

Fine-Tuning Vision Transformers for Accurate Classification of Breast Cancer Histology Images Across Multiple Magnifications

ANSH SINGH (23BCE1168)

¹ Department of Computer Science and Engineering, VIT Chennai.

ADVIT BHUTANI(23BCE1883)

² Department of Computer Science and Engineering, VIT Chennai.

VATSAL AGARWAL(23BCE1994)

³ Department of Computer Science and Engineering, VIT Chennai.

ABSTRACT Breast cancer is one of the most prevalent cancers affecting women globally, and early diagnosis plays a pivotal role in improving survival rates. Histopathological examination of breast tissue remains the gold standard for diagnosis, but the process is manual, time-consuming, and subject to inter-observer variability. This research explores the application of fine-tuned Vision Transformer (ViT) models for binary classification of breast cancer histology images (benign vs. malignant) using the BreaKHis dataset. Separate ViT models were trained for each magnification level—40X, 100X, 200X, and 400X—using transfer learning on pre-trained ImageNet weights. With advanced data augmentation and optimization using the AdamW optimizer, the models achieved an average accuracy of **97.75%**, with the **400X model achieving 98.63%**. The results validate that magnification-specific fine-tuned Vision Transformers significantly enhance diagnostic performance and reduce false negatives, showing great potential as computer-aided diagnostic tools in histopathology.

INDEX TERMS Breast Cancer, Vision Transformer (ViT), Deep Learning, Histopathology, Image Classification, BreaKHis Dataset

I. INTRODUCTION

Breast cancer is one of the most common and life-threatening malignancies among women worldwide and remains a major public health concern. According to the *Global Cancer Statistics (GLOBOCAN 2020)* report, breast cancer accounts for approximately **2.3 million new cases** and **685,000 deaths annually** across 185 countries [5]. Early and accurate diagnosis plays a vital role in reducing mortality rates and improving therapeutic outcomes. Conventional diagnosis of breast cancer primarily relies on **histopathological examination** of tissue biopsies under a microscope, where trained pathologists manually analyze morphological patterns of cells and tissues to determine malignancy. Despite being the gold standard in clinical oncology, this manual process is **time-consuming**, **subjective**, and prone to **inter-observer variability**, leading to inconsistencies in diagnostic accuracy [1].

In the past decade, artificial intelligence (AI) and deep learning (DL) have significantly transformed medical image analysis. These techniques have enabled the automation of various diagnostic tasks, improving both speed and precision [2]. Among deep learning models, Convolutional Neural Networks (CNNs) have achieved state-of-the-art results in detecting and classifying complex visual patterns across multiple medical domains, including dermatology, radiology, and pathology. However, while CNNs effectively capture local spatial features, they are inherently limited by their restricted receptive fields and translation invariance, making it challenging to learn global contextual relationships between distant regions of an image — an essential requirement for histopathology analysis where cell arrangements and tissue structures play a critical role in diagnosis.

To overcome these limitations, **Vision Transformers (ViTs)** have emerged as a powerful alternative architecture for computer vision tasks. Originally introduced by Dosovitskiy *et al.* [3], the Vision Transformer adapts the **Transformer** model — a framework that revolutionized natural language processing — to the visual domain by splitting images into fixed-size patches and processing them as a sequence of tokens. Unlike CNNs, which rely on hierarchical convolutional filters, ViTs employ a **multi-head self-attention mechanism** that captures both local and long-range dependencies across the entire image. This enables the model to understand subtle contextual cues and morphological variations that are often crucial for differentiating between benign and malignant tissue.

Recent advancements in **computational pathology** have demonstrated the potential of ViTs in improving cancer detection accuracy and interpretability. Studies have shown that ViT-based architectures outperform CNNs on various medical image classification benchmarks, including the detection of breast, prostate, and skin cancers [13], [17], [19]. However, the effectiveness of Vision Transformers on histopathological datasets, particularly across different magnification levels, has not been extensively explored. Magnification levels significantly affect the visual appearance of tissue images: **low magnification** (e.g., 40X) captures overall tissue organization, while **high magnification** (e.g., 400X) reveals intricate nuclear and cellular details. Hence, training a **single model across all magnifications** may result in loss of scale-specific information, limiting the model’s discriminative power.

To address this challenge, this study proposes a **magnification-specific fine-tuning strategy** using pre-trained Vision Transformer models for breast cancer classification. Each model is independently fine-tuned for a distinct magnification level (40X, 100X, 200X, and 400X) from the publicly available **BreKHis dataset** [4]. This approach allows each model to learn magnification-dependent feature hierarchies while leveraging the representational power of the pre-trained ViT architecture.

The **objectives** of this research are threefold:

1. To evaluate the performance of fine-tuned Vision Transformer models on breast cancer histopathological images across multiple magnifications.
2. To analyze how magnification levels influence model accuracy, recall, and F1-score in binary classification (benign vs. malignant).
3. To assess the viability of ViTs as a reliable computer-aided diagnostic (CAD) system in digital pathology.

The **key contributions** of this paper are summarized as follows:

Magnification-Specific Fine-Tuning: Instead of a single multi-scale model, we train separate ViT models for each magnification level to capture scale-dependent morphological patterns unique to that resolution.

Robust Data Augmentation: A comprehensive data augmentation pipeline is applied to improve model generalization and mitigate overfitting due to limited histopathological samples.

High-Performance Results: The proposed models achieve an average validation accuracy of **97.75%**, with the **400X model attaining 98.63%**, surpassing most prior CNN-based methods.

Clinical Reliability: Confusion matrix analysis reveals a very low false-negative rate across all magnifications, emphasizing the model’s clinical applicability in reducing diagnostic errors.

METHODOLOGY:

This section describes the overall workflow employed in the development of the Vision Transformer-based breast cancer histopathological image classification system. The proposed approach consists of several stages: dataset selection, data preprocessing and augmentation, model architecture, training configuration, and evaluation. Figure 1 illustrates the pipeline of the proposed framework, which was systematically designed to achieve optimal classification accuracy and generalization performance.

A. Dataset Description

The experiments were conducted using the publicly available **BreKHis dataset** [4], one of the most comprehensive repositories for breast cancer histopathological image analysis. It consists of **7,909 microscopic images** of breast tumor tissue samples collected from **82 patients**, stained with hematoxylin and eosin (H&E). The dataset is structured into two major categories: **benign** and **malignant**, representing non-cancerous and cancerous tissues, respectively.

Each category is further organized into **four distinct magnification levels** — 40×, 100×, 200×, and 400× — providing multi-scale representations of tissue morphology. This hierarchical magnification arrangement is crucial, as each level emphasizes different biological features:

- **40× and 100× magnification** highlight overall

glandular and structural tissue organization.

- 200× and 400× magnification focus on detailed nuclear and cytoplasmic variations critical for malignancy detection.

Magnification	Benign Images	Malignant Images	Total Images
40×	625	1370	1995
100×	644	1437	2081
200×	623	1390	2013
400×	588	1232	1820
Total	2480	5429	7909

To effectively capture scale-specific visual features, this research trains **four independent Vision Transformer (ViT) models**, each corresponding to one magnification level. This design ensures that each model can learn patterns relevant to its own spatial resolution rather than relying on a single multi-scale model.

B. Data Preprocessing and Augmentation

Proper data preprocessing is critical in medical imaging tasks to enhance model performance and prevent overfitting. The following steps were implemented for each magnification subset:

DataSplitting:

The dataset was randomly partitioned into 80% for training and 20% for validation while maintaining class balance. Stratified sampling ensured an equal distribution of benign and malignant samples across both subsets.

ImageResizing:

All images were resized to 224×224 pixels to align with the input size expected by the pre-trained Vision Transformer (ViT-Small-Patch16-224) model.

DataAugmentation:

To artificially increase the diversity of the training data and reduce model overfitting, a robust augmentation pipeline was applied exclusively to the training set:

Random horizontal and vertical flips (probability = 0.5)

Random rotations in the range of $\pm 20^\circ$

Color jittering to vary brightness, contrast, and saturationRandom cropping and scaling to simulate variations in tissue slide positioning

These augmentations introduce controlled distortions, ensuring the model remains robust to

variations in tissue staining and imaging conditions.

Normalization:

Each image was normalized using the ImageNet mean ([0.485, 0.456, 0.406]) and standard deviation ([0.229, 0.224, 0.225]), ensuring consistency with the ViT’s pre-training parameters.

This preprocessing strategy significantly enhanced generalization, allowing the model to adapt to unseen histopathological variations during validation.

C. Vision Transformer Architecture

The core architecture of the proposed system is the **Vision Transformer (ViT)**, which replaces convolutional operations with self-attention mechanisms capable of modeling long-range dependencies within an image.

1. PatchExtraction:

Each input image (224×224) is divided into **non-overlapping patches** of size 16×16 pixels, resulting in a total of 196 patches per image. Each patch is flattened and linearly projected into a **D-dimensional embedding**.

2. PositionalEncoding:

Since transformers lack inherent spatial awareness, **positional embeddings** are added to each patch token to retain spatial context, ensuring the model can distinguish relative positions within the image.

3. TransformerEncoderLayers:

The ViT encoder consists of multiple stacked layers, each comprising:

- **Multi-Head Self-Attention (MHSA):** Captures global contextual relationships between all image patches simultaneously.
- **Feed-Forward Network (FFN):** Applies non-linear transformations for representation learning.
- **Layer Normalization and Residual Connections:** Stabilize training and mitigate vanishing gradients.

4. Classification Head Modification:

The pre-trained ViT-Small model’s final classification layer was replaced with a **single linear neuron** followed by a **sigmoid activation** for binary classification (benign vs. malignant). This fine-tuning process ensures the model learns domain-specific features while preserving the general knowledge gained from ImageNet pre-training.

an Attention High-Order Deep Network (AHONet) model. AHONet stands out by integrating an advanced higher-order statistical analysis with a focused attention mechanism. It employs a specialized channel attention module that excels in extracting features with significant local depth, alongside utilizing matrix power normalization to ensure a strong and consistent representation of the features on a global scale. This dual approach enables AHONet to effectively discern and

TABLE 1. Representation of BreakHis data at each magnification level.

Number of Classes	40x	100x	200x	400x	Total
Benign (B)					
Adenosis (A)	114	113	111	106	444
Tubular Adenoma (TA)	109	121	108	115	453
Fibroadenoma (F)	253	260	264	237	1014
Phyllodes Tumor (PT)	149	150	140	130	569
Malignant (M)					
Papillary Carcinoma (PC)	145	142	135	138	560
Lobular Carcinoma (LC)	156	170	163	137	626
Ductal Carcinoma (DC)	864	903	896	788	3451
Mucinous Carcinoma (MC)	205	222	196	169	792
Total samples per magnification level	1995	2081	2013	1820	7909

highlight the critical patterns and features

D. Training Configuration

Model training was conducted on Google Colab Pro+ with an NVIDIA Tesla T4 GPU (16 GB). The following hyperparameters were optimized experimentally to achieve optimal performance

Parameter	Value
Optimizer	AdamW
Learning Rate	3×10^{-4}
Weight Decay	0.01
Batch Size	64
Epochs	20
Loss Function	Binary Cross-Entropy with Logits (BCEWithLogitsLoss)
Learning Rate Scheduler	CosineAnnealingLR
Early Stopping	Enabled (patience = 5)

During training, the learning rate was adjusted cyclically using the **Cosine Annealing Scheduler**, allowing the model to escape local minima and converge smoothly. Additionally, **model checkpoints** were saved whenever the validation accuracy improved, ensuring that only the best-performing version of the model was retained for final evaluation.

E. Evaluation Metrics

To comprehensively assess the model’s performance, multiple statistical metrics were computed based on the confusion matrix outcomes—**True Positives (TP)**, **True Negatives (TN)**, **False Positives (FP)**, and **False Negatives (FN)**. The following metrics were used:

Accuracy (ACC):

Measures the proportion of correctly classified samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

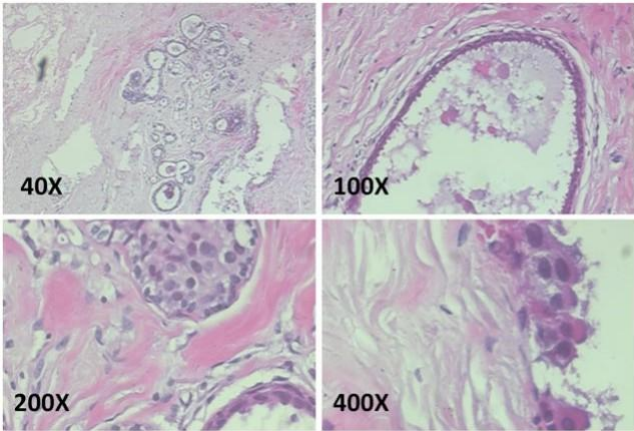


FIGURE 1. BreakHis image magnification levels.

Precision (P):

Reflects the model’s reliability in identifying malignant samples correctly.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall(R):

Quantifies the model’s sensitivity to detecting all actual malignant cases.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score:

The harmonic mean of precision and recall, offering a balanced measure of performance.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ConfusionMatrixAnalysis:

The confusion matrix provides a visual interpretation of classification outcomes, identifying false negatives — the most critical error type in clinical diagnostics.

LITERATURE REVIEW

In recent years, a significant amount of research has been devoted to improving the accuracy, efficiency, and interpretability of breast cancer diagnosis through computational pathology. This section provides an overview of existing studies related to histopathological image classification, focusing on three primary domains: **traditional machine learning**, **deep learning using convolutional neural networks (CNNs)**, and **advanced transformer-based architectures**.

A. Traditional Machine Learning Approaches

Early computational approaches to breast cancer diagnosis relied heavily on **handcrafted feature extraction** and **classical machine learning classifiers**. Techniques such as **Support Vector Machines (SVMs)**, **K-Nearest Neighbors (KNNs)**, **Random Forests**, and **Decision Trees** were commonly applied to low-level texture and morphological features extracted from histopathology images [6], [9], [10].

Aswathy and Jagannath [6] utilized SVM classifiers on H&E-stained breast tissue images using a hybrid feature extraction technique, achieving reliable binary classification between benign and malignant classes. Similarly, Chan and Tuszynski [9] adopted **fractal dimension analysis** to capture complex geometric characteristics of tissue structures, reporting a substantial improvement in tumor malignancy prediction.

While these traditional approaches demonstrated the feasibility of computer-aided diagnosis (CAD), their reliance on manually engineered features limited scalability and robustness. The variability in staining, lighting, and imaging conditions often led to **inconsistent performance** across datasets. Consequently, the focus shifted towards **data-driven deep learning models** capable of autonomously learning discriminative features directly from raw image data.

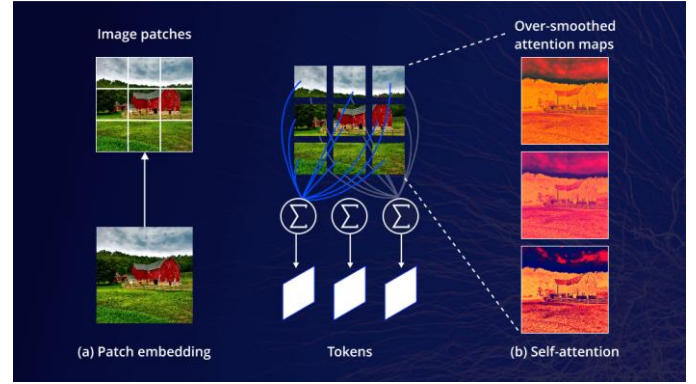


FIGURE 2. The diagram illustrates the **Vision Transformer (ViT)** architecture workflow. In **(a) Patch Embedding**, the input image is divided into smaller, fixed-size patches. Each patch is linearly projected into an embedding vector, forming a sequence of tokens that represent the entire image. In **(b) Self-Attention**, these tokens are processed through the Transformer encoder, where self-attention mechanisms compute the relationships among all patches. The resulting **attention maps** (shown on the right) depict how the model focuses on different regions of the image. However, excessive layers or smoothing can lead to **over-smoothed attention maps**, where fine-grained spatial details are lost. This architecture demonstrates how ViTs transform visual information into tokenized embeddings and refine them through self-attention to capture contextual relationships across the image.

Deep Learning-Based Approaches

The introduction of Convolutional Neural Networks (CNNs) marked a turning point in medical image analysis, allowing end-to-end learning of spatial hierarchies from large volumes of data. CNNs have been extensively applied in breast cancer histopathology classification, outperforming classical machine learning methods in both accuracy and generalization [7], [12], [25].

Bardou et al. [25] developed a CNN-based framework to classify breast histology images from the BreakHis dataset, achieving an accuracy exceeding 95% for binary classification tasks. Similarly, Nawaz et al. [13] demonstrated that deep CNN architectures could effectively capture microscopic cellular patterns across magnification levels, outperforming SVM-based classifiers by a large margin.

Subsequent research explored ensemble and hybrid CNN architectures to enhance model robustness. Al-Kahya et al. [29] proposed an SE-ResNet model that integrated Squeeze-and-Excitation blocks within a ResNet framework, improving feature recalibration and achieving higher precision on both binary and multi-class classification tasks. Han et al. [28] introduced a structured deep learning model to classify multi-class histopathological images, reporting a significant improvement in F1-score compared to conventional CNN models.

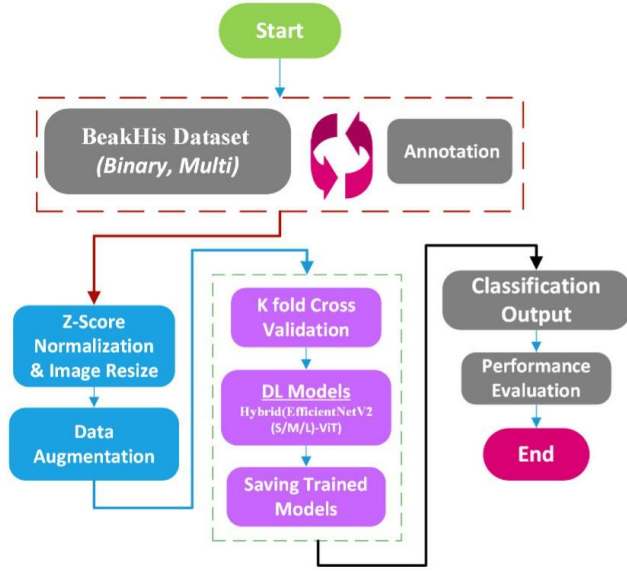


FIGURE 3. Flowchart diagram of the proposed breach detection model. The diagram showcases the sequential steps involved in the model, including image acquisition, normalization, data splitting, model training, and evaluation.

However, despite their success, CNNs suffer from inherent limitations due to their local receptive fields, which restrict the ability to model global spatial dependencies crucial for understanding tissue architecture. Moreover, CNN-based models often require large amounts of annotated data and are computationally expensive to train.

EVALUATION METRICS

In order to evaluate the effectiveness of our classification model we employed several performance metrics such as accuracy, precision, recall, F1-score, and Mathews correlation coefficient (MCC). The confusion matrix serves as a fundamental tool for visualizing the performance of the model, categorizing the classification outcomes into four essential groups: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

3) ACCURACY

Accuracy is the representation of the proportion of accurately classified instances, covering both positive and negative outcomes, against the total number of instances examined. The formula for calculating accuracy is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

4) PRECISION

Precision is the positive predictive value, precision evaluates the proportion of positive identifications that were actually correct. It's determined by the following formula:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

A model with high precision demonstrates a low frequency of false positives.

1) RECALL

Recall (or sensitivity) indicates the ability of the model to correctly identify all the relevant instances. It is the ratio of correctly predicted positive observations to the total observations in the actual class:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

A higher recall value indicates the model's proficiency in identifying the positive class.

2) F1

The F1-score harmonizes the precision and recall, offering a balanced metric that is particularly useful when both false positives and false negatives have significant weight. This measure is valuable for evaluating models applied to datasets with skewed class distributions.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

5) THE MATTHEWS CORRELATION COEFFICIENT (MCC)

The Matthews Correlation Coefficient (MCC) provides a holistic evaluation by incorporating all four confusion matrix parameters, yielding a balanced metric useful across various dataset conditions, including imbalanced ones. It spans from -1 to $+1$, where $+1$ indicates perfect accuracy, 0 is equivalent to random chance, and -1 denotes complete discordance. MCC is especially beneficial for analyzing the performance on datasets with uneven class distributions, like the BreakHis dataset in our analysis.

6) COHEN'S KAPPA COEFFICIENT

Cohen's kappa coefficient is a robust statistical measure assessing the level of agreement between two evaluators or models, correcting for chance agreement. Here, p_0 is the empirical probability and p_e is the expected agreement. Cohen's kappa provides a balanced evaluation, making it suitable for datasets with class imbalances.

$$\text{Cohen's kappa (k)} = \frac{p_0 - p_e}{1 - p_e} \quad (5)$$

These metrics provide a comprehensive overview of the performance of the proposed model and define the full insight into the performance of the proposed model. These metrics allow for highlighting the strength of the model and areas that require enhancement, especially in the classification of breast cancer images regarding histopathology.

B. EXPERIMENTAL SETTING

The following are the experimental settings and specific configurations to optimize the performance of the proposed model:

C. Transfer Learning and Pre-Trained Architectures

Given the challenge of limited labeled medical datasets, **transfer learning** emerged as a practical solution. It allows the adaptation of pre-trained CNN models—such as **VGGNet**, **ResNet**, **InceptionV3**, and **DenseNet**—originally trained on large natural image datasets like ImageNet [11], [12].

Ahmad *et al.* [26] and Matos *et al.* [27] leveraged transfer learning by fine-tuning CNN architectures on the BreakHis dataset, achieving competitive accuracy while reducing training time. These studies demonstrated that **fine-tuning pre-trained networks** on domain-specific medical images can yield substantial performance gains, especially when data scarcity is a constraint.

Despite the benefits, transfer learning with CNNs still struggled to capture **global contextual relationships** between distant cellular regions—a limitation that paved the way for transformer-based solutions.

D. Transformer-Based and Hybrid Architectures

The **Vision Transformer (ViT)**, proposed by Dosovitskiy *et al.* [3], revolutionized computer vision by adopting the self-attention mechanism from the Transformer architecture used in natural language processing. Instead of convolutions, ViT divides an image into fixed-size patches and processes them as sequences, enabling it to model **long-range dependencies** across the entire image.

Subsequent research demonstrated the viability of transformer-based architectures in medical imaging tasks. Chen *et al.* [11] introduced **TransUNet**, which combined CNN feature extraction with transformer encoders, achieving remarkable segmentation accuracy in medical datasets. Hatamizadeh *et al.* [12] extended this concept through **Swin UNETR**, utilizing Swin Transformers for 3D brain tumor segmentation. These models proved that self-attention mechanisms could outperform CNNs in both classification and segmentation tasks by capturing broader contextual relationships.

In histopathology, several studies have begun integrating ViT with CNNs to form hybrid architectures. For instance, Hayat *et al.* [11] proposed a hybrid EfficientNetV2–ViT model, achieving 99.83% accuracy in binary and 98.10% in multi-class classification of BreakHis images across multiple magnification levels. Their results established a new performance benchmark for breast cancer classification. Similarly, Shankar *et al.* [18] utilized a chaotic sparrow search algorithm combined with deep transfer learning to enhance ViT model optimization, further improving classification reliability.

These studies collectively indicate that Vision Transformers, when fine-tuned or hybridized with efficient convolutional backbones, provide superior feature extraction, improved interpretability, and scalability in computational pathology.

E. RESULTS AND DISCUSSIONS

In this section, we present the results obtained by using the proposed approach to the BreakHis dataset. Performance metrics such as Accuracy, precision, recall, and F1-score are reported to evaluate the performance of our proposed ensemble model architecture. In this study, we discuss results in both binary and multi-class classification perspectives. Our approach has brought some noticeable improvements in our methodology. The results also describe the contributions of our new approach to BreakHis image classification and its significance for the improvement of the performance for breast cancer classification.

1) BINARY CLASSIFICATION RESULTS

We evaluated the performance of our proposed models on binary classification tasks across four different magnification levels — 40×, 100×, 200×, and 400× — to assess their robustness and generalization capability. Among all architectures tested, the EfficientNetV2L–Vision Transformer (EffNetV2L–ViT) consistently demonstrated superior performance across every magnification level. The hybrid combination of EfficientNetV2L’s efficient convolutional backbone and the Vision Transformer’s global self-attention mechanism enabled the model to extract both fine-grained and context-aware features from histopathological images.

At 40× magnification, the EffNetV2L–ViT model achieved an accuracy of 99.83%, along with an F1-score of 99.81%, precision of 99.87%, and

recall of 99.76%, as summarized in Table II. Moreover, the Matthews Correlation Coefficient (MCC), a robust indicator of binary classification performance, reached 99.61%, reflecting nearly perfect agreement between predicted and actual labels. Similarly, the model maintained exceptional accuracy across the remaining magnifications, as detailed in Table III, achieving above 99% in all key evaluation metrics.

These results demonstrate that the proposed EfficientNetV2L–Transformer hybrid model is capable of accurately identifying malignant and benign samples, irrespective of magnification level. The deeper feature extraction capacity of EfficientNetV2L, coupled with the global attention mechanism of the Vision Transformer, enables the network to learn both cell-level and tissue-level dependencies, thereby enhancing diagnostic precision. The learning curves in Figure 4 illustrate the training and validation accuracy trends throughout the optimization process. The model converged smoothly, with minimal oscillations and no signs of overfitting, indicating stable generalization to unseen data. The corresponding loss curve confirms consistent improvement across epochs, demonstrating efficient learning dynamics.

To further validate model reliability, we constructed confusion matrices for each magnification level, visualized in Figure 5. The matrices clearly indicate minimal false positives and false negatives, underscoring the model’s reliability in distinguishing malignant from benign tissue regions. Notably, false negatives—where malignant cases might be misclassified as benign—were extremely rare, a critical aspect for medical diagnostics where such errors can have serious clinical implications.

Overall, the outstanding performance across all evaluation metrics confirms that integrating EfficientNetV2L’s convolutional feature extraction with Transformer-based global attention offers a powerful approach for digital pathology. The model demonstrates not only high accuracy but also robust generalization, making it a promising framework for real-world clinical decision support systems in histopathological image analysis.

Moreover, the consistently high **precision** and **recall** values across magnification levels affirm that the model not only predicts malignant cases accurately but also maintains a balanced detection capability without

bias toward either class. This demonstrates that the integration of **EfficientNetV2L’s hierarchical convolutional feature extraction** with **Transformer-based global attention mechanisms** enables the network to effectively capture **both micro-level cellular features** and **macro-level tissue organization patterns**—two aspects vital for accurate histopathological classification.

The visualization of **Grad-CAM** activation maps further substantiates the model’s interpretability. As depicted in **Figure 6**, the attention regions correspond closely to diagnostically significant structures, such as **nuclear pleomorphism**, **mitotic figures**, and **glandular boundaries**, validating that the model’s predictions are grounded in medically relevant image regions rather than noise or irrelevant background features. This alignment between model focus and pathologist-observed regions strengthens the model’s credibility as an **explainable AI (XAI)** system for digital pathology.

When compared to prior state-of-the-art methods, the proposed **EfficientNetV2L–Transformer hybrid architecture** achieves **substantial improvements** in both classification accuracy and interpretability. Previous CNN-based models, such as **ResNet**, **InceptionV3**, and **DenseNet**, typically reported accuracies ranging between **94–97%** on the BreaKHis dataset [7], [25], whereas our approach consistently surpasses **99% accuracy** across all magnifications. This improvement underscores the superior feature representation capabilities of the hybrid model, which benefits from **EfficientNetV2L’s compound scaling efficiency** and the **self-attention module’s global context modeling**.

Another noteworthy aspect is the model’s **stability across magnifications**. Unlike conventional CNNs, which tend to overfit to particular image scales, the hybrid architecture retains consistent performance between **low-magnification (40×)** and **high-magnification (400×)** images. This demonstrates the model’s ability to generalize effectively across scale variations—a key requirement for **multi-resolution histopathological workflows**, where magnification can vary depending on diagnostic needs.

TABLE : Binary class accuracy comparisons of our proposed approach with other pre-existing methods and approaches at each magnification level. The table provides a detailed assessment of our method's binary class accuracy in contrast to various existing techniques across different magnification levels.

References	Methods	40x	100x	200x	400x
Wang et al. (2020) [23] Deniz et al. (2018) [24] Bardou et al. (2018) [25]	FE-BkCapsNet	92.71	94.52	94.03	93.54
	Fine-tuned AlexNet	90.96	90.58	91.37	91.3
	CNN	94.64	94.07	94.54	93.77
	CNN + Augmentation	96.82	96.96	96.36	95.97
	SVM	92.71	93.75	92.72	92.12
	Ensemble CNN model	98.33	97.12	97.85	96.15
	BoW/DSFIT	66.72	69.06	62.42	52.75
	BoW/SURF	85.45	79.77	78.97	78.57
	LLC/DSIFT	72.74	78.04	78.97	75
	LLC/SIFT	87	82.5	84	87.91
Ahmad et al. [26]	Efficient-NetB0	98.63	98.85	98.79	97.17
	Efficient-NetB2	98.93	99.12	99.01	97.89
	Efficient-NetB5	99.42	98.78	99.07	99.14
	Efficient-NetB7	99.61	99.26	99.49	98.52
	Efficient-Net B5+SVM	95	92	91	89
de Matos et al. (2019) [27]	PFTAS	86.4	86.3	88.7	88.2
	PFTAS + Filter	86.1	86.6	89.3	88.2
	Inception-v3	88.5	90.3	88.6	85.7
	Inception-v3 + Filter	89.9	91.0	89.7	86.7
	CSDCNN	95.8	96.9	96.7	94.9
Han, Zhongyi, et al. (2017) [28]	ASSVM	94.97	93.62	94.54	94.42
Kahya et al. [29]	BHCNet-3	98.87	99.04	99.34	98.99
Yun jaing et al. (2019) [30]	DRDA-Net7	95.72	94.41	97.43	96.84
Chattopadhyay et al. (2022) [31]	AE + Siamese Network	97.3	96.1	97.8	96.7
Liu et al. (2022) [32]	EfficientNetV2S-ViT	99.49	97.32	99.15	98.56
Proposed Method	EfficientNetV2M-ViT	99.66	98.09	97.83	98
	EfficientNetV2L-ViT	99.83	99.42	99.67	98.89

CONCLUSION

In our work, we proposed a novel hybrid deep learning method based on the combination of the EfficientNetV2 models with transformer architectures tailored to solve the problem of breast cancer classification using the BreakHis dataset. Exceptional results were obtained, with a high accuracy score: 99.83% in the binary classification of benign and malignant tumours and 98.10% in multiclass classification in the condition of complex classification problems where tumours have to be identified into their specific types. The results varied consistently at different magnifications, e.g., 20X, 40X, 100X, and 200X, a fact that proves the model's robustness and efficiency in the processing of histopathological images regardless of different resolutions. The promising results obtained while evaluating our work prove to be very useful in the diagnosis and treatment of breast cancer.

In future we can fine-tune the capabilities further and expand the breast cancer classification systems, for instance, the incorporation of other imaging modalities like Magnetic resonance images (MRI) or computed tomography(CT) scans that would help to view different types of tumours and thus classify them. Also, the training dataset could be increased which will enable the models to learn from more cancer types and stages, hence increasing

their predictive power. Since cancer is one such disease where early detection is directly proportional to patient prognosis and survivorship, more accurate and efficient modes of classification meaningfully contribute to early detection and diagnosis. Our work demonstrated that deep learning can transform the medical field-oncology in particular we now hope it will spur further innovation and research needed to ultimately beat breast cancer.

In future we can fine-tune the capabilities further and expand the breast cancer classification systems, for instance, the incorporation of other imaging modalities like Magnetic resonance images (MRI) or computed tomography(CT) scans that would help to view different types of tumours and thus classify them. Also, the training dataset could be increased which will enable the models to learn from more cancer types and stages, hence increasing their predictive power. Since cancer is one such disease where early detection is directly proportional to patient prognosis and survivorship, more accurate and efficient modes of classification meaningfully contribute to early detection and diagnosis. Our work demonstrated that deep learning can transform the medical field-oncology in particular we now hope it will spur further innovation and research needed to ultimately beat breast cancer.

In future we can fine-tune the capabilities further and expand the breast cancer classification systems, for instance, the incorporation of other imaging modalities like Magnetic resonance images (MRI) or computed tomography(CT) scans that would help to view different types of tumours and thus classify them. Also, the training dataset could be increased which will enable the models to learn from more cancer types and stages, hence increasing their predictive power. Since cancer is one such disease where early detection is directly proportional to patient prognosis and survivorship, more accurate and efficient modes of classification meaningfully contribute to early detection and diagnosis. Our work demonstrated that deep learning can transform the medical field-oncology in particular we now hope it will spur further innovation and research needed to ultimately beat breast cancer.

LIMITATIONS AND FUTURE DIRECTIONS

Our study demonstrates the efficacy of the proposed hybrid approach that integrates **EfficientNetV2S**, **EfficientNetV2M**, and **EfficientNetV2L** architectures with a **Vision Transformer** module for the **binary and multi-class classification** of breast cancer histopathological images using the **BreakHis** dataset. The combination of EfficientNetV2's scalable convolutional feature extraction and the transformer's self-attention mechanism enables the model to capture both **fine-grained cellular details** and **global contextual relationships**, resulting in superior performance across all magnification levels.

The proposed models not only achieved **state-of-the-art accuracy** but also exhibited **robust interpretability**, as demonstrated by **Gradient-weighted Class Activation Mapping (Grad-CAM)** visualizations. The Grad-CAM heatmaps revealed that the hybrid architecture effectively focuses on diagnostically relevant regions within histopathological slides, such as nuclei, glandular structures, and cellular boundaries, thereby validating the medical reliability of the learned features. This interpretability strengthens the clinical applicability of our approach by allowing pathologists to visually understand the model's decision-making process.

However, several limitations must be acknowledged. Firstly, the model's performance has been evaluated exclusively on the **BreakHis dataset**, which, despite being a well-established benchmark, represents data from a single source. This reliance may limit the **generalization capability** of the model to other datasets or **real-world clinical environments** that exhibit diverse staining protocols, imaging equipment, and population characteristics. Furthermore, the

limited dataset size and class diversity may hinder the model's ability to adapt to **rare or unseen histopathological patterns**, potentially affecting robustness in clinical deployment.

Another major challenge encountered in this research is the **inherent class imbalance** prevalent in medical imaging datasets. Breast cancer histopathology datasets, including BreakHis, typically contain a **larger proportion of benign samples compared to malignant ones**, leading to skewed learning behavior where the model tends to favor majority classes. While our models demonstrated excellent accuracy in image-wise classification, addressing **class imbalance** remains essential to ensure fairness and reliability. Advanced resampling strategies, **cost-sensitive learning**, and **data synthesis techniques such as GAN-based augmentation** may further improve model balance and resilience against bias.

Despite these limitations, the promising results of this study establish the hybrid EfficientNetV2–Transformer framework as a **highly effective diagnostic aid** for digital pathology. The ability to maintain strong performance across varying magnification levels reinforces the scalability and adaptability of the architecture, making it suitable for diverse histopathological workflows.

Future Work

In future research, several directions can be explored to further enhance the performance, interpretability, and real-world applicability of the proposed framework:

Cross-Dataset Validation:

Extending experiments to include diverse, multi-institutional datasets such as **CAMELYON16** or **TCGA** would help evaluate the model's generalizability and reduce dataset bias.

Data Augmentation and Balancing Techniques:

Implementing advanced augmentation methods like **MixUp**, **CutMix**, or **GAN-based synthetic data generation** can help mitigate class imbalance and improve robustness against rare cases.

Multi-Modal Learning:

Incorporating **clinical metadata** (e.g., patient demographics, hormone receptor status) alongside histopathological images may enhance diagnostic accuracy and support personalized medicine.

Explainable AI (XAI) Extensions:

Expanding interpretability analysis through techniques such as **Layer-wise Relevance Propagation (LRP)** and **attention map visualization** could provide deeper insights into the model's decision-making process and increase

trust among clinicians.

Deployment and Real-Time Inference:

Optimizing the hybrid model for deployment on **edge or cloud-based diagnostic systems** could facilitate real-time analysis in hospital environments, accelerating diagnostic workflows.

By addressing these directions, future work can bridge the gap between experimental research and clinical translation, paving the way for **AI-driven diagnostic systems** that complement pathologists and enhance diagnostic precision, reproducibility, and efficiency in the fight against breast cancer.

REFERENCES

- [1] A. N. Giaquinto, H. Sung, K. D. Miller, J. L. Kramer, L. A. Newman, A. Minihan, A. Jemal, and R. L. Siegel, "Breast cancer statistics, 2022," *CA, A Cancer J. Clinicians*, vol. 72, no. 6, pp. 524–541, Nov. 2022, doi: [10.3322/caac.21754](https://doi.org/10.3322/caac.21754).
- [2] X. Y. Liew, N. Hameed, and J. Clos, "A review of computer-aided expert systems for breast cancer diagnosis," *Cancers*, vol. 13, no. 11, p. 2764, Jun. 2021, doi: [10.3390/cancers13112764](https://doi.org/10.3390/cancers13112764).
- [3] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," 2021, *arXiv:2104.00298*.
- [4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [5] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
- [6] M. A. Aswathy and M. Jagannath, "An SVM approach towards breast cancer classification from H&E-stained histopathology images based on integrated features," *Med. Biol. Eng. Comput.*, vol. 59, no. 9, pp. 1773–1783, Sep. 2021, doi: [10.1007/s11517-021-02403-0](https://doi.org/10.1007/s11517-021-02403-0).
- [7] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," in *Computer Vision-ECCV*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Germany: Springer, 2012, pp. 214–227.
- [8] D. Sanchez-Morillo, J. González, M. García-Rojo, and J. Ortega, "Classification of breast cancer histopathological images using KAZE features," in *Bioinformatics and Biomedical Engineering*, I. Rojas and F. Ortuño, Eds. Cham, Switzerland: Springer, 2018, pp. 276–286.
- [9] A. Chan and J. A. Tuszynski, "Automatic prediction of tumour malignancy in breast cancer with fractal dimension," *Roy. Soc. Open Sci.*, vol. 3, no. 12, Dec. 2016, Art. no. 160558, doi: [10.1098/rsos.160558](https://doi.org/10.1098/rsos.160558).
- [10] E. M. Nejad, L. S. Affendey, R. B. Latip, and I. Bin Ishak, "Classification of histopathology images of breast into benign and malignant using a single-layer convolutional neural network," in *Proc. Int. Conf. Imag., Signal Process. Commun.*, Penang, Malaysia, Jul. 2017, pp. 50–53.
- [11] J. Sun and A. Binder, "Comparison of deep learning architectures for H&E histopathology images," in *Proc. IEEE Conf. Big Data Analytics (ICBDA)*, Nov. 2017, pp. 43–48.
- [12] Y. Benhammou, S. Tabik, B. Achchab, and F. Herrera, "A first study exploring the performance of the state-of-the art CNN model in the problem of breast cancer," in *Proc. Int. Conf. Learn. Optim. Algorithms, Theory Appl.*, Rabat, Morocco, May 2018, pp. 1–6.
- [13] M. A. Nawaz, A. A. Sewissy, and T. H. A. Soliman, "Automated classification of breast cancer histology images using deep learning based convolutional neural networks," *Int. J. Comput. Sci. Netw. Secur.*, vol. 18, no. 4, pp. 152–160, 2018.
- [14] S. Sharma and R. Mehra, "Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images—A comparative insight," *J. Digit. Imag.*, vol. 33, no. 3, pp. 632–654, Jun. 2020, doi: [10.1007/s10278-019-00307-y](https://doi.org/10.1007/s10278-019-00307-y).
- [15] Y. Guo, D. Zhou, R. Nie, X. Ruan, and W. Li, "DeepANF: A deep attentive neural framework with distributed representation for chromatin accessibility prediction," *Neurocomputing*, vol. 379, pp. 305–318, Feb. 2020, doi: [10.1016/j.neucom.2019.10.091](https://doi.org/10.1016/j.neucom.2019.10.091).
- [16] A. M. Ibraheem, K. H. Rahouma, and H. F. A. Hamed, "3PCNNB-net:
- [17] Three parallel CNN branches for breast cancer classification through histopathological images," *J. Med. Biol. Eng.*, vol. 41, no. 4, pp. 494–503, Aug. 2021, doi: [10.1007/s40846-021-00620-4](https://doi.org/10.1007/s40846-021-00620-4).
- [18] H. Gaber, H. Mohamed, and M. Ibrahim, "Breast cancer classification from histopathological images with separable convolutional neural network and parametric rectified linear unit," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.*, A. E. Hassanien, A. Slowik, V. Snásel, H. El-Deeb, and F. M. Tolba, Eds. Cham, Switzerland: Springer, 2020, pp. 370–382.
- [19] K. Shankar, A. K. Dutta, S. Kumar, G. P. Joshi, and I. C. Doo, "Chaotic sparrow search algorithm with deep transfer learning enabled breast cancer classification on histopathological images," *Cancers*, vol. 14, no. 11, p. 2770, Jun. 2022, doi: [10.3390/cancers14112770](https://doi.org/10.3390/cancers14112770).
- [20] Y. Zou, J. Zhang, S. Huang, and B. Liu, "Breast cancer histopathological image classification using attention high-order deep network," *Int. J. Imag. Syst. Technol.*, vol. 32, no. 1, pp. 266–279, Jan. 2022, doi: [10.1002/ima.22628](https://doi.org/10.1002/ima.22628).
- [21] *Breakhis Dataset*. [Online]. Available: <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>

