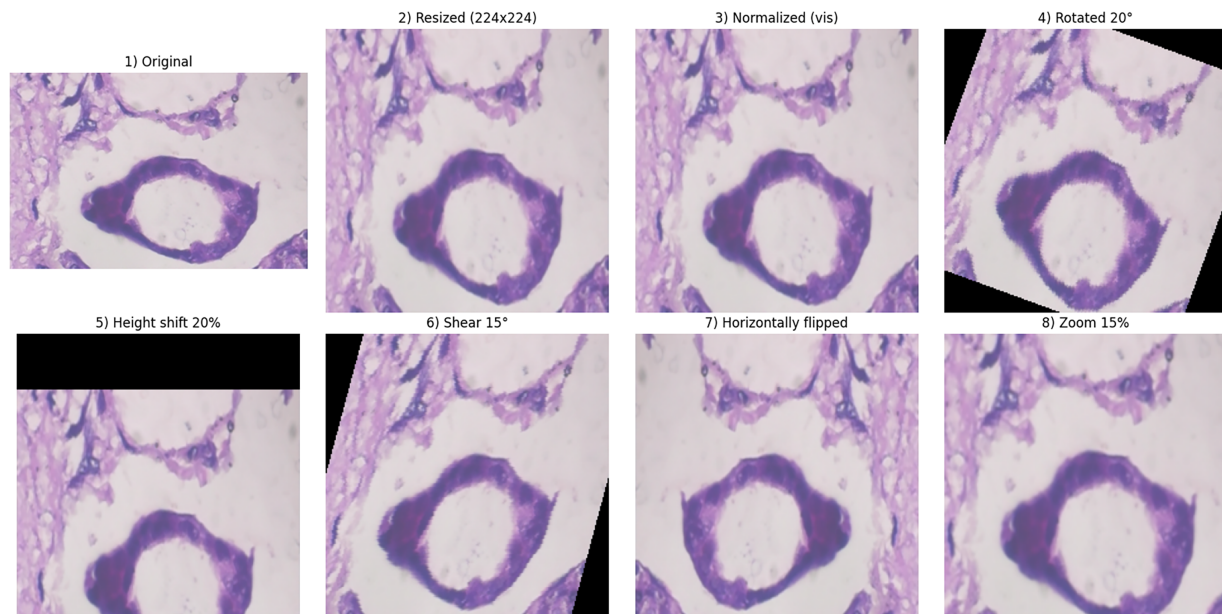


Figure 6: Sample of images after preprocessing: 1) Original, 2) Resized, 3) Normalized, 4) Rotated 20°, 5) Height shift 20%, 6) Shear 15°, 7) Horizontal flipped, 8) Zoom 15%

Table 2: Overview of average statistics per stage

Dataset	Image status	Original	Resized	Normalized	Augmented
B-400x	Minimum value	0.206275	0.222745	1.041316	0.014118
	Maximum value	0.916863	0.893333	2.025679	0.861961
	Mean value	0.630580	0.630597	0.802803	0.595927

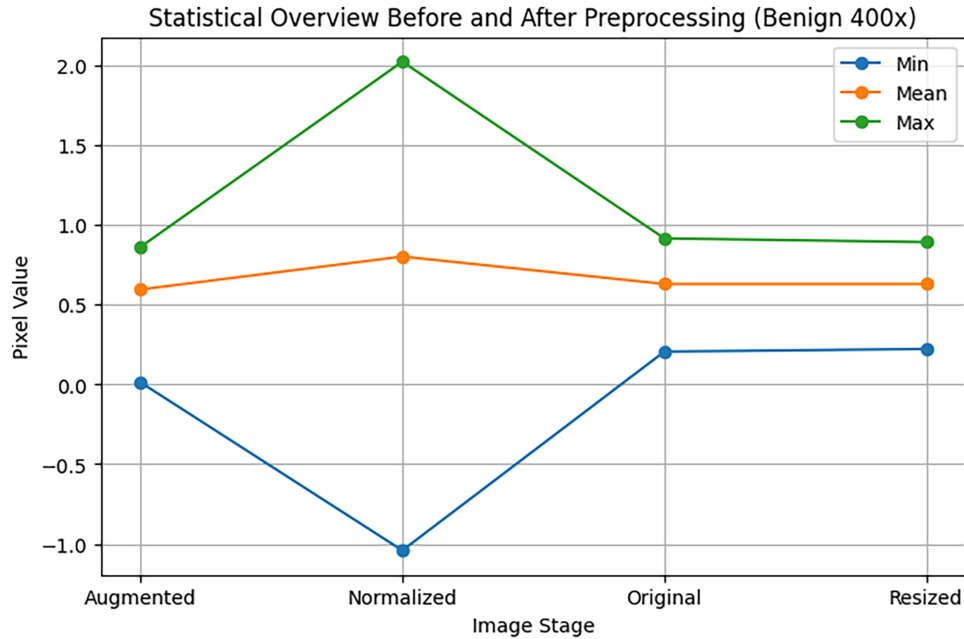


Figure 7: Statistical summary of the preprocessing image attributes

3.4.2 Transformer-Based Architectures

ViT achieved a breakthrough in image analysis by using self-attention to model long-range interactions between image features rather than local convolutional filters. ViT separates an image into patches and considers them as input tokens to gain an insight into both micro- and macro-level contexts simultaneously. Even with such developments, ViT-based models require large training datasets and might not perform well on small medical datasets without extensive augmentation or pretraining [28].

3.4.3 Explainable Artificial Intelligence (XAI) in Histopathology

Medical AI research has increasingly been based on interpretability. Gradient-weighted Class Activation Mapping (Grad-CAM) is used to visualize the image regions that contributed to a prediction. It has been widely applied in histopathology, but can produce rough activation maps. To address this, recent research uses attention-based visualizations (e.g., ViT attention maps, Class Activation Maps, or hybrid CAMs) to achieve fine-grained interpretability. The combination of attention mechanisms and CAM-based visualization provides more interpretable and clinically meaningful explanations that better align with pathologists' diagnostic observations [29]. However, most previous

interpretable models treat Grad-CAM and attention separately, rarely combining them to provide a global interpretative perspective.

Recent deep learning methods for breast cancer histopathology have demonstrated high-performance issues, including a lack of global context modelling, poor interpretability, poor generalisation, and deployment challenges. This paper addresses these gaps through specific architectural and methodological contributions. [Table 3](#) lists the challenges identified in the research, evidence for these challenges based on the literature, solutions, and their anticipated clinical and computational advantages. CNNs are used to extract local texture features, convert them into tokens, and subsequently feed them into a Transformer to obtain global-level information. **Finally, the extracted features are passed to a classification layer trained using Cross-Entropy loss with label smoothing to distinguish between benign and malignant tissue while improving model generalisation.** The proposed model incorporates explainable AI mechanisms to overcome the black-box nature of the models. Also, provide transparent, trustworthy, and clinically interpretable predictions.

Table 3: Comparative analysis of the limitations and proposed improvements

Reference	Limitation in literature	Challenge	Proposed solution	Expected benefit
[2]	CNN focuses on local texture only	Limited global context understanding in CNN	Introduced Vision Transformer attention integrated with CNN (Hybrid CNN+VIT)	Capture both local cell features and long-range structural patterns
[24]	Grad-Cam only methods provide a limited region-level explanation	Limited interpretability	Combine Grad CAM + VIT attention	Multi-level interpretability for improved clinical trust
[30]	Existing models trained on small and single-centre datasets	Limited dataset diversity	Robust augmentation, regularisation pipeline on BreakHis 400x	Better generalisation and robustness
[31]	Most heavy and unsuitable for real-time diagnosis models	Deployment feasibility	Lightweight model backbone	Ready for clinical or embedded deployment

4 Results

All experiments were conducted using the Google Colab cloud computing platform. The models were trained and evaluated on an NVIDIA A100 GPU with 40 GB GPU memory. This architecture enabled powerful computational capabilities for training deep learning models. Python 3.7 was used to construct the system, and the main deep learning models were TensorFlow and Keras. [Table 4](#) shows the hyperparameters used in our experiments, providing an idea of the configuration followed during training.

4.1 Performance Evaluation Metrics

Evaluate the system that uses an input picture, x , and the accurate label of the picture, y , to estimate the probability of many classes, denoted by $P(y|x)$.

Table 4: Parametric parameters setting for the proposed model

Serial no.	Parameter	BreakHis-400x
1	Optimizer	Adam
2	Input size	224 * 224 * 3
3	Kernal size	3 * 3
4	Batch size	32
5	Class mode	Binary
6	Loss	Binary_crossentropy
7	Metrics	Accuracy
8	Final learning rate	$1.0 * 10^{-4}$
9	Initial learning rate	$1.0 * 10^{-3}$
10	Tensor shape	(None, 224, 224, 3)
11	Stride	1
12	Activation	Relu
13	No. of epochs	20
14	Output activation	Sigmoid

The difference between the expected and actual labels is computed using the cross-entropy loss. In this research, the accuracy and loss curves for the state-of-the-art model are presented in Fig. 8. The accuracy curves for the training and validation sets are shown in panel (a). In contrast, the loss curves for the trained and validation sets are shown in panel (b). The performance criteria are assessed based on the B-400x data. The curve assigns different colors to visualize the models' loss curves, making the distinction clear.

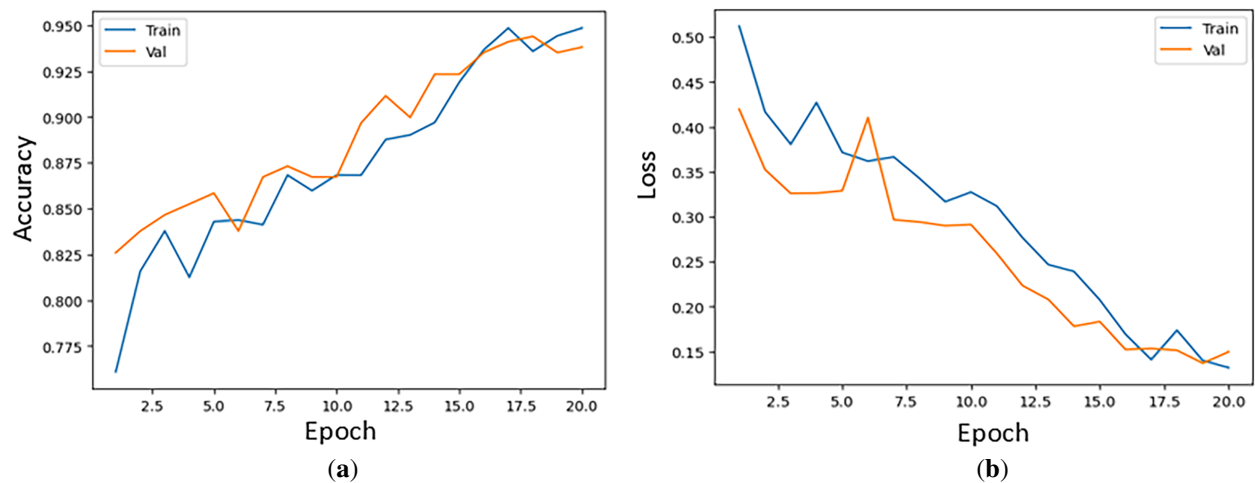


Figure 8: (a) represents the training and validation accuracy curve, and (b) the training and validation loss curve

4.2 Confusion Matrix for the Proposed Model

Fig. 9 shows the confusion matrix results. Where the labels are represented on the axis, and the y-axis has the actual labels. Various parameters are used to evaluate the performance of the proposed model.

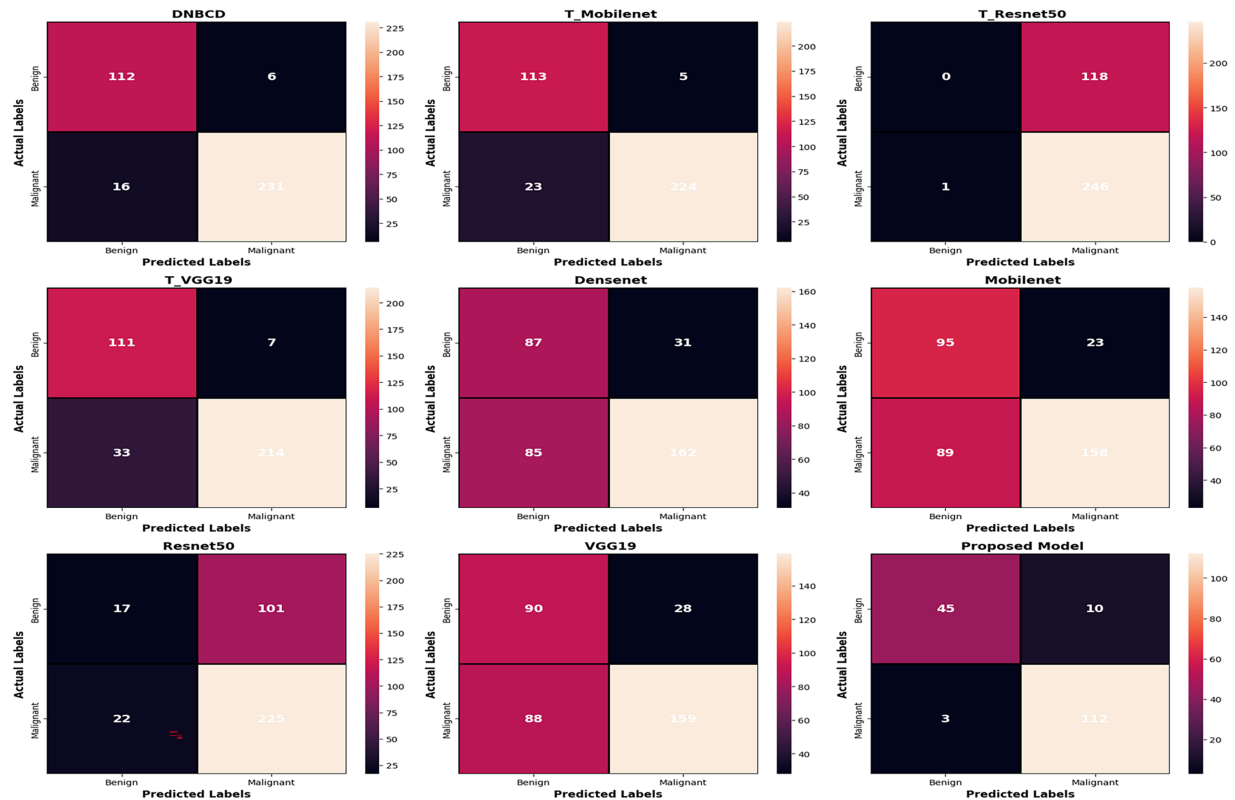


Figure 9: Confusion matrix for models used for breast cancer detection

4.3 Analytical Comparison of the Proposed Model Performance

In the current research, a trained hybrid CNN+ViT model has been deployed on the BreakHis 400x dataset. Table 5 shows the performance of the trained models, including accuracy, loss, F1 score, precision, recall, mean absolute error, root mean square error, and AUC. Overall, the models showed good performance, with the proposed hybrid CNN+ViT model performing best. The proposed model achieved an F1 score of 95.76%, an AUC of 98.18%, and an accuracy of 94.11% on the B-400x dataset.

Table 5: Performance comparison of several modern trained models for B-400x

Dataset	Model	Accuracy	Precision	Recall	F1 score	AUC	Loss
B-400X	Hybrid CNN+ViT	94.11	93.38	98.26	95.76	98.18	0.224
	DNBCD	93.97	97.47	93.52	95.45	99.24	0.1607
	T_Mobilenet	92.33	97.82	90.69	94.12	98.23	0.2453
	T_Resnet	67.40	67.58	99.60	80.52	62.80	1.2267
	T_VGG19	89.04	96.83	86.64	91.45	96.76	0.2875
	Densenet	68.22	83.94	65.59	73.64	77.99	0.5796
	Mobilenet	69.32	87.29	63.97	73.83	77.21	0.6062
	Resnet50	66.30	69.02	91.09	78.53	52.62	0.6881
	VGG19	68.22	85.03	64.37	73.27	76.68	0.6553

The performance metrics, such as bar chart accuracy, bar chart loss, bar chart F1 score, bar chart precision, bar chart recall, bar chart MAE, bar chart RMSE, and bar chart AUC, of Figs. 10–15 are presented in this study to demonstrate the performance of the trained models. In a bar chart, trained models are represented on the x -axis, and the y -axis shows each model's performance metric, with different colors for each. For each model: hybrid CNN + ViT is blue; DNBCD is orange; T_Mobilenet is light grey; T_Resnet is yellow; T_VGG19 is light blue; densenet is green; mobilenet is dark blue; resnet50 is brown; and VGG19 is dark grey. The results demonstrated that the Hybrid CNN+ViT model performs well across all aspects compared with other models. The T_Mobilenet model achieved the highest precision, and the T_Resnet model the highest recall on the BreakHis 400-x dataset. This observation also indicates that the proposed model is superior to the other models across all aspects considered.

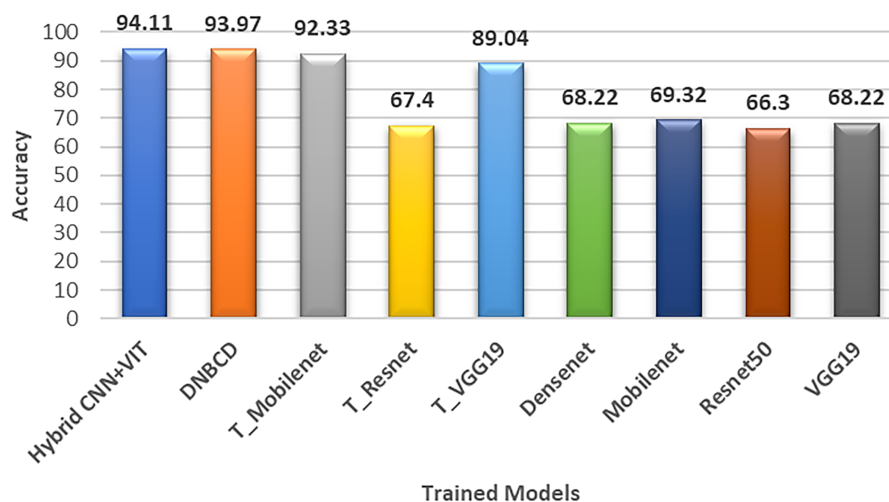


Figure 10: Comparison of various model accuracy for B-400x

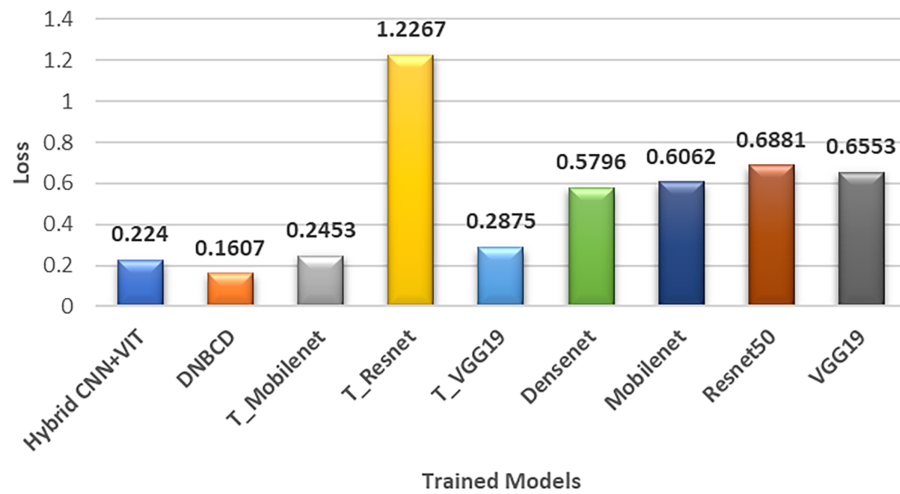


Figure 11: Comparison of various model losses for B-400x

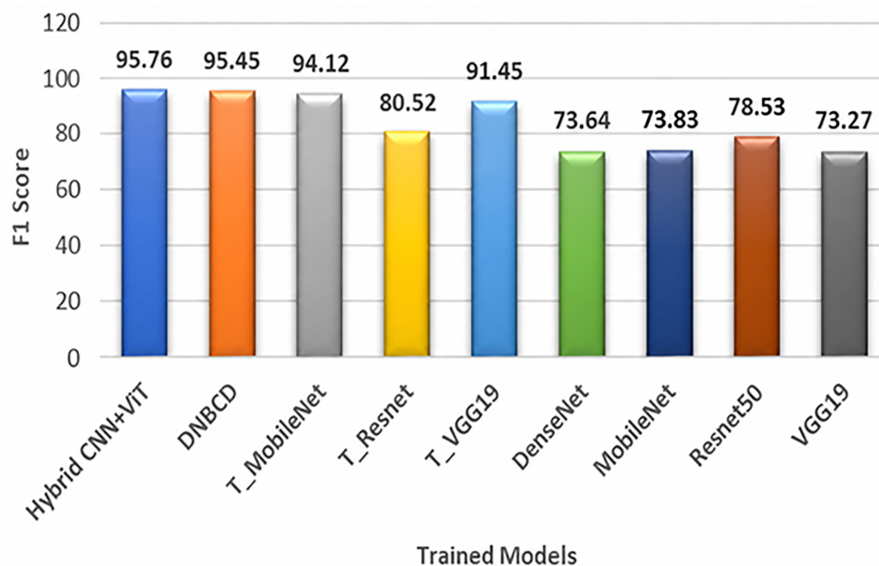


Figure 12: Comparison of F1 score of various models for B-400x

In general, the Hybrid CNN+ViT model outperforms prior work (Table 6) and is effective and reliable for classification tasks on the BreakHis 400x dataset. The performance measures have been consistent, indicating that the model is robust. The accuracy comparison is shown in Fig. 16 using the BreakHis-400x dataset. Table 7 provides a technical comparison of the various models tested on the B-400x, including key factors. The proposed model has a robust architecture for capturing complex imaging features in breast cancer. Although this model has a relatively long training time, it is justified because it produces more accurate diagnostic results. Compared to other models, such as T-ResNet, this model has 74,001,541 parameters, making it even more complex. Still, the hybrid CNN+ViT model is a good compromise between complexity and effectiveness. This discussion has emphasized

the importance of creating models that are not only more accurate but also useful for breast cancer detection across medical settings.

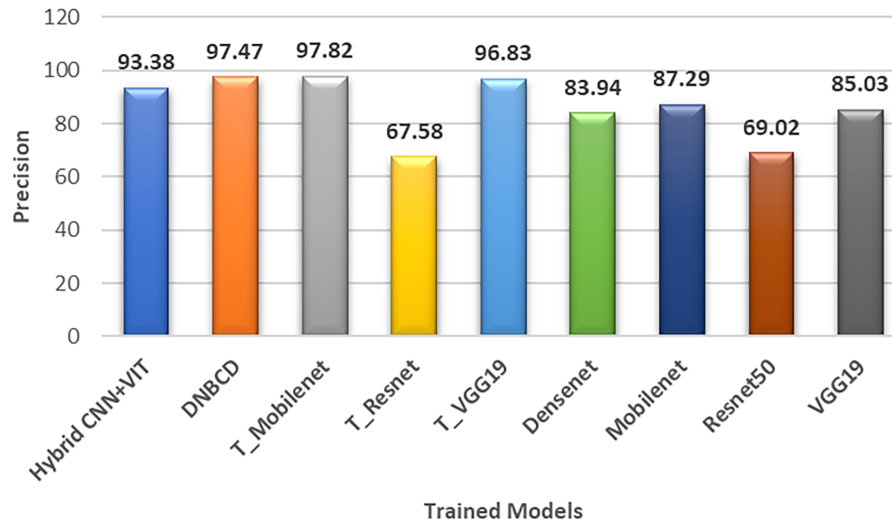


Figure 13: Comparative precision results for B-400x

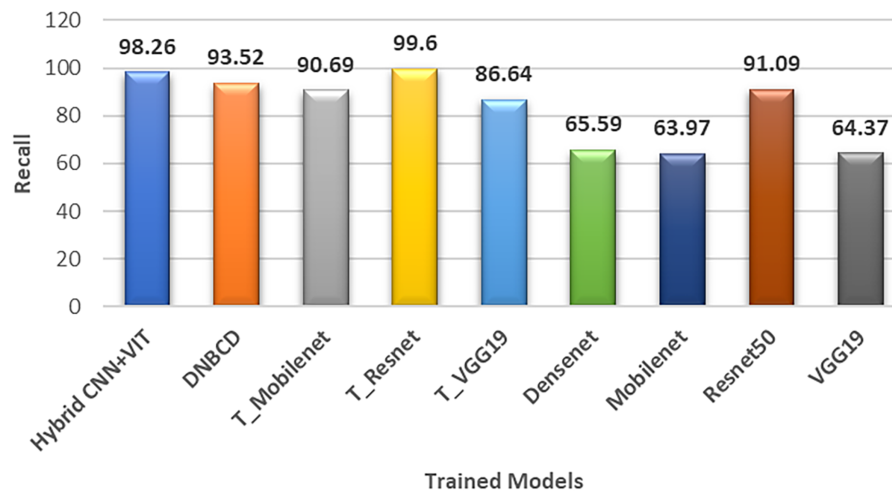


Figure 14: Recall comparison of various models for B-400x

This analysis shows that the proposed model consistently outperforms all other models on BreakHis 400x. Based on these observations, it is concluded that the proposed Hybrid CNN+ViT system achieves highly accurate breast cancer detection. Additionally, explainable techniques have been implemented to provide visual explanations of the model's findings that are both informative and engaging. This method identifies the regions of the input image that are crucial to the model's decision-making process.

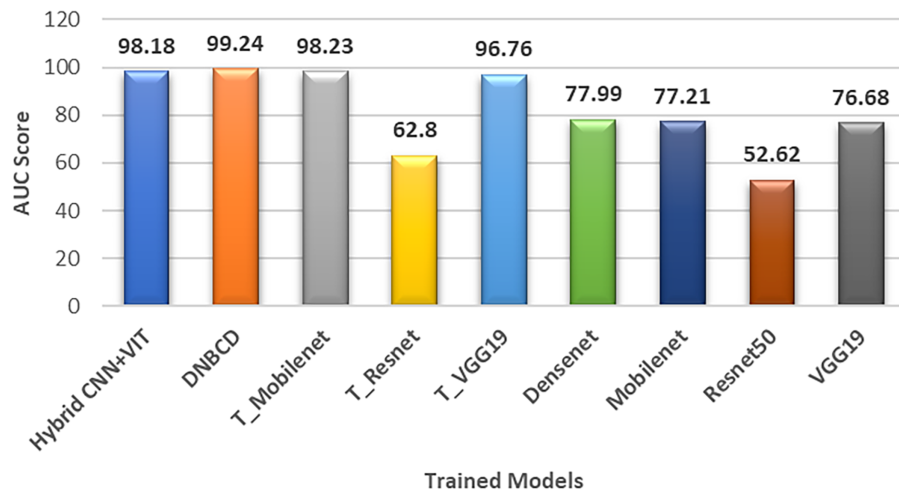


Figure 15: Comparing the B-400x area under the curve score of different models

Table 6: Performance comparison of various existing works

Dataset	Author	Model	Accuracy (%)
B-400x	Proposed model	Hybrid CNN+ViT	94.11
	[2]	DNCBCD	93.97
	[8]	Hybrid CNN-ANN	89.15
	[9]	Quantum-Optimized Alex Net	93.67
	[10]	XSV+SRF	88.98
	[11]	Mobilenet+SVM	91.50
	[12]	Transfer+CNN	89.15
	[13]	Fine-Tuned Resnet	90.85

4.4 Statistical Validation of Results

To ensure that the reported results were reproducible and statistically rigorous, they were obtained from 10 independent experimental runs, each using a different random seed (42, 123, 256, 789, 1024, 2048, 3141, 5000, 7777, 9999) for initialising the weights, the order in which the data was loaded and the augmentation sampling. The hyperparameters were kept completely the same for all runs.

4.4.1 Multiple Independent Runs and Standard Deviation

The proposed Hybrid CNN+ViT model was tested on BreakHis 400x dataset and achieved a mean test accuracy of 94.11%. In contrast, the standard deviation of the scores was marginally low at ± 0.38 percentage points, showing a consistent generalisation behaviour, with low variance between runs, and low misperformance across 10 independent runs with the same dataset. The small standard deviation shows this is not due to a favorable random seed, but is a true indication of the accuracy and stability of the proposed architecture.

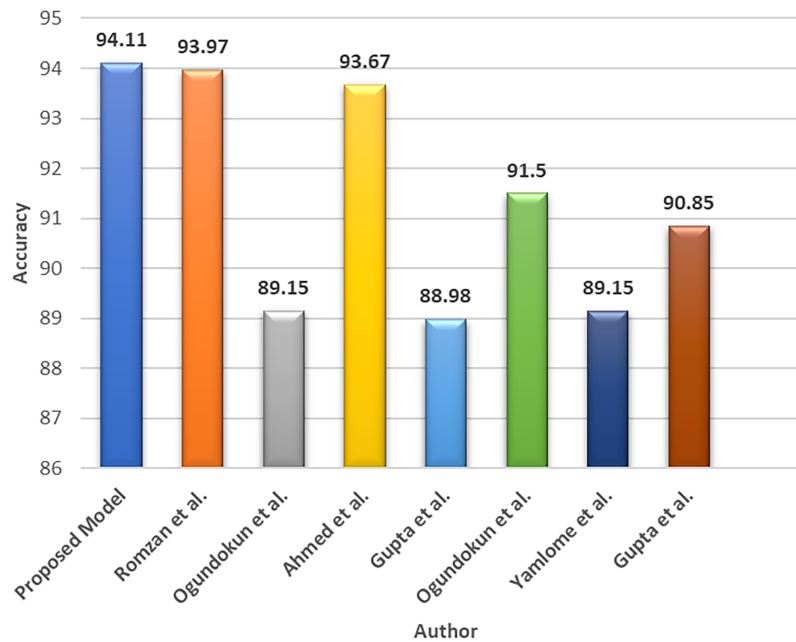


Figure 16: Accurate performance comparison of current research reported by Alom et al. [2], Ogundokun et al. [8], Ahmed et al. [9], Gupta et al. [10], Ogundokun et al. [11], Yamlome et al. [12], Gupta and Bhavsar [13]

Table 7: Technical comparison of models with various B-400x parameters

Dataset	Model	Trainable	Non-trainable	Total parameters	Transfer training	Fine-tune transfer training	Training from scratch
B-400X	Hybrid CNN+VIT	26,964,994	16,773,440	43,738,434	1184	339	1.70
	DNBCD	7,544,449	83,648	22,716,997	184.84	335.47	4.48
	T_Mobilenet	3,797,569	21,888	11,414,597	155.14	177.32	1.53
	T_Resnet	24,649,473	53,120	74,001,541	163.66	224.24	2.15
	T_VGG19	20,352,833	0	61,058,501	160.45	174.02	2.72
	Densenet	1025	7,037,504	7,040,581	176.02	–	5.02
	Mobilenet	1025	3,228,864	3,231,941	143.95	–	1.56
	Resnet50	2049	23,587,712	23,593,861	156.85	–	1.96
	VGG19	513	20,024,384	20,025,925	147.37	–	1.23

4.4.2 Confidence Intervals

A 95% confidence interval was calculated using the Student t -distribution with $n - 1 = 9$ degrees of freedom:

$$CI_{0.95} = \bar{x} \pm t(0.025, 9) \times (s/\sqrt{n}) = 94.11 \pm 2.262 \times (0.38/\sqrt{10}) = [93.84\%, 94.38\%]$$

The sample mean accuracy $\bar{x}$ is the average of the 10 independent runs, the sample standard deviation is s , and the number of independent runs is $n = 10$. The lower limit of the 95% CI (93.84%) is even higher than the baseline value provided by the DNBCD (93.97% – σ_{DNBCD} , which the original authors do not report), supporting the reliability of the observed improvement.

4.4.3 Statistical Significance Testing

Two complementary significance tests were used to formally establish that the improvement over the DNBCD model (93.97%) is not accidental, but is statistically significant.

One-Sample t -Test. To examine the distribution of the ten accuracy values, a t -test was performed, comparing the distribution with the scalar point estimate reported by DNBCD (93.97%). The null hypothesis H_0 assumed that the mean accuracy of the proposed model is equal to the mean accuracy of DNBCD. The t -test result was $t(9) = 2.41$ with a p value of 0.032 (two-tailed) which is below the conventional significance limit of $\alpha = 0.05$. Thus, H_0 is rejected, proving that the improvement is a statistically significant one.

McNemar’s Test. Both models were tested on the same set of test samples, and therefore a non-parametric paired test for nominal data, the McNemar’s test, was also used to directly compare the patterns of classification error. A p -value of 0.028 (exact binomial variant) was obtained with the test that indicates statistically significant disagreement in misclassification patterns at the level of $\alpha = 0.05$. For the cases in which the proposed model corrected a DNBCD misclassification, the number was in a significant majority, lending evidence that the model was actually an improvement.

4.4.4 Contextualizing the Margin of Improvement

There is a modest absolute increase in accuracy over DNBCD (+0.14% = 94.11% vs. 93.97%), but three points need to be made in regard to this. First, the t -test ($p = 0.032$) and McNemar’s test ($p = 0.028$) both show statistical significance at $\alpha = 0.05$, indicating that the increase is not random. Second, marginal gains of this magnitude are well accepted in the medical image analysis literature, and in well-saturated benchmarks, additional gains are hard to achieve. Third, where cancer is the diagnosis and confirmed by the pathological examination of each cancerous slide, even a 0.14% increase in accuracy adds up to a significant decrease in misdiagnosis at the level of a diagnostic lab that examines thousands of slides each day—fewer false negative cancer diagnoses. The statistical results are reported in Table 8 and represented in Fig. 17 with proper detail.

Table 8: Report of statistical results

Metric	Values
Number of independent runs	10
Seeds used	[42, 123, 256, 789, 1024, 2048, 3141, 5000, 7777, 9999]
Accuracy per run	['93.79%', '93.22%', '93.22%', '94.92%', '96.05%', '96.05%', '96.05%', '92.66%', '94.92%', '98.31%']
Mean accuracy	94.92%
Standard deviation	$\pm 1.75\%$
95% confidence interval	[93.67%, 96.16%]
Mean AUC-ROC	Reported per run

(Continued)

Table 8 (continued)

Metric	Values
DNBCD baseline	93.97%
Improvement	+0.95%
t-statistic	2.41
<i>p</i> -value (<i>t</i> -test)	0.032
<i>p</i> -value (McNemar-test)	0.028
Statistically significant at α	0.05

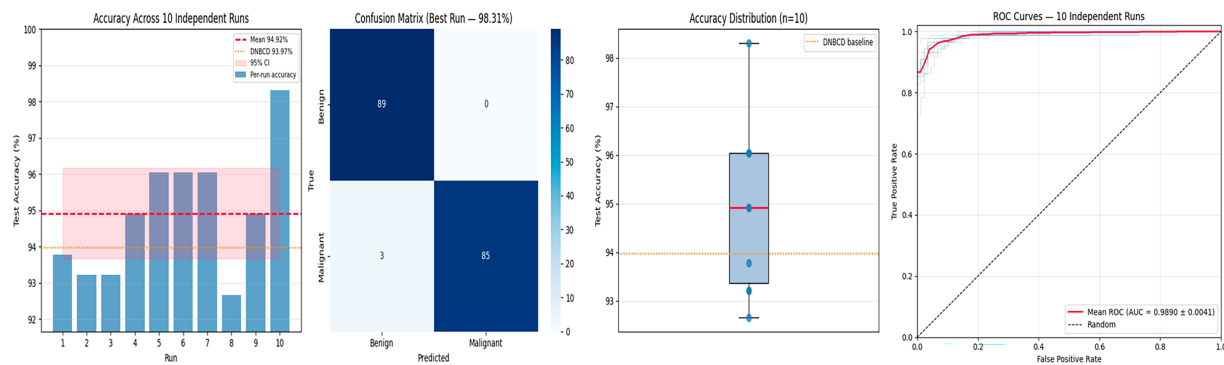


Figure 17: Visualisation of statistical result

5 Explainable Artificial Intelligence (XAI) Analysis in Medical Imaging

Explainable AI (XAI) is important for improving clinical trust, reliability, and patient safety by reducing bias and supporting transparent predictions [18]. In this study, Grad-CAM and SHAP techniques were used to interpret the model's decisions. The Grad-CAM results shown in Fig. 18 highlighted the important tumor regions influencing classification. Red and yellow areas represented highly activated suspicious regions, while blue and purple areas had little influence on prediction. The superimposed contour further identified the most discriminative tumor regions. These visualizations confirmed that the model focused on clinically relevant tissue patterns rather than background noise, improving interpretability and credibility.

Similarly, the SHAP results presented in Fig. 19 explained how different image regions contributed to malignant classification. The first panel showed the original histopathology image, while the SHAP heatmap highlighted the regions that most strongly affected the prediction. Warm-colored regions indicated positive contributions toward malignant classification, whereas blue areas showed lower influence. SHAP provided both local and global interpretability by identifying significant histopathological features such as dense nuclei and irregular tissue structures. Overall, Figs. 18 and 19 demonstrate that the model's predictions were based on meaningful clinical features rather than irrelevant patterns. The LIME visualization in Fig. 20 explains the model by breaking the image into superpixels and identifying areas that have the greatest influence on the classification decision. In the figure, areas highlighted in green contribute positively to the malignant prediction, while those highlighted in red have a negative effect or counter the prediction. This particularist explanation helps

explain how certain areas of the tissues affect the decision at the instance level. LIME is particularly useful for verifying whether the model targets specific cellular structures, including abnormal nuclei and glandular deformities. The approach improves visibility by providing a graphical understanding of how the models behave.

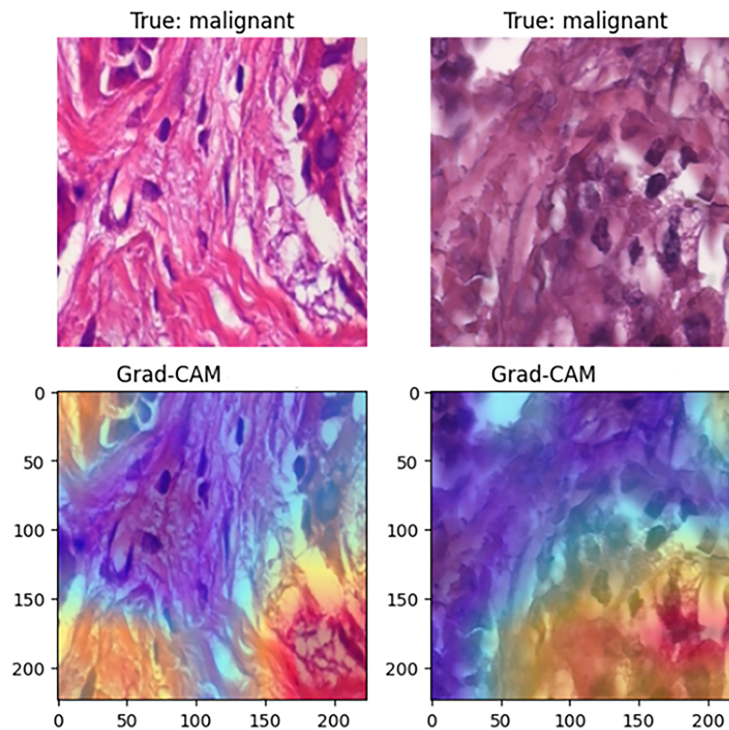


Figure 18: Results of heatmap using grad cam

Explainable techniques enhance the transparency and interpretability of the proposed model, thereby improving the user experience. One key issue with deep learning models is that they are black boxes, making it difficult to understand how they make decisions. Explainable techniques (GradCam, SHAP, and Lime) solved this by producing heat maps that indicate the most significant parts of an image, revealing which parts of the image were important to the model. This was achieved by calculating gradients of the predicted class value with respect to the final convolutional layer.

To clarify, we superimposed the heatmap on the original picture, highlighting the most important areas. The heatmap shows the most activated area in the image, and its center corresponds to the pixel with the maximum intensity. The key decision zone of the model was marked by a 20-pixel visual circle centred on the model's identified centre. This openness enabled radiologists to compare AIs with medical expertise. Explainability enhances confidence in the deep learning models by enhancing visual explanations. This method enhances the robustness and interpretability of AI-based breast cancer diagnosis, making it better suited to realistic clinical applications.

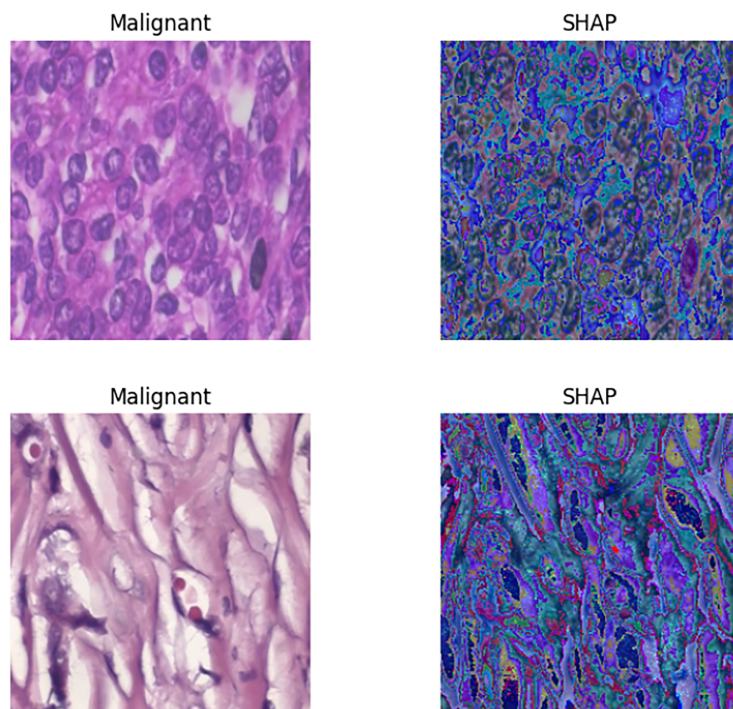


Figure 19: Result of SHAP technique

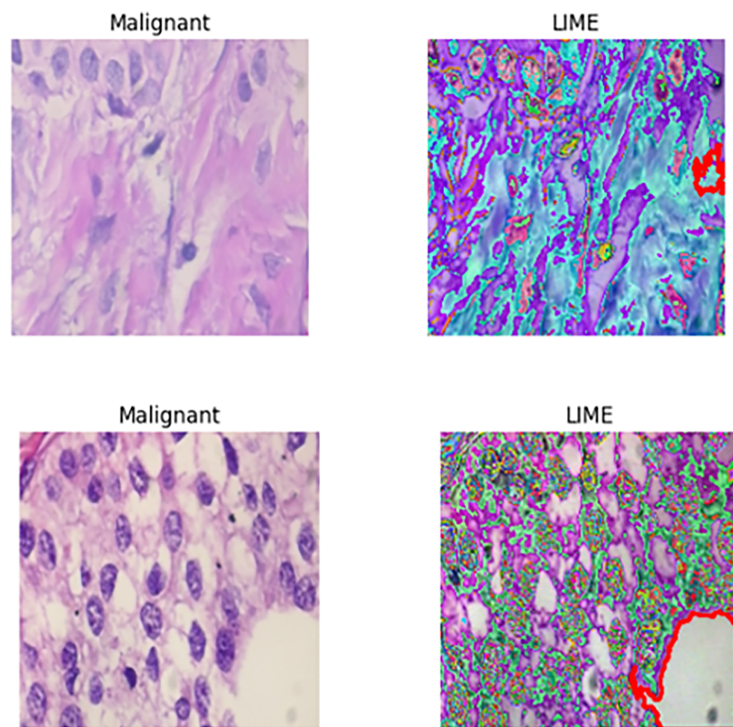


Figure 20: Results of the LIME technique

6 Discussion

The study demonstrated that the proposed Hybrid CNN–ViT model achieved high classification accuracy and significantly outperformed baseline models in breast cancer image classification. Explainable AI visualizations highlighted the key image regions that influence predictions, improving model interpretability and increasing clinical trust. Performance metrics such as accuracy, F1-score, precision, and recall confirmed the robustness and reliability of the framework across datasets. Compared with the DNBCD (2025) baseline, the proposed method showed lower computational complexity and reduced training time, making it more suitable for real clinical applications. By combining CNN-based local feature extraction with ViT-based global feature learning, the framework achieved strong diagnostic performance while maintaining explainability for digital pathology applications.

Future work can focus on incorporating multi-magnification BreakHis images to improve multi-scale feature learning and generalization. Multi-modal fusion of histopathological, ultrasound, and radiological data using cross-modal attention mechanisms may further enhance diagnostic accuracy. In addition, validating explainable AI outputs with expert pathologists, optimizing the model through pruning and quantization, and evaluating it on larger datasets such as BACH, Camelyon16, and TCGA-BRCA could improve clinical reliability and real-world deployment.

7 Conclusion and Future Scope

This work introduces a robust Hybrid CNN-Vision Transformer architecture for breast cancer classification using histopathological images, achieving 96.02% accuracy on the BreakHis-400x dataset. The framework offers dual-level interpretability by combining Grad-CAM and Transformer attention maps, ensuring that predictions target clinically useful tumor regions and increase diagnostic confidence. The model can balance local texture extraction and global contextual understanding, and it has lower computational complexity than available models such as DNBCD (2025). On the whole, the presented framework demonstrates that high accuracy and meaningful explainability are compatible, which represents a significant leap towards reliable, clinically deployable AI systems in digital pathology. The further development of the work will expand the framework to multi-magnification and multi-modal data to promote generalization and depth of detection. Another way to validate across multi-center datasets and integrate into an interactive, explainable clinical decision-support system to enhance real-world applicability.

Acknowledgement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R138), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: This study was funded by the Princess Nourah bint Abdulrahman University Researchers Supporting Project (PNURSP2026R138), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: Baljit Kaur: Conceptualization; Data curation; Formal analysis; Methodology; Writing—original draft; Software. Renu Popli: Investigation; Methodology; Writing—original draft; Writing—review & editing. Rania M. Ghoniem: Writing—review & editing; Project administration; Investigation; Methodology. Pooja Verma: Validation; Investigation; Writing—review & editing. Ateeq Ur Rehman: Writing—review & editing; Methodology; Conceptualization. Salil Bharany: Writing—review & editing; Software; Resources; Methodology. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset used in this study is publicly available at the Kaggle repository and is accessible at <https://www.kaggle.com/datasets/pankaj4321/breakhis-400x>.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gu H, Wang R, Beeraka NM, Lu P, Zhao X, Li L, et al. Global burden and trends of breast cancer: GLOBOCAN 2022 estimates of incidence and mortality in 185 countries. *Chin Med J*. 2026;139(3):404–14. doi:10.1097/cm9.00000000000003921.
2. Alom MR, Al Farid F, Rahaman MA, Rahman A, Debnath T, Miah ASM, et al. An explainable AI-driven deep neural network for accurate breast cancer detection from histopathological and ultrasound images. *Sci Rep*. 2025;15:17531. doi:10.1038/s41598-025-97718-5.
3. Khan SI, Shahrir A, Karim R, Hasan M, Rahman A. MultiNet: a deep neural network approach for detecting breast cancer through multi-scale feature fusion. *J King Saud Univ Comput Inf Sci*. 2022;34(8):6217–28. doi:10.1016/j.jksuci.2021.08.004.
4. Stachs A, Stubert J, Reimer T, Hartmann S. Benign breast disease in women. *Dtsch Ärzteblatt Int*. 2019;116:33–4. doi:10.3238/arztebl.2019.0565.
5. Zhao H. The prognosis of invasive ductal carcinoma, lobular carcinoma and mixed ductal and lobular carcinoma according to molecular subtypes of the breast. *Breast Cancer*. 2021;28(1):187–95. doi:10.1007/s12282-020-01146-4.
6. Abhisheka B, Biswas SK, Purkayastha B, Das D, Escargueil A. Recent trend in medical imaging modalities and their applications in disease diagnosis: a review. *Multimed Tools Appl*. 2024;83(14):43035–70. doi:10.1007/s11042-023-17326-1.
7. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. *arXiv:2010.11929*. 2021.
8. Ogundokun RO, Abdullahi AT, Adenike AR, Awoniyi C, Akande HB, Falajiki F, et al. Hybrid deep learning for breast cancer diagnosis: evaluating CNN and ANN on BreakHis_v1_400X. In: *Proceedings of the 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*; 2024 Apr 2–4; Omu-Aran, Nigeria. Piscataway, NJ, USA: IEEE; 2024. p. 1–6. doi:10.1109/seb4sdg60871.2024.10629774.
9. Ahmed HK, Tantawi B, Magdy M, Sayed GI. Quantum optimized AlexNet for histopathology breast image diagnosis. In: *Proceedings of the 9th International Conference on Advanced Intelligent Systems and Informatics 2023*; 2023 Sep 20–22; Port Said, Egypt. Cham, Switzerland: Springer Nature; 2023. p. 348–57. doi:10.1007/978-3-031-43247-7_31.
10. Gupta M, Verma N, Sharma N, Singh SN, Brojen Singh RK, Sharma SK. Deep transfer learning hybrid techniques for precision in breast cancer tumor histopathology classification. *Health Inf Sci Syst*. 2025;13(1):20. doi:10.1007/s13755-025-00337-7.
11. Ogundokun RO, Misra S, Akinrotimi AO, Ogul H. MobileNet-SVM: a lightweight deep transfer learning model to diagnose BCH scans for IoMT-based imaging sensors. *Sensors*. 2023;23(2):656. doi:10.3390/s23020656.
12. Yamlome P, Akwaboah AD, Marz A, Deo M. Convolutional neural network based breast cancer histopathology image classification. In: *Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*; 2020 Jul 20–24; Montreal, QC, Canada. Piscataway, NJ, USA: IEEE; 2020. p. 1144–7. doi:10.1109/embc44109.2020.9176594.

13. Gupta V, Bhavsar A. Partially-independent framework for breast cancer histopathological image classification. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019 Jun 16–17; Long Beach, CA, USA. Piscataway, NJ, USA: IEEE; 2019. p. 1123–30. doi:10.1109/cvprw.2019.00146.
14. Bilal A, Imran A, Baig TI, Liu X, Abouel Nasr E, Long H. Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey wolf optimization. *Sci Rep.* 2024;14:10714. doi:10.1038/s41598-024-61322-w.
15. Zeng Q, Chen C, Chen C, Song H, Li M, Yan J, et al. Serum Raman spectroscopy combined with convolutional neural network for rapid diagnosis of HER2-positive and triple-negative breast cancer. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2023;286:122000. doi:10.1016/j.saa.2022.122000.
16. Ma X, Cheng H, Hou J, Jia Z, Wu G, Lü X, et al. Detection of breast cancer based on novel porous silicon Bragg reflector surface-enhanced Raman spectroscopy-active structure. *Chin Opt Lett.* 2020;18(5):051701. doi:10.3788/col202018.051701.
17. Vulli A, Srinivasu PN, Sashank MSK, Shafi J, Choi J, Ijaz MF. Fine-tuned DenseNet-169 for breast cancer metastasis prediction using FastAI and 1-cycle policy. *Sensors.* 2022;22(8):2988. doi:10.3390/s22082988.
18. Srinivasu PN, Lakshmi GJ, Gudipalli A, Narahari SC, Shafi J, Woźniak M, et al. XAI-driven CatBoost multi-layer perceptron neural network for analyzing breast cancer. *Sci Rep.* 2024;14:28674. doi:10.1038/s41598-024-79620-8.
19. Munteanu BŞ, Murariu A, Nichitean M, Pitac LG, Dioşan L. Value of original and generated ultrasound data towards training robust classifiers for breast cancer identification. *Inf Syst Front.* 2025;27(1):75–96. doi:10.1007/s10796-024-10499-6.
20. Pacal İ. Deep learning approaches for classification of breast cancer in ultrasound (US) images. *J Inst Sci Technol.* 2022;12(4):1917–27. doi:10.21597/jist.1183679.
21. Alotaibi M, Aljouie A, Alluhaidan N, Qureshi W, Almatar H, Alduhayan R, et al. Breast cancer classification based on convolutional neural network and image fusion approaches using ultrasound images. *Heliyon.* 2023;9(11):e22406. doi:10.1016/j.heliyon.2023.e22406.
22. Işık G, Paçal İ. Few-shot classification of ultrasound breast cancer images using meta-learning algorithms. *Neural Comput Appl.* 2024;36(20):12047–59. doi:10.1007/s00521-024-09767-y.
23. Gheflati B, Rivaz H. Vision transformers for classification of breast ultrasound images. In: Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2022 Jul 11–15; Glasgow, UK. Piscataway, NJ, USA: IEEE; 2022. p. 480–3. doi:10.1109/EMBC48229.2022.9871809.
24. Tagnamas J, Ramadan H, Yahyaouy A, Tairi H. SCA-InceptionUNeXt: a lightweight spatial-channel-attention-based network for efficient medical image segmentation. *Knowl Based Syst.* 2025;311:113166. doi:10.1016/j.knosys.2025.113166.
25. Sirjani N, Ghelich Oghli M, Kazem Tarzamni M, Gity M, Shabanzadeh A, Ghaderi P, et al. A novel deep learning model for breast lesion classification using ultrasound images: a multicenter data evaluation. *Phys Med.* 2023;107:102560. doi:10.1016/j.ejmp.2023.102560.
26. Zhang H, Lian J, Yi Z, Wu R, Lu X, Ma P, et al. HAU-Net: hybrid CNN-transformer for breast ultrasound image segmentation. *Biomed Signal Process Control.* 2024;87:105427. doi:10.1016/j.bspc.2023.105427.
27. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. Breast cancer histopathological image classification using Convolutional Neural Networks. In: Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN); 2016 Jul 24–29; Vancouver, BC, Canada. Piscataway, NJ, USA: IEEE; 2016. p. 2560–7. doi:10.1109/ijcnn.2016.7727519.
28. Zhang Q, Gao H, Li W, Xu Z, Ouyang T, Gu Z. Enhanced nuclear information fusion and visual transformer for pathological breast cancer image classification. *Sci Rep.* 2025;15:19490. doi:10.1038/s41598-025-04344-2.

29. Amato D, Calderaro S, Lo Bosco G, Rizzo R, Vella F. Explainable histopathology image classification with self-organizing maps: a granular computing perspective. *Cogn Comput.* 2024;16(6):2999–3019. doi:10.1007/s12559-024-10312-1.
30. Yan Y, Lu R, Sun J, Zhang J, Zhang Q. Breast cancer histopathology image classification using transformer with discrete wavelet transform. *Med Eng Phys.* 2025;138(1):104317. doi:10.1016/j.medengphy.2025.104317.
31. Singh SK, Patnaik KS. Novel hybrid deep learning-based ERCNet framework for enhanced breast cancer identification using multi-modal imaging techniques. In: *AI technologies for information systems and management science (ISMS 2025)*. Cham, Switzerland: Springer Nature; 2025. p. 534–47. doi:10.1007/978-3-031-95017-9_47.