



Multi-Stage Feature Fusion for Breast Cancer Histopathology using EfficientNetV2 and Vision Transformers

Errissya Rasywir¹, Yovi Pratama², Yessi Hartiwi³, Fachruddin⁴

^{1,2}Informatic Engineering, Faculty of Computer Science, Dinamika Bangsa University, Jambi, Indonesia

^{3,4}Information System, Faculty of Computer Science, Dinamika Bangsa University, Jambi, Indonesia

*errissya.rasywir@gmail.com

Abstract

Early detection of breast cancer is essential for improving patient survival and reducing mortality. Although histopathology-based diagnosis provides definitive clinical evidence, the complex and heterogeneous nature of breast cancer histopathology (BCH) images poses significant challenges for accurate classification. Conventional deep learning models often struggle to capture both local tissue characteristics and global contextual patterns simultaneously. To address this limitation, this study proposes a Multi-Stage Feature Fusion (MSFF) model that integrates EfficientNetV2, a CNN local texture branch, and a Vision Transformer (ViT). EfficientNetV2 extracts high-level semantic features, the CNN branch captures fine-grained local texture information, and the ViT models long-range contextual dependencies through self-attention. These complementary representations are fused through an adaptive multi-stage fusion strategy to enhance discriminative capability for binary BCH classification. An ablation study is conducted to systematically evaluate the contribution of each architectural component. Experimental results on the BreakHis dataset demonstrate that the proposed model achieves the best overall balanced classification performance, with an accuracy of 93.04%, precision of 93.08%, recall of 92.69%, F1-score of 92.89%, and an AUC of 0.9303 at the optimal 200× magnification level. The findings confirm that the proposed fusion strategy provides balanced classification performance, stable optimization behaviour, and strong generalization capability, effectively addressing BCH image heterogeneity and offering strong potential for clinical decision-support applications.

Keywords: BCH; deep learning; EfficientNetV2; feature fusion; vision transformer

How to Cite: E. Rasywir, Y. Pratama, Y. Hartiwi, and F. Fachruddin, "Multi-Stage Feature Fusion for Breast Cancer Histopathology using EfficientNetV2 and Vision Transformers", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1514–1525, Oct. 2026, doi: [10.29207/resti.v10i5.7769](https://doi.org/10.29207/resti.v10i5.7769).

Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.7769>

Received: April 28, 2026

Accepted: September 5, 2026

Available Online: October 17, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.ikatanahli.org/)

1. Introduction

Early detection of breast cancer relies heavily on histopathological diagnosis, which is an important method for accurately analyzing breast tissue [1-3]. However, Breast Cancer Histopathology (BCH) image analysis faces considerable challenges [2], [4]. The complexity of BCH images often complicates this process. This challenge is exacerbated by the heterogeneous nature of cancer cells, such as the irregular size and shape of the cell nucleus, rough chromatin at the edges of the nucleus, and the presence of clearly visible nucleoli [5-7]. An accurate classification remains a major obstacle in image-based diagnosis. Several deep learning architectures, such as convolutional neural networks (CNN) [8], [9], U-Net [10-13], VGG16 [9], [14], [15], and ResNet50 [14] have been proven to provide promising results in medical

image. However, although this method is effective, challenges in BCH image classification still require more sophisticated approaches to handle image complexity [16-18].

Previous research shows that image augmentation techniques can improve model generalization by enriching the training data [19-21]. Therefore, it is important to explore the combination of Convolutional and Vision Transformer equipped with more techniques, to achieve significant improvements in classification accuracy. Against this background, this research aims to develop an effective hybrid model for a more accurate classification of BCH images. EfficientNetV2 is an advanced CNN architecture designed to improve the efficiency and accuracy of image classification. Although EfficientNetV2 has been applied in various medical applications, such as CT and

MRI imaging, its application to breast cancer histopathology images is still relatively new and requires further exploration [22-25]. ViT has the advantage of capturing long-range context and complex spatial patterns in images, which are often missed by traditional CNNs [10], [13], [26-31]. ViT has been successfully used for the classification of natural images, satellite images, and other medical images, including cancer detection [32], [33]. Breast cancer histopathology classification has been widely studied using CNN-based transfer learning, patch-level classification, and transformer-based models. Recent studies reported strong performance using ViT-based comparative models and supervised contrastive learning on BCH datasets [25], [34], [35]. However, most existing methods still rely on a single dominant feature extractor and do not explicitly combine semantic, local-texture, and global-context representations within one unified architecture. Table 1 summarizes the methodological focus and remaining limitations of related studies. EfficientNet-based architectures are attractive for medical image classification because they balance representation capacity and computational efficiency. In BCH analysis, this is important because histopathology images contain magnification-dependent texture variation and complex tissue morphology. However, EfficientNet-based models generally remain CNN-centric and may not sufficiently capture long-range contextual dependencies. Therefore, this study uses EfficientNetV2-B0 as a semantic feature branch rather

than as a standalone classifier, and complements it with CNN-based local texture extraction and ViT-based global context modelling[35], [36]. Vision Transformer models are increasingly used in histopathology because self-attention can capture long-range spatial relationships that convolutional operations may miss. Recent BCH and pathology studies show that ViT-based models can provide strong classification performance and useful contextual representation [34], [35], [37-39]. Nevertheless, transformer-only models may overlook fine-grained local texture patterns that are important for distinguishing benign and malignant tissue. This motivates the use of ViT as a global-context branch within a hybrid fusion framework. Hybrid CNN–Transformer models show that combining local and global information can improve pathology classification. Methods such as KAT and LongMIL support local–global modelling, but they mainly operate at the WSI level. In contrast, this study applies local–global fusion at the BCH image level by integrating EfficientNetV2, CNN local-texture extraction, and ViT global-context modelling with adaptive feature reweighting. Existing BCH studies demonstrate the effectiveness of CNNs, EfficientNet-like backbones, and ViT models, but many still rely on single-stream encoders or limited hybrid fusion. As summarized in Table 1, few studies explicitly combine semantic, local-texture, and global-context features within one adaptive multi-stage fusion framework. Therefore, this study proposes MSFF and validates its branch contributions through component-wise ablation.

Table 1. Comparison of Existing Deep Learning Approaches for Breast Cancer Histopathology Analysis

Study	CNN	ViT	Hybrid Fusion	MSFF	Key Result	Limitation / Research Gap
Review of breast histopathology image analysis [40]	✓	✗	✗	✗	Comprehensive review	Predates recent ViT and hybrid architectures
BCH classification [41]	✓	✗	✗	✗	Demonstrated CNN effectiveness	No global-context modelling
Automatic breast cancer detection [1]	✓	✗	✓	✗	Improved detection performance	Not EfficientNetV2- or ViT-based
Breast cancer diagnosis from histopathology images [4]	✓	✗	✗	✗	Effective diagnostic classification	No branch-level feature fusion
Mitotic detection [42]	✓	✗	✓	✗	Improved mitosis detection	Detection task, not BCH classification
IDC detection / representation learning [43]	✗	✓	✗	✗	Strong representation learning	Transformer-only architecture
BCH image classification	✓	✗	✗	✗	90% image-level accuracy	Single-stream representation learning
Autonomous breast cancer diagnosis [44]	✓	✗	✗	✗	Automated diagnosis capability	No explicit local–global fusion
Patch/slide pathology classification [38]	✗	✓	✗	✗	+8.21 ROC-AUC over ViT	Not BCH-specific
Breast histopathology classification [34]	✓	✓	✗	✗	ViT achieved 94% validation accuracy	No feature fusion mechanism

This study proposes a hybrid deep learning approach that combines EfficientNetV2, Convolutional Neck Section, and Vision Transformer with feature fusion to enhance the classification accuracy of breast cancer histopathology images (benign and malignant). Unlike previous methods that rely solely on either CNNs or Transformers, the proposed approach utilizes three-stage feature extraction, including Bottleneck Feature

Extraction (EfficientNetV2) for capturing global features efficiently, Neck Section Feature Enhancement (CNN) for refining and enhancing local details, and Vision Transformer (ViT) for learning long-range dependencies and spatial relationships. The features from the three branches are first concatenated and then adaptively reweighted in the feature fusion layer, producing a comprehensive representation before

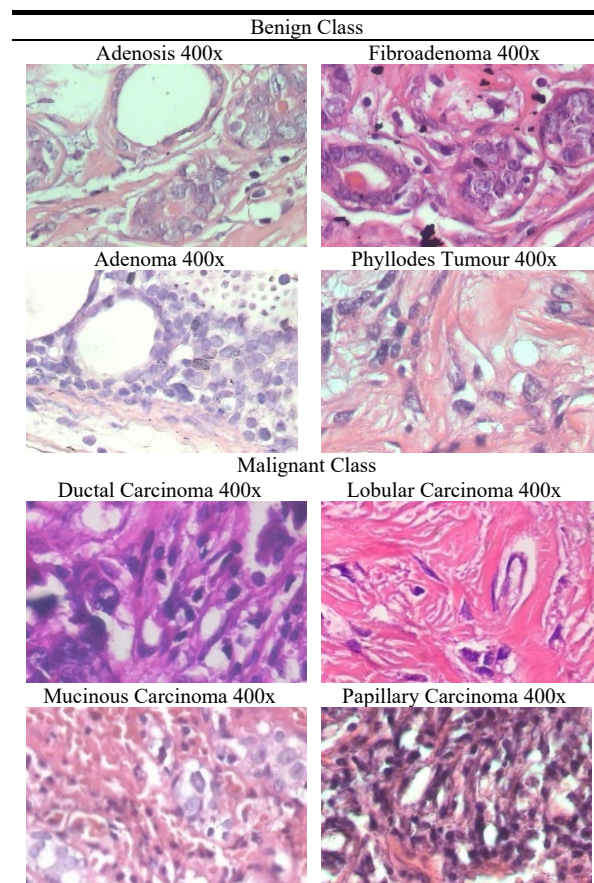
classification using a dense sigmoid layer. Image augmentation is also incorporated to improve generalization and address class imbalance.

2. Methods

2.1 Dataset Preparation and Augmentation

The experiments used the BreakHis dataset, which contains benign and malignant BCH images acquired at 40 $\times$, 100 $\times$, 200 $\times$, and 400 $\times$ magnifications.

Table 2. Display of BCH Image Classes



The dataset includes four benign subtypes (adenosis, fibroadenoma, adenoma, and phyllodes tumour) and four malignant subtypes (ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma), as shown in Table 2. Due to visual similarity across tissue patterns and magnification variation, robust feature extraction is required. Data

augmentation was applied to the training images, increasing the dataset from 7,909 to 31,636 images while preserving class distribution.

Table 3. Class-wise Distribution Before and After Data Augmentation

Class	Original Images	After Augmentation
Benign	2480	9920
Malignant	5429	21716
Total	7909	31636

As shown in Table 3, the original dataset contains 2,480 benign images and 5,429 malignant images. Following augmentation, the number of benign and malignant samples increased proportionally to 9,920 and 21,716 images, respectively. This augmentation strategy increases data diversity while maintaining the inherent class distribution of the original dataset.

Table 4. Magnification-wise Distribution Before and After Data Augmentation

Magnification	Original Images	After Augmentation
40 $\times$	1995	7980
100 $\times$	2081	8324
200 $\times$	2013	8052
400 $\times$	1820	7280
Total	7909	31636

As shown in Table 4, data augmentation increased the BreakHis dataset fourfold, from 7,909 to 31,636 images, while preserving the original distribution across magnification levels (40 $\times$, 100 $\times$, 200 $\times$, and 400 $\times$). The 100 $\times$ subset remained the largest, while 400 $\times$ was the smallest. This approach enhances training data diversity without disproportionately affecting any magnification level, supporting consistent evaluation of the proposed MSFF model.

Table 5. Detailed Data Augmentation Configuration

Augmentation	Parameter
Rotation	$\pm 20^\circ$
Horizontal Flip	0.5
Width Shift	0.1
Height Shift	0.1
Zoom	0.2

As summarized in Table 5, rotation, horizontal flipping, spatial shifting, and zooming were applied during training. These transformations increase intra-class variability and improve the model's ability to generalize to unseen BCH images without altering the underlying pathological characteristics.

Figure 1. Distribution of BCH Images Across the Five Official BreakHis Folds and Corresponding Train-Test Partitions.

To ensure a reliable and unbiased evaluation, the official BreakHis fold partition protocol was adopted in this study. Figure 1 illustrates the five-fold evaluation protocol adopted in this study. The right panel shows the train–test partition for each evaluation round, where one-fold is used for testing and the remaining folds are used for training and validation. This confirms that all five official BreakHis folds were evaluated and the final performance was reported as the average across the five evaluation rounds, reducing selection bias and improving result reliability.

2.2 Overview of the Proposed MSFF Architecture

The proposed Multi-Stage Feature Fusion (MSFF) architecture integrates three parallel feature extraction branches for BCH classification, as shown in Figure 2. The EfficientNetV2-B0 branch extracts high-level semantic representations, the CNN local texture branch captures fine-grained spatial and texture patterns, and the Vision Transformer branch models global contextual dependencies through self-attention.

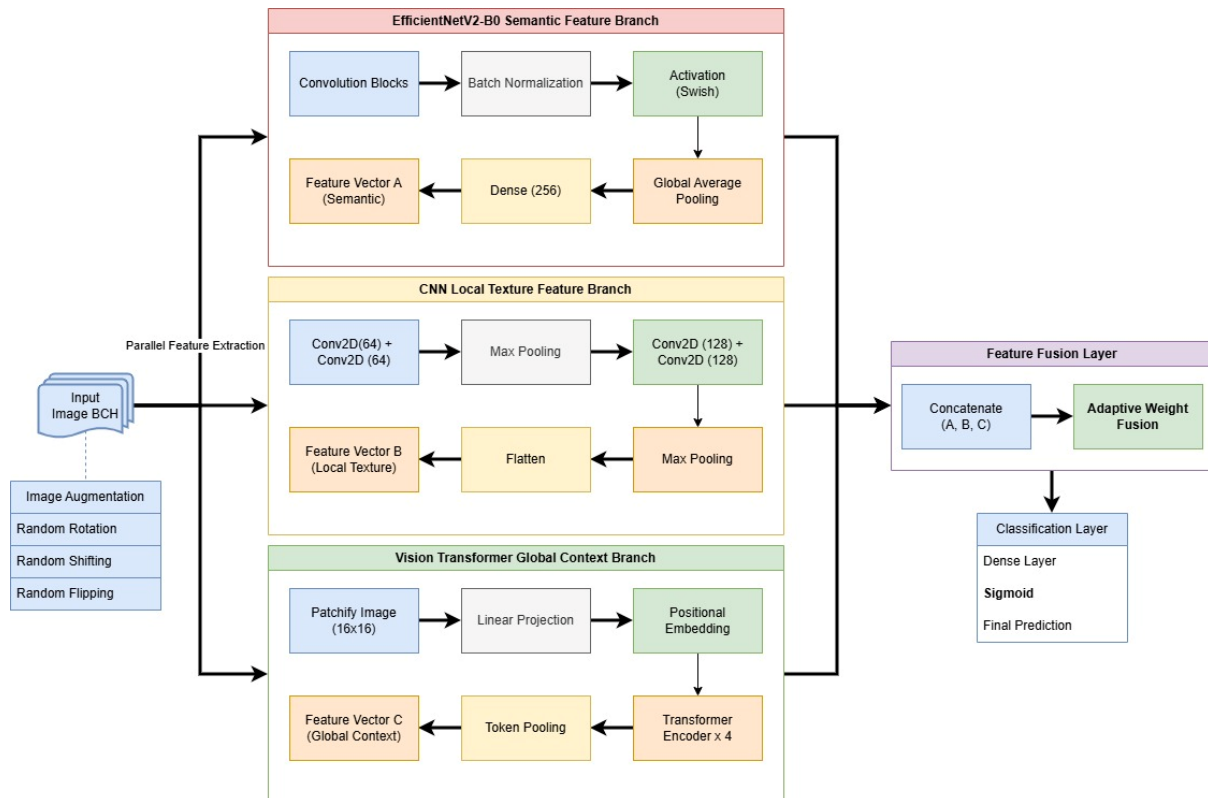


Figure 2. Proposed Multi-Stage Feature Fusion Architecture for Breast Cancer Histopathology Classification Integrating EfficientNetV2, a CNN-Based Neck, and a Vision Transformer.

The feature vectors from the three branches are concatenated and further refined using an adaptive weight fusion module to form a unified representation. This fused representation is then passed to a fully connected classification layer with sigmoid activation for benign–malignant prediction. By explicitly combining semantic, local-texture, and global-context information, the proposed architecture is designed to improve discriminative capability and address the heterogeneous visual patterns of BCH images.

2.3 EfficientNetV2-Based Semantic Feature Extraction

EfficientNetV2-B0 is used as the semantic feature extraction branch due to its efficient compound scaling and strong representational capability. It extracts high-level BCH features related to cellular density, glandular structure, and overall tissue morphology.

These semantic features provide a compact representation of the input image, but they may not fully

capture fine-grained local texture and long-range contextual relationships. Therefore, additional CNN and ViT branches are introduced to complement EfficientNetV2.

2.4 CNN-Based Local Texture Feature Extraction

An independent CNN local texture branch operates in parallel with the EfficientNetV2 branch to capture fine-grained spatial and texture patterns. This branch uses stacked convolutional and pooling layers to refine local tissue details that are important for distinguishing benign and malignant BCH images.

However, local texture information alone may be insufficient for robust class separation, motivating its integration with global context modelling.

2.5 ViT-Based Global Context Modelling

A lightweight ViT branch is incorporated to capture long-range contextual dependencies across image

patches. The input image is divided into fixed-size patches, embedded, and processed through transformer encoder layers with self-attention. This branch complements CNN-based local feature extraction by

modelling broader tissue-level relationships. The detailed layer configuration of all branches is summarized in Table 6.

Table 6. Detailed Layer Configuration and Functional Description of the Proposed Model.

Level	Block / Branch	Layer / Operation	Output Shape	Functional Description
0	Input	BCH Image (128×128×3)	(128,128,3)	RGB histopathology image
1A	EfficientNetV2-B0 Semantic Feature Branch	Data Augmentation → EfficientNetV2-B0 Backbone	(4,4,1280)	Extracts high-level semantic representations
2A		GAP → BatchNorm → Dense(256)	(256,)	Generates semantic feature vector A
1B	CNN Local Texture Feature Branch	Conv64 → Conv64 → MaxPool	(14,14,64)	Captures local texture patterns
2B		Conv128 → Conv128 → MaxPool	(5,5,128)	Refines local texture representations
3B		Flatten	(3200,)	Produces local texture feature vector B
1C	Vision Transformer Global Context Branch	Patchify + Embedding	(64,256)	Converts image patches into token embeddings
2C		Transformer Encoder ×4	(64,256)	Models global contextual dependencies
3C		Token Pooling	(256,)	Produces global context feature vector C
4	Feature Fusion Layer	Concatenate(A,B,C) → Adaptive Weight Fusion	(3712,)	Fuses semantic, local texture, and global contextual representations
5	Classification Layer	Dense(1) + Sigmoid	(1,)	Produces benign/malignant prediction

2.6 Adaptive Multi-Stage Feature Fusion Strategy

To overcome the limitations, this study proposes a Multi-Stage Feature Fusion Model that integrates EfficientNetV2, a Convolutional Neck Section, and Vision Transformer (ViT) to enhance BCH classification accuracy. The input BCH image $X \in \mathbb{R}^{H \times W \times C}$ undergoes basic preprocessing and data augmentation before being fed into the EfficientNetV2-B0 backbone. EfficientNetV2-B0 serves as the primary bottleneck for extracting high-level semantic representations using compound scaling, enabling efficient feature learning with reduced computational cost. The feature extraction process can be formulated as shown in Equation 1.

$$F_s = \phi_{\text{EffNet}}(X) \quad (1)$$

ϕ_{EffNet} ($\cdot$) denotes the EfficientNetV2 feature extractor and F_s represents the resulting semantic feature map. This process consists of convolution, normalization, and nonlinear activation operations as shown in Equation 2.

$$F_s = \sigma(W * X + b) \quad (2)$$

Where W is the convolution kernel, $*$ denotes convolution, b is the bias term, and $\sigma(\cdot)$ represents the activation function (Swish). To enhance local and boundary-aware representations, in parallel with the EfficientNetV2 semantic feature branch, an independent CNN-based local texture branch is employed to capture fine-grained spatial and texture information. This branch applies stacked convolutional and max-pooling operations directly to the input image to preserve local discriminative patterns relevant to BCH classification. This neck applies stacked convolutional and max-pooling layers directly on the feature maps to preserve spatial structure while emphasizing discriminative local patterns. The refined

local feature representation is expressed as shown in Equation 3.

$$F_l = \phi_{\text{CNN}}(X) \quad (3)$$

ϕ_{CNN} ($\cdot$) denotes the convolutional and pooling operations in the neck section. The convolution and pooling operations are defined as shown in Equations 4 and 5.

$$F_l = \sigma(W_l * X + b_l) \quad (4)$$

$$F_l = \text{MaxPool}(F_l) \quad (5)$$

Flattening is applied only after the CNN local texture branch processing to convert spatial feature maps into a vector representation suitable for fusion. In parallel, the input image is processed by a Vision Transformer to capture long-range contextual dependencies. The image is divided into non-overlapping patches, each of which is linearly projected into an embedding space as shown in Equation 6.

$$z_i = X_i E + b \quad (6)$$

X_i denotes the i -the image patch, E is the learnable projection matrix, and z_i represents the patch embedding. Positional embeddings are added before passing the tokens through transformer encoder layers utilizing multi-head self-attention, as shown in Equation 7.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

The final global contextual representation is obtained as shown in Equation 8.

$$F_g = T(z) \quad (8)$$

$T(\cdot)$ denotes the transformer encoding process. The semantic (F_s), local (F_l), and global (F_g) feature vectors are fused to form a unified representation as shown in Equation 9. In the implemented MSFF model, the

semantic, local, and global feature vectors are fused through feature concatenation followed by adaptive weighting. First, the three feature vectors are concatenated to preserve complementary semantic, spatial, and contextual information.

$$F = [F_s \parallel F_l \parallel F_g] \quad (9)$$

The concatenated representation is then refined using an adaptive weighting mechanism to emphasize more informative feature dimensions before classification as shown in Equation 10.

$$F = \alpha F_s \parallel \beta F_l \parallel \gamma F_g \quad (10)$$

The learnable weights α, β, γ are normalized via softmax as shown in Equation 11.

$$[\alpha, \beta, \gamma] = \text{softmax}(W_a * [F_s \parallel F_l \parallel F_g] + b_a) \quad (11)$$

The adaptive weighting mechanism operates at the branch level, where α, β , and γ represent learnable importance coefficients assigned to the semantic, local texture, and global contextual branches, respectively. The fused feature representation F is passed to a fully connected classification layer with Sigmoid activation to predict class probabilities as shown in Equation 12.

$$\hat{y} = \text{Sigmoid}(W_c * F + b_c) \quad (12)$$

W_c denotes the weight matrix of the classification layer, b_c is the bias term, and Sigmoid converts the fused representation into a probability score for binary BCH classification. The trained model is ready to predict unseen data and evaluate classification performance. The proposed architecture integrates complementary semantic, spatial, and contextual features through multi-branch fusion. EfficientNetV2-B0 extracts high-level semantic representations, while the CNN local texture branch refines spatial and boundary-aware features for distinguishing benign and malignant tissues. In parallel, the Vision Transformer captures long-range contextual relationships through self-attention. The branch outputs are concatenated and adaptively reweighted, producing a unified embedding that enhances discriminative capability and learning stability.

2.7 Training and Evaluation Protocol

To ensure a fair comparison among all model variants, a unified training protocol was employed. The experiments used the official five-fold BreakHis partition. In each evaluation round, one-fold was used as the test set, while the remaining folds were used for

training and validation. This procedure was repeated across all five folds, and the final results were reported as the average performance across the five evaluation rounds. Data augmentation was applied only to the training images, while the corresponding test set remained unchanged.

Model performance was evaluated using multiple complementary metrics, including Accuracy, Precision, Recall, F1-score, Sensitivity, Specificity, Matthews Correlation Coefficient (MCC), and ROC-AUC. Accuracy and F1-score were used to assess overall classification performance, while ROC-AUC was employed to evaluate threshold-independent class separability. To provide a more detailed analysis of classification behaviour, confusion-matrix-based metrics, including True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), and False Negative Rate (FNR), were also reported. In addition, qualitative prediction analysis was conducted to examine correctly classified and misclassified BCH samples across different tissue patterns and magnification levels.

2.8 Reproducibility

To ensure reproducibility, all experiments were implemented using Python with TensorFlow/Keras. Input BCH images were resized to 128×128 pixels and processed using a batch size of 28. All models were trained for a maximum of 25 epochs using the Adam optimizer with an initial learning rate of 0.001 and binary cross-entropy loss.

Early stopping with a patience of five epochs and learning-rate reduction on plateau were applied during training. A fixed random seed of 51432 was used for Python, NumPy, and TensorFlow operations. The classification head used dropout rates of 0.4, 0.3, and 0.2. Training and evaluation were conducted using Google Colab equipped with an NVIDIA A100 GPU (40 GB VRAM).

3. Results and Discussions

3.1 Overall Performance of the Proposed Model

This subsection presents the overall classification performance of the proposed MSFF model on the BreakHis dataset. As shown in Table 7, MSFF achieved the best-balanced performance among all evaluated variants, with 93.04% accuracy, 93.08% precision, 92.69% recall, 92.89% F1-score, and 0.9303 AUC.

Table 7. Ablation Study

Model Variant	Acc.	Pre.	Recall	F1	AUC	Key Finding
C: CNN-only	83.93	89.26	87.84	88.54	0.8167	Strong local texture modelling; limited global context.
V: ViT-only	70.72	70.72	100.00	82.85	0.7847	High sensitivity; poor class balance and specificity.
C+V: CNN+ViT	85.40	89.01	90.54	89.77	0.8238	Local-global fusion improves overall performance.
E:EfficientNetV2	82.71	85.00	89.47	87.18	0.8953	Strong semantic representation, baseline performance.
E+C	82.86	80.95	89.47	85.00	0.8204	Local refinement but limited performance gain.
E+V	90.97	75.00	78.95	76.92	0.9587	Highest AUC and strong global separability.
MSFF	93.04	93.08	92.69	92.89	0.9303	Best balanced performance through semantic, local, and global feature fusion.

*E = baseline; E+C = baseline + CNN local texture branch; E+V = baseline + Vision Transformer; MSFF = Multi-Stage Feature Fusion.

Table 7 confirms the complementary roles of the three branches. CNN-only showed strong precision and F1-score but lower AUC, while ViT-only achieved perfect

sensitivity but poor class balance. CNN+ViT improved performance, whereas E+C provided marginal gains and E+V achieved the highest AUC.

Table 8. Diagnostic Metrics Table per Model Variant

Model Variant	Sen.	Spe.	FPR	FNR	Bal. Acc.	MCC	Diagnostic Interpretation
CNN-only	87.84	74.56	25.44	12.16	81.20	0.6221	Good balance between sensitivity and specificity.
ViT-only	100.00	0.00	100.00	0.00	50.00	0.0000	Detects positives but poorly distinguishes negatives.
CNN+ViT	90.54	73.08	26.92	9.46	81.81	0.6463	Improved balance through local–global fusion.
E	89.47	66.67	33.33	10.53	78.07	0.5804	Strong sensitivity but moderate specificity.
E+C	89.47	55.56	44.44	10.53	72.51	0.4857	Increased false positives reduce diagnostic reliability.
E+V	78.95	44.44	55.56	21.05	61.70	0.2418	High separability; weak classification balance.
MSFF	92.69	93.37	6.63	7.31	93.03	0.8607	Best diagnostic reliability, sensitivity–specificity balance.

Overall, MSFF achieved the best accuracy, precision, recall, and F1-score through complementary semantic, local, and global features. Table 8 further supports this finding, with MSFF achieving the highest balanced accuracy (93.03%) and MCC (0.8607), and the lowest FPR (6.63%) and FNR (7.31%). Overall, these results demonstrate that MSFF improves classification performance and diagnostic reliability through complementary multi-branch feature fusion.

3.2 Ablation Study on Multi-Stage Feature Fusion

Figure 3 shows stable learning behaviour of the EfficientNetV2 baseline, with training and validation accuracy converging around 82–83% and smoothly decreasing loss curves, indicating reasonable generalization without severe overfitting. However, its moderate convergence suggests limited discriminative capability beyond local texture features.

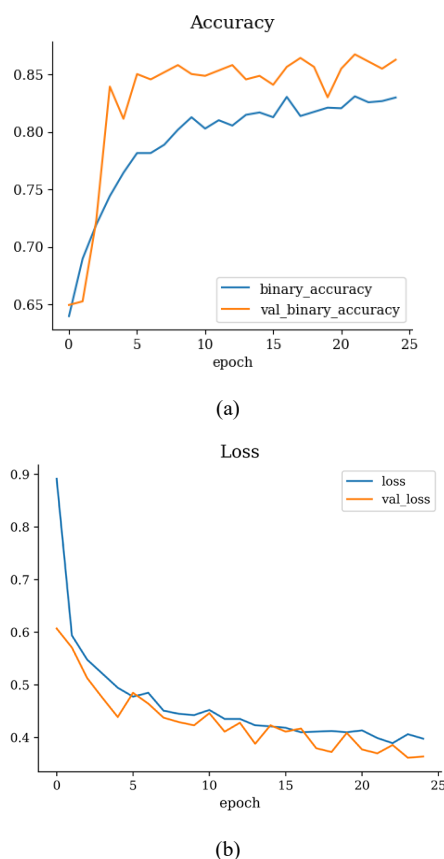


Figure 3. (a) Training and Validation Accuracy, (b) Loss Curves of the EfficientNetV2 Baseline Model.

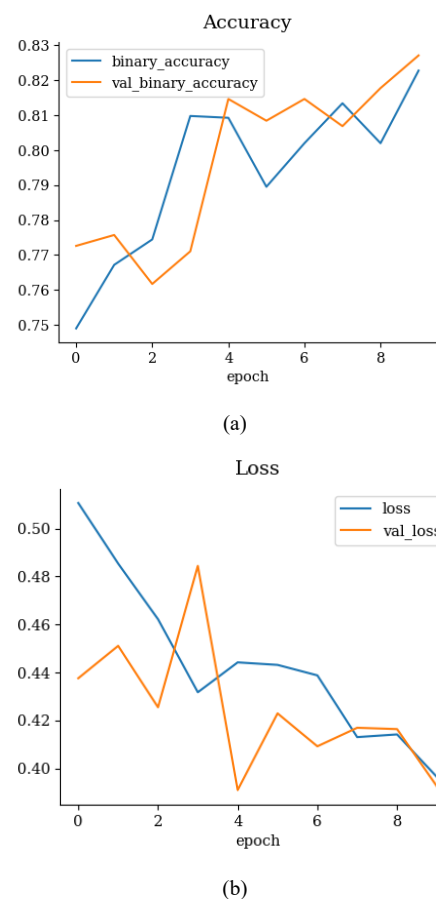


Figure 4. (a) Training and Validation Accuracy, (b) Loss Curves of the EfficientNetV2 with CNN local texture branch.

Figure 4 shows that adding the CNN local texture branch results in a learning pattern similar to that of the baseline model, with training and validation accuracy converging at nearly the same level. The loss curves exhibit mild fluctuations, suggesting slightly higher optimization sensitivity due to the additional convolutional refinement layers. Although this variant preserves high recall and improves local feature aggregation, the limited gain in convergence level and lower AUC indicate that local enhancement alone does not substantially improve overall class separability.

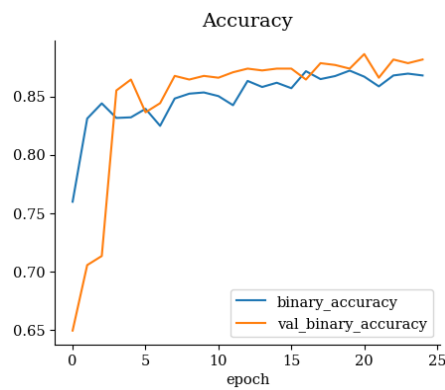
Table 9 further shows that MSFF achieved the highest validation accuracy (90.97%), lowest validation loss (0.2528), and a small train–validation gap (-0.79%), indicating stable convergence with limited overfitting. Although ViT-only showed the lowest validation-loss

variation, its low validation accuracy and poor class performance, convergence stability, and diagnostic balance limited its standalone reliability. Overall, MSFF provided the best combination of validation

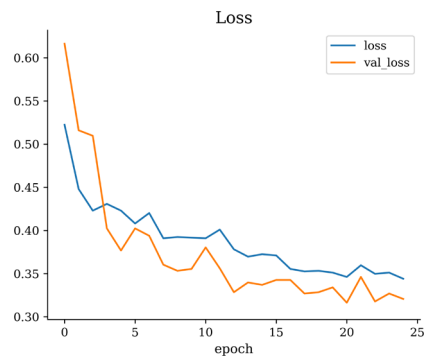
Table 9. Training Stability

Model	Best Acc. (%)	Final Acc. (%)	Best Loss	Best Epoch	Train Acc. (%)	T-V Gap (%)	Loss SD	Interpretation
CNN-only	87.63	87.03	0.3577	19	86.82	-0.21	0.0264	↑ Local, ↓ Global
ViT-only	67.36	67.36	0.5895	1	67.43	0.07	0.0055	↑ Recall, ↓ Balance
CNN+ViT	87.33	85.84	0.3875	10	86.18	0.33	0.0199	↑ Local-global
E	86.76	86.29	0.3586	22	83.00	-3.29	0.0103	↑ Stable
E+C	82.71	82.71	0.3909	10	82.28	-0.43	0.0117	↑ Stable, ↓ Gain
E+V	88.63	88.16	0.3137	21	86.80	-1.36	0.0137	↑ Performance
MSFF	90.97	89.10	0.2528	18	88.31	-0.79	0.0099	↑ Best overall

↑ = strength/improvement; ↓ = limitation; Acc. = validation accuracy unless otherwise specified; T-V Gap = final training accuracy – final validation accuracy; Loss SD = standard deviation of validation loss over the last five epochs



(a)

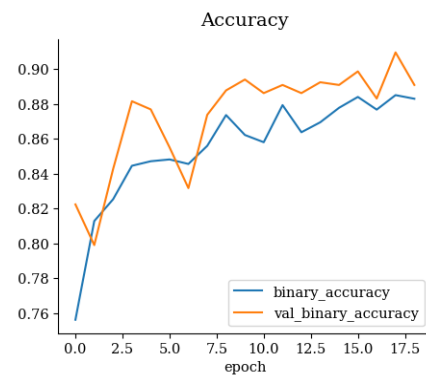


(b)

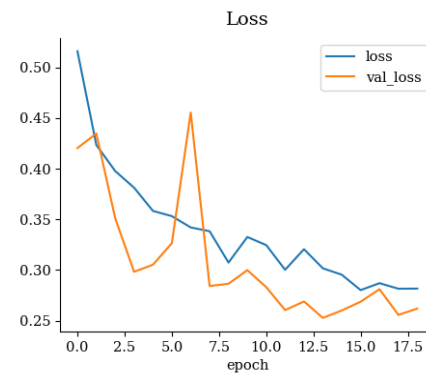
Figure 5. (a) Training and Validation Accuracy, (b) Loss Curves of the EfficientNetV2 with Vision Transformer.

Figure 5 shows stable convergence of the E+V model, with training and validation accuracy increasing steadily and remaining closely aligned, indicating limited overfitting. Both loss curves also decrease consistently with minor fluctuations. Consistent with Table 9, E+V achieved 88.63% best validation accuracy and 0.3137 validation loss, demonstrating good generalization from combined semantic and global contextual features.

Figure 6 shows stable convergence of MSFF, with closely aligned training and validation accuracy and progressively decreasing losses, indicating good generalization and limited overfitting.



(a)



(b)

Figure 6. (a) Training and Validation Accuracy, (b) Loss Curves of the Proposed Multi-Stage Feature Fusion Model.

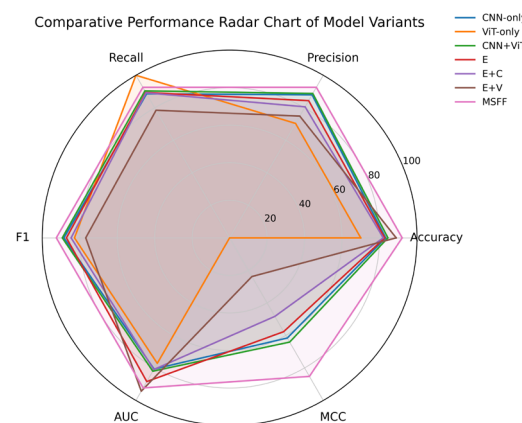


Figure 7. Comparative performance radar chart of model variants based on accuracy, precision, recall, F1-score, AUC, and MCC.

Figure 7 compares all variants across six metrics. MSFF shows the most balanced overall performance, while E+V achieves the highest AUC but lower precision, recall, F1-score, and MCC. MSFF consistently performs well across metrics, confirming the benefit of integrating semantic, local, and global features. In contrast, ViT-only shows poor MCC and balanced accuracy despite high recall, indicating biased malignant predictions.

3.3 Effect of Multi-Stage Feature Fusion Mechanism

Table 10 summarizes the relative improvements of MSFF over the ablation variants, highlighting the contribution of semantic, local-texture, and global-context feature integration.

Table 10. Relative improvement

Comparison	Δ Acc. (%)	Δ Pre. (%)	Δ Sen. (%)	Δ F1 (%)	Δ AUC
MSFF vs C	9.11	3.82	4.85	4.35	0.1136
MSFF vs V	22.32	22.36	-7.31	10.04	0.1456
MSFF vs C+V	7.64	4.07	2.15	3.12	0.1065
MSFF vs E	10.33	8.08	3.22	5.71	0.0350
MSFF vs E+C	10.18	12.13	3.22	7.89	0.1099
MSFF vs E+V	2.07	18.08	13.74	15.97	-0.028

MSFF improves most threshold-based metrics across all variants. The largest accuracy gain occurs over ViT-only, while the largest F1-score gain is observed over E+V. Although E+V achieves a slightly higher AUC, MSFF provides better precision, sensitivity, and F1-score, indicating more balanced classification.

Table 11. Converged Adaptive Fusion Weights

Fold	α Efficient NetV2	β CNN	γ ViT
Fold 1	0.45	0.32	0.23
Fold 2	0.35	0.40	0.25
Fold 3	0.34	0.33	0.33
Fold 4	0.38	0.27	0.35
Fold 5	0.40	0.22	0.38
Mean $\pm$ SD	0.384 ± 0.044	0.308 ± 0.068	0.308 ± 0.065

Table 11 shows that EfficientNetV2 receives the highest mean fusion weight ($\alpha = 0.384 \pm 0.044$), while CNN ($\beta = 0.308 \pm 0.068$) and ViT ($\gamma = 0.308 \pm 0.065$) contribute comparably. This balanced distribution indicates that MSFF benefits from complementary semantic, local, and global features rather than a single dominant branch.

Table 12. Summary of Learning Curve Characteristics of Model

Model	Trend	Loss Trend	Interpretation
C	Stable $\uparrow$ (~ 0.84 – 0.88)	Gradual $\downarrow$	Strong local, $\downarrow$ context
V	Stagnant (~ 0.67)	Limited $\downarrow$	$\downarrow$ Balance
C+V	Stable $\uparrow$ (~ 0.85 – 0.87)	Gradual $\downarrow$	$\uparrow$ Local–global
E	Gradual $\uparrow$ (~ 0.82 – 0.86)	Smooth $\downarrow$	Strong local, $\downarrow$ global
E+C	Baseline-like	Mild fluctuation	$\uparrow$ Sensitivity, $\downarrow$ separation
E+V	Rapid $\uparrow$ (~ 0.90)	Unstable $\rightarrow$	$\uparrow$ Global, sensitive
MSFF	Consistent $\uparrow$ (~ 0.89 – 0.91)	Stable $\downarrow$	$\uparrow$ Balanced, robust

Table 12 summarizes the learning behaviour of all variants. Single and partial-fusion models show limitations in contextual modelling, class balance, or stability, whereas MSFF achieves the highest consistent accuracy with smooth loss reduction, supporting the effectiveness of adaptive multi-stage fusion.

3.4 Effect of Image Magnification on Classification Performance

Tables 13 and 14 summarize MSFF performance across different magnification levels. As shown in Table 13, classification performance improves from $40\times$ to $200\times$ and declines at $400\times$. The highest performance is achieved at $200\times$, with 93.04% accuracy and 92.89% F1-score. To provide a clearer comparison, Table 14 summarizes the ranking of each magnification level based on classification accuracy and F1-score.

Table 13. Performance Across Different Magnification Level

M	Acc.(%)	Pre. (%)	Sen. (%)	F1-Score (%)
$40\times$	82.14	90.00	85.71	87.80
$100\times$	85.71	90.00	90.00	90.00
$200\times$	93.04	93.08	92.69	92.89
$400\times$	71.43	80.95	80.95	80.95

Table 14 confirms that $200\times$ achieves the highest overall rank, providing the best balance between cellular detail and tissue context. Performance declines at $400\times$, likely due to greater texture similarity and reduced global context, consistent with the error analysis in Table 16.

Table 14. Rank Across Different Magnification Levels

M	Rank by Acc	Rank by F1	Mean (%)
$40\times$	3	3	86.41
$100\times$	2	2	88.93
$200\times$	1	1	92.93
$400\times$	4	4	78.57

3.5 Confusion Matrix Analysis

Figure 8 presents the MSFF confusion matrix at $200\times$, the best-performing magnification. The model correctly classified 465 benign and 444 malignant samples, with 33 false positives and 35 false negatives, indicating balanced class recognition.

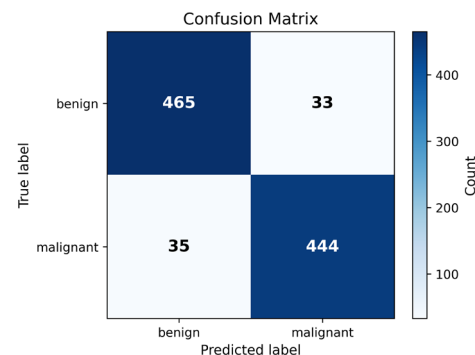


Figure 8. Confusion matrix of the proposed MSFF

The corresponding quantitative metrics are summarized in Table 15, where the proposed model achieved 93.04% accuracy, 92.69% sensitivity, 93.37% specificity, and 93.08% precision.

Table 15. Classification Performance of Proposed MSFF Model

Evaluation Metric	Result (%)
Accuracy	93.040
Recall / True Positive Rate (TPR)	92.693
Specificity / True Negative Rate (TNR)	93.373
Precision	93.082
False Positive Rate (FPR)	6.627
False Negative Rate (FNR)	7.307

3.6 Qualitative Analysis and Error Inspection

This analysis aims to investigate representative misclassified cases and identify possible factors contributing to prediction errors in challenging BCH samples.

Table 16. Representative Misclassified BCH Samples

ID	TL	PL	M	Reason
S-01	Malignant	Benign	400×	Texture similarity
S-02	Benign	Malignant	200×	Staining variation

*ID: sample-ID, TL: True Label, PL: Predicted Label, M: Magnification.

Table 16 summarizes representative errors, linking 400× errors to texture similarity and reduced global

context, and 200× errors to staining variation, supporting the superior balance at 200×.

3.7 Comparative Benchmarking

MSFF is compared with recent BCH methods using common evaluation metrics. Due to differing experimental settings, the results serve as performance references.

Table 17 summarizes the comparison, while Figure 9 compares accuracy. Figure 9 shows that MSFF achieves the highest accuracy (93.04%) among the compared BCH methods, supporting the effectiveness of semantic, local-texture, and global-context fusion.

Figure 10 presents representative predictions, with correct cases shown in green and errors in red. Misclassifications mainly occur in morphologically ambiguous samples, including dense stromal regions, irregular glandular structures, staining variations, and uniform malignant textures. Overall, most samples are correctly classified, indicating effective local and global feature learning.

Table 17. The Benchmarking of BCH Classification Using Deep Learning

Model	Dataset	Accuracy	Precision	Recall	F1-Score	AUC	Augmentation
U-Net [11]	BreakHis	88%	87%	89%	88%	0.93	Rotation, Cropping
VGG16 [40]	BreakHis	90%	89%	91%	90%	0.94	Random Flip
ResNet50[45]	BreakHis	91.00%	90.00%	90.00%	90%	0.92	Random Rotation, Noise
DarkNet[45]	BreakHis	89%	88%	90%	89%	0.91	Random Shift
LeNet[45]	BreakHis	85%	84%	86%	85%	0.88	None
EfficientNet[45]	BreakHis	92%	91%	90%	90.50%	0.95	Random Rotation, Shifting
ViT [43]	BreakHis	91.50%	90%	92%	91%	0.94	Random Flipping, Cropping
Our Model	BreakHis	93.04%	93.08%	92.69%	92.89%	0.9303	Rotation, Shifting, Flipping

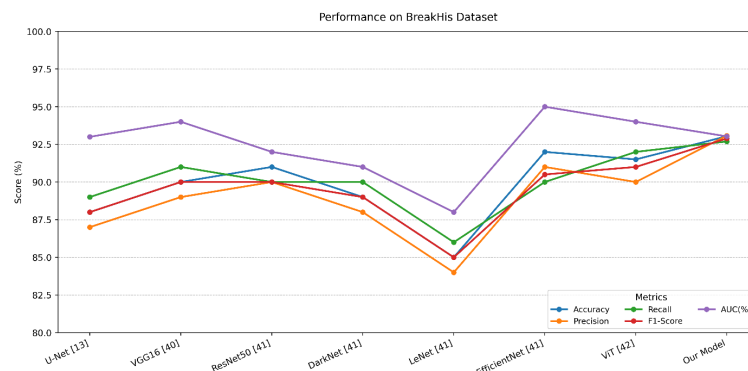


Figure 9. Line Chart of Multi-Stage Feature Fusion Model for BCH Classification

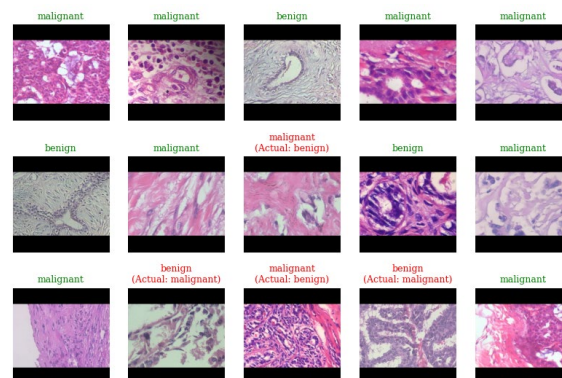


Figure 10. Qualitative Visualization of Classification Results Produced by the Proposed Multi-Stage Feature Fusion Model on Representative BCH Test Samples.

4. Conclusions

The proposed MSFF model effectively integrates semantic, local-texture, and global-context features for BCH classification. On BreakHis at 200× magnification, MSFF achieved 93.04% accuracy, 93.08% precision, 92.69% recall, 92.89% F1-score, and 0.9303 AUC. Ablation results confirm the complementary contribution of the three branches, while augmentation and five-fold evaluation support model robustness and generalization. These findings demonstrate the potential of MSFF for computer-aided breast cancer diagnosis. Future work will explore color normalization, domain adaptation, and advanced transformer-based learning to improve cross-dataset robustness.

Declarations

AI Usage Statement

The authors used ChatGPT (OpenAI) solely for language editing and manuscript refinement. All scientific content, analyses, interpretations, and conclusions remain the authors' responsibility.

Author Contributions

Errissya Rasywir: conceptualization, methodology, formal analysis, validation, supervision; Yovi Pratama: methodology, software, investigation; Yessi Hartiwi: project administration; Fachruddin: investigation, supervision.

Funding Statement

The authors acknowledge the support of the Ministry of Education, Culture, Research, and Technology of Indonesia through the BIMA research program (Contract No. 057/LL10/PG.AK/2024) and the Dinamika Bangsa Foundation.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Clearance

This study used the publicly available BreakHis dataset without involving new human participants or animal subjects and complied with applicable ethical standards for secondary data analysis.

Data and Code Availability

The BreakHis dataset is publicly available. The implementation code and related experimental materials are available from the corresponding author upon reasonable request.

References

- [1] A. Bouziane, S. Boumali, N. Berkane, and F. S. Guendouz, "A Hybrid Approach for Automatic Breast Cancer Detection," in *2020 International Conference on e-Health and Bioengineering (EHB)*, Oct. 2020, pp. 1–4. doi: <https://doi.org/10.1109/EHB50910.2020.9280280>.
- [2] A. J. Shephard et al., "An Automated Pipeline for Tumour-Infiltrating Lymphocyte Scoring in Breast Cancer," in *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, May 2024, pp. 1–5. doi: <https://doi.org/10.1109/ISBI56570.2024.10635302>.
- [3] A. J. Suresh and L. S. Nair, "Mitosis Detection In Breast Histopathology Image Using Ensemble Features Fed CNN Model," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jul. 2023, pp. 1–5. doi: <https://doi.org/10.1109/ICCCNT56998.2023.10307867>.
- [4] A. Labrada and B. D. Barkana, "Breast Cancer Diagnosis from Histopathology Images using Supervised Algorithms," in *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*, Jul. 2022, pp. 102–107. doi: <https://doi.org/10.1109/CBMS55023.2022.00025>.
- [5] Y. Zhang, K. Xia, C. Li, B. Wei, and B. Zhang, "Review of Breast Cancer Pathological Image Processing," *BioMed Research International*, vol. 2021, no. 1, p. 1994764, Jan. 2021, doi: <https://doi.org/10.1155/2021/1994764>.
- [6] Y. Feng, L. Zhang, and J. Mo, "Deep Manifold Preserving Autoencoder for Classifying Breast Cancer Histopathological Images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 1, pp. 91–101, Feb. 2020, doi: <https://doi.org/10.1109/TCBB.2018.2858763>.
- [7] Y. Benmabrouk, M. Gasmi, H. Bendjenna, and A. Nadjah, "Semantic segmentation of breast cancer histopathology images using deep learning," in *2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, Oct. 2022, pp. 1–7. doi: <https://doi.org/10.1109/PAIS56586.2022.9946874>.
- [8] M. J. Hasan et al., "Nuclei Segmentation in ER-IHC Stained Histopathology Images using Mask R-CNN," in *2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Nov. 2022, pp. 1–4. doi: <https://doi.org/10.1109/ISPACS57703.2022.10082832>.
- [9] S. Iqbal and A. N. Qureshi, "A Heteromorphic Deep CNN Framework for Medical Image Segmentation Using Local Binary Pattern," *IEEE Access*, vol. 10, pp. 63466–63480, 2022, doi: <https://doi.org/10.1109/ACCESS.2022.3183331>.
- [10] T. P. T. Armand, S. Bhattacharjee, H. -K. Choi, and H. -C. Kim, "Transformers Effectiveness in Medical Image Segmentation: A Comparative Analysis of UNet-Based Architectures," in *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)*, Feb. 2024, pp. 238–242. doi: <https://doi.org/10.1109/ICAIIIC60209.2024.10463435>.
- [11] M. A. Ortega-Ruiz, C. Karabağ, E. Roman-Rangel, and C. C. Reyes-Aldasoro, "DRD-UNet, a UNet-Like Architecture for Multi-Class Breast Cancer Semantic Segmentation," *IEEE Access*, vol. 12, pp. 40412–40424, 2024, doi: <https://doi.org/10.1109/ACCESS.2024.3377428>.
- [12] L. Zhang, Z. Yang, M. Ju, and H. Liu, "MTC-TransUNet: A Multi-Scale Mixed Convolution TransUNet for Medical Image Segmentation," in *2023 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Oct. 2023, pp. 1–7. doi: <https://doi.org/10.1109/CISP-BMEI60920.2023.10373226>.
- [13] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," Feb. 08, 2021, arXiv: 2102.04306. doi: <https://doi.org/10.48550/arXiv.2102.04306>.
- [14] A. M. Alhassan, "Driving Training-Based Optimization-Multitask Fuzzy C-Means (DTBO-MFCM) Image Segmentation and Robust Deep Learning Algorithm for Multicenter Breast Histopathological Images," *IEEE Access*, vol. 11, pp. 136350–136360, 2023, doi: <https://doi.org/10.1109/ACCESS.2023.3335667>.
- [15] L. Wang, T. Song, T. Katayama, X. Jiang, T. Shimamoto, and J. -S. Leu, "Deep Regional Metastases Segmentation for Patient-Level Lymph Node Status Classification," *IEEE Access*, vol. 9, pp. 129293–129302, 2021, doi: <https://doi.org/10.1109/ACCESS.2021.3113036>.
- [16] Z. Xing, L. Zhu, L. Yu, Z. Xing, and L. Wan, "Hybrid Masked Image Modeling for 3D Medical Image Segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 2115–2125, Apr. 2024, doi: <https://doi.org/10.1109/JBHI.2024.3360239>.

- [17] Y. Ling, Y. Wang, W. Dai, J. Yu, P. Liang, and D. Kong, "MTANet: Multi-Task Attention Network for Automatic Medical Image Segmentation and Classification," *IEEE Transactions on Medical Imaging*, vol. 43, no. 2, pp. 674–685, Feb. 2024, doi: <https://doi.org/10.1109/TMI.2023.3317088>.
- [18] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved Automatic Detection and Segmentation of Cell Nuclei in Histopathology Images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, Apr. 2010, doi: <https://doi.org/10.1109/TBME.2009.2035102>.
- [19] B. D. Basaran, W. Zhang, M. Qiao, B. Kainz, P. M. Matthews, and W. Bai, "LesionMix: A Lesion-Level Data Augmentation Method for Medical Image Segmentation," Aug. 17, 2023, *arXiv*: 2308.09026, doi: <https://doi.org/10.48550/arXiv.2308.09026>.
- [20] D. DECAHNG, "Semi-Supervised Medical Image Segmentation with Multimodal Data Augmentation and Self-Supervised Learning," in *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Dec. 2024, pp. 1–4, doi: <https://doi.org/10.1109/ICCWAMTIP64812.2024.10873760>.
- [21] X. Liu, K. Ono, and R. Bise, "Mixing Data Augmentation with Preserving Foreground Regions in Medical Image Segmentation," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2023, pp. 1–5, doi: <https://doi.org/10.1109/ISBI53787.2023.10230495>.
- [22] C. Raghuram and M. Thenmozhi, "Short Review on Contrastive Learning-based Segmentation Techniques for Medical Image Processing," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, Jun. 2023, pp. 290–296, doi: <https://doi.org/10.1109/APSIT58554.2023.10201707>.
- [23] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "Methods for Nuclei Detection, Segmentation, and Classification in Digital Histopathology: A Review—Current Status and Future Potential," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 97–114, 2014, doi: <https://doi.org/10.1109/RBME.2013.2295804>.
- [24] H. Fatakdwala et al., "Expectation–Maximization-Driven Geodesic Active Contour With Overlap Resolution (EMaGACOR): Application to Lymphocyte Segmentation on Breast Cancer Histopathology," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 7, pp. 1676–1689, Jul. 2010, doi: <https://doi.org/10.1109/TBME.2010.2041232>.
- [25] J. Albarracin, F. Cano, E. Romero, and A. Cruz-Roa, "A comparative analysis between two convolutional networks architectures for semantic segmentation of histopathology breast cancer images," in *2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM)*, Nov. 2023, pp. 1–5, doi: <https://doi.org/10.1109/SIPAIM56729.2023.10373477>.
- [26] S. Zheng et al., "Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers," Jul. 25, 2021, *arXiv*: 2012.15840, doi: <https://doi.org/10.48550/arXiv.2012.15840>.
- [27] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," Oct. 28, 2021, *arXiv*: 2105.15203, doi: <https://doi.org/10.48550/arXiv.2105.15203>.
- [28] H. Thisanake, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic Segmentation using Vision Transformers: A survey," May 05, 2023, *arXiv*: arXiv:2305.03273, doi: <https://doi.org/10.48550/arXiv.2305.03273>.
- [29] Z. Liang, K. Zhao, G. Liang, S. Li, Y. Wu, and Y. Zhou, "MAXFormer: Enhanced transformer for medical image segmentation with multi-attention and multi-scale features fusion," *Knowledge-Based Systems*, vol. 280, p. 110987, Nov. 2023, doi: <https://doi.org/10.1016/j.knosys.2023.110987>.
- [30] J. Yang, L. Jiao, R. Shang, X. Liu, R. Li, and L. Xu, "EPT-Net: Edge Perception Transformer for 3D Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 11, pp. 3229–3243, Nov. 2023, doi: <https://doi.org/10.1109/TMI.2023.3278461>.
- [31] Z. Ji, Z. Chen, and X. Ma, "Grouped multi-scale vision transformer for medical image segmentation," *Sci Rep*, vol. 15, no. 1, p. 11122, Apr. 2025, doi: <https://doi.org/10.1038/s41598-025-95361-8>.
- [32] E. Rasywir et al., "Robust SegFormer-Based Cervical Neoplastic Lesion Segmentation Using Lesion-Specific Augmentation and Anatomical Label Optimization," *IEEE Access*, vol. 13, pp. 205915–205931, 2025, doi: <https://doi.org/10.1109/ACCESS.2025.3639572>.
- [33] E. Rasywir, S. Nurmaini, M. Fachrurrozi, S. Assegaff, M. N. Rachmatullah, and Y. Pratama, "CerviNetFormer: Leveraging Convolutional and Vision Transformer Ensembles for Enhancing Accuracy in Cervical Pre-cancer Lesion Segmentation," *IAENG IJCS*, vol. 52, no. 12, pp. 4842–4853, 2025. [Online]. Available: https://www.iaeng.org/IJCS/issues_v52/issue_12/IJCS_52_12_28.pdf.
- [34] Z. Eskandari, A. Eslamian, and N. Munia, "Evaluating Deep Learning Models for Breast Cancer Classification: A Comparative Study," presented at the SPIE Medical Imaging 2025, 2025, doi: <https://doi.org/10.1117/12.3047441>.
- [35] M. M. Sani, "Classification of Breast Cancer Histopathology Images using a Modified Supervised Contrastive Learning Method," 2024, doi: <https://doi.org/10.48550/arXiv.2405.03642>.
- [36] M. Behzadpour, B. L. Ortiz, E. Azizi, and K. Wu, "Breast Tumor Classification Using EfficientNet Deep Learning Model," *arXiv preprint*, 2024, doi: <https://doi.org/10.48550/arXiv.2411.17870>.
- [37] M. Shiri, "Supervised Contrastive Vision Transformer for Breast Histopathological Image Classification," 2024, doi: <https://doi.org/10.48550/arXiv.2404.11052>.
- [38] R. Rawat and F. Ahmed, "Efficient Breast and Ovarian Cancer Classification via ViT-Based Preprocessing and Transfer Learning," *arXiv preprint*, 2025, doi: <https://doi.org/10.48550/arXiv.2509.18553>.
- [39] P. Afshin, D. Helminiak, and T. Lu, "Breast Cancer Classification in Deep Ultraviolet Fluorescence Images Using a Patch-Level Vision Transformer Framework," presented at the IEEE EMBC 2025, 2025, doi: <https://doi.org/10.1109/EMBC58623.2025.11253275>.
- [40] X. Zhou et al., "A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks," *IEEE Access*, vol. 8, pp. 90931–90956, 2020, doi: <https://doi.org/10.1109/ACCESS.2020.2993788>.
- [41] E. Egriboz, B. Gokcen, and G. Bilgin, "Classification of Breast Cancer Histopathology Images by Cell-Centered Deep Learning Approach," in *2020 28th Signal Processing and Communications Applications Conference (SIU)*, Oct. 2020, pp. 1–4, doi: <https://doi.org/10.1109/SIU49456.2020.9302066>.
- [42] M. V. Deepak Naik, S. Adarsh, S. Vijayakumar, A. P. Nambiar, and L. S. Nair, "Hybrid Feature Set based Mitotic Detection in Breast Histopathology Images," in *2022 International Conference on Inventive Computation Technologies (ICICT)*, Jul. 2022, pp. 281–290, doi: <https://doi.org/10.1109/ICICT54344.2022.9850552>.
- [43] Kiran, P. B D, D. R H, K. N. Abdul Sattar, and S. Kumar D S, "Breast Histopathology Image Embedding's in Invasive Ductal Carcinoma Detection using Self-Supervised Vision Transformers Techniques," in *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, Oct. 2023, pp. 1–5, doi: <https://doi.org/10.1109/EASCT59475.2023.10393320>.
- [44] M. Son, W. Jun, and S. Lee, "Autonomous Diagnosis System of Breast Cancer," in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2024, pp. 1–6, doi: <https://doi.org/10.1109/ICCE59016.2024.10444471>.
- [45] K. S. Gandle and D. B. Kshirsagar, "Breast Cancer Categories, Analysis, Detection: Systematic Review for Histopathological Images," in *2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)*, Feb. 2024, pp. 1–6, doi: <https://doi.org/10.1109/ICIPTM59628.2024.10563307>.