WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 8, Number 3, Year 2026
DOI: https://doi.org/10.62051/ijcsit.v8n3.07
Journal homepage: https://wepub.org/index.php/IJCSIT/index

HGCTransformer: Hybrid Gated CNN-Transformer for Breast Cancer Image Classification

Shengpei Ye *

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan 454000, China

*Corresponding Author: Shengpei Ye

ABSTRACT

Breast cancer is the most common malignant tumor among women, and histopathological images play a crucial role in the differential diagnosis of benign and malignant lesions. Although Convolutional Neural Networks (CNNs) and Transformers have been widely used in medical image classification, CNNs often struggle to capture global structural patterns, while Transformers may underperform in modeling fine-grained local features. To address this, we propose a Hybrid Gated CNN-Transformer (HGCTransformer) model for breast tumor histopathological image classification. The model introduces a Dual-Branch Convolution and Attention Residual Module (DBCARM) into the Transformer block to integrate local texture and global contextual information. Additionally, a Gated Multi-Scale Feed-forward Network (GMSFN) is designed to enhance the discrimination of multi-scale malignant features, such as nuclear atypia and architectural disarray. Experimental results on public breast histopathology dataset indicate that the proposed method achieves a classification accuracy of 99.76%, underscoring its potential for computer-aided diagnosis of breast cancer.

KEYWORDS

Breast cancer classification; Histopathological image; Convolutional neural network; Transformer

1. INTRODUCTION

Breast cancer is the most commonly diagnosed cancer and a leading cause of cancer-related death among women worldwide. According to the GLOBOCAN 2022 estimates by the International Agency for Research on Cancer (IARC) and the World Health Organization (WHO), there were 2.31 million new cases (11.6% of all cancers) and 665,684 deaths (6.9% of cancer mortality) in 2022, with global incidence projected to rise by 77% by 2050 [1]. Early symptoms—such as breast lumps, skin changes, or nipple discharge—are often subtle, leading to delayed diagnosis. As the disease progresses, tumors may invade local tissues and metastasize to regional lymph nodes or distant organs [2]. Although non-invasive imaging modalities like mammography are widely used, their capacity to capture tumor heterogeneity remains limited [3], resulting in non-negligible false-negative and false-positive rates. Here, 'negative' refers to benign tumors, while 'positive' indicates malignant cancer. Therefore, histopathological examination of hematoxylin and eosin (H&E)-stained tissue sections, obtained via biopsy or surgical resection, remains the diagnostic gold standard [4]. Given the high resolution and complexity of whole-slide images (WSIs), automated analysis tools have become increasingly important to support consistent and efficient diagnosis.

Diagnostic errors in breast cancer can have serious clinical consequences. False-negative diagnoses may delay appropriate treatment, allowing early-stage cancers to progress to advanced stages and

reducing survival outcomes. Conversely, false-positive results may lead to unnecessary interventions—including surgery, chemotherapy, or radiation therapy—imposing significant physical, psychological, and economic burdens on patients [5]. Moreover, inter-observer variability among physicians contributes to diagnostic inconsistency, particularly in differentiating borderline lesions such as atypical ductal hyperplasia from ductal carcinoma in situ [6]. These challenges underscore the need for robust, reproducible, and objective computational tools to assist physicians in achieving accurate and reliable diagnoses.

Traditional machine learning methods relying on handcrafted features often exhibit limited representational capacity and poor generalization in histopathological image analysis [7]. Deep learning has emerged as a powerful alternative, with CNNs excelling at capturing local morphological patterns, while Transformers model long-range dependencies and global context through self-attention. However, CNNs are constrained by fixed receptive fields, limiting their ability to perceive global tissue architecture, whereas Transformers lack the inductive biases of CNNs—such as local connectivity and translation equivariance—and require large amounts of annotated data to prevent overfitting, making them less suitable for data-scarce medical applications [8].

To address these limitations, an ideal model should simultaneously preserve the strong local inductive biases of CNNs and harness the global contextual modeling capability of Transformers, while being data-efficient and robust in medical scenarios. The main contributions of this paper are as follows:

(1) We propose a Hybrid Gated CNN-Transformer (HGCTransformer) that effectively integrates fine-grained nuclear morphological details with global glandular and tissue-level contextual information. By synergizing CNNs and Transformers, our model enables a more comprehensive and discriminative feature representation, thereby addressing the limitations of existing methods in modeling multi-scale pathological structures.

(2) The architecture is built upon Hybrid Convolutional and Attention (HybridCA) Blocks, each employing a dual pre-layer normalization strategy. Channel-wise normalization is applied at each spatial location to standardize feature scales across channels while preserving spatial structure. When combined with residual connections throughout the network, this design effectively retains low-level features and accelerates model convergence.

(3) We introduce a Dual-Branch Convolution and Attention Residual Module (DBCARM), which employs a dual-branch architecture: the global branch uses adaptive average pooling and 1×1 convolutions to capture glandular and tissue-level structures, while the local branch uses depth-wise separable convolutions to extract fine-grained nuclear details. Features from both branches are fused via a residual connection, which enhances the model's ability to differentiate between benign and malignant tissues.

(4) The Gated Multi-Scale Feed-forward Network (GMSFN) leverages three parallel depth-wise convolutional branches with dilation rates of 1, 2, and 3 to extract multi-scale morphological features, capturing both cellular-level details and broader tissue patterns. These features are integrated through element-wise multiplication followed by a GELU activation, thereby improving robustness in breast cancer image classification.

The remainder of this paper is structured as follows: Section 2 reviews related work on methodologies and loss functions for breast cancer histopathological image classification. Section 3 presents the proposed HGCTransformer framework, detailing its architectural innovations and the evaluation metrics used. Section 4 reports experimental results, including datasets, training protocols, data augmentation strategies, ablation studies, cross-dataset generalization, and comparative analysis. Finally, Section 5 concludes the paper with a summary and outlook on future work.

2. RELATED WORK

2.1. Deep Learning for Breast Cancer Classification

Breast cancer histopathological image classification plays a critical role in computer-aided diagnosis, with classification—distinguishing benign from malignant tissue regions—being the most clinically relevant task. Early approaches relied on handcrafted features such as color histograms, texture descriptors (e.g., Gray-Level Co-occurrence Matrix, GLCM [9]), Gabor filters and wavelet-based representations [10], combined with classical classifiers like SVMs or decision trees. While feasible on small datasets, these methods suffer from poor generalization across staining variations and complex tissue morphologies.

With the rise of deep learning, Convolutional Neural Networks (CNNs) have become the dominant framework for automatic feature learning in histopathology. They excel at capturing local morphological cues such as nuclear atypia, glandular architecture, and stromal patterns. For instance, RANet [11] enhances sensitivity to rare malignant samples through an anomaly detection module, while msSE-ResNet [12] achieves strong performance via multi-scale attention mechanisms. However, the lack of standardized evaluation and frequent neglect of class imbalance raise concerns about the reliability and generalizability of existing methods.

However, CNNs are limited by fixed receptive fields, hindering their ability to model long-range spatial dependencies and global tissue organization—key factors in differentiating invasive carcinoma from benign mimics. To address this challenge, Transformer-based models have gained popularity. The Vision Transformer(ViT) [13] enables global context aggregation through self-attention on image patches. Variants like CvT [14] and MaxViT [15] integrate convolutional priors to preserve locality and enhance multi-granularity perception, while lightweight designs such as SepViT [16] reduce computation via decoupled attention.

More recently, hybrid CNN-Transformer architectures have emerged to combine the strengths of both paradigms. Models like MedViT [17] and HCANet [18] integrate local convolution blocks with global attention modules to enhance feature interaction across scales. Although these hybrids show promise in low-data regimes, they are often evaluated under inconsistent protocols, and many overlook the prevalent class imbalance between benign and malignant samples.

However, few studies report consistent evaluation protocols for the binary subtask, and many overlook the impact of class imbalance between benign and malignant samples. These gaps motivate the need for more robust, interpretable, and clinically aligned frameworks.

2.2. Loss Function

In breast cancer image classification, training datasets often exhibit class imbalance between benign and malignant samples. Using standard cross-entropy loss directly may bias the model toward the majority class. To formalize this issue, we first define the standard binary cross-entropy(BCE) loss [19] as Equation(1):

$$L_{BCE} = -[y \log(p) + (1-y) \log(1-p)] \tag{1}$$

Where $y \in \{0,1\}$ is the ground-truth label and $p \in (0,1)$ is the predicted probability for the positive class (malignant). This formulation treats all classes equally, which can lead to suboptimal performance when class distributions are highly imbalanced.

To address this limitation, we adopt a weighted binary cross-entropy loss, in which class-specific weights are inversely proportional to the class frequencies in the training set. This strategy effectively amplifies the gradient contribution of minority class samples, thereby improving class-balanced learning.

Specifically, the class weights are calculated as follows. Let N denote the total number of training samples, $C = 2$ the number of classes (benign and malignant), and n_i the number of samples in class $i \in \{0,1\}$. The weight for class i is defined as Equation (2):

$$w_i = \frac{N}{C \cdot n_i} \tag{2}$$

This strategy assigns higher weights to underrepresented classes, increasing their influence during gradient updates [20]. The final loss function is formulated as Equation (3):

$$L_{wce} = -\sum_{i=0}^{1} w_i \cdot (y_i \log p_i + (1 - y_i) \log(1 - p_i)) \tag{3}$$

Where $y_i \in \{0,1\}$ is the ground-truth label of class i and p_i represents the predicted probability output by the model for class i. By incorporating class-dependent weights into the loss function, our method mitigates the bias introduced by data imbalance and enhances model generalization on minority-class samples.

3. METHODOLOGY

3.1. Overview of HGCTransformer

Figure 1 illustrates the hierarchical architecture of HGCTransformer. The input image $x \in \mathbb{R}^{B\times3\times224\times224}$ is first processed by a custom-designed backbone ensemble with variable depth and adaptive residual connectivity, referred to as the Adaptive Backbone Ensemble. The ensemble is defined as M_{θ}^{r}, where θ denotes network depth and r indicates the presence(1) or absence(0) of residual connectivity. This ensemble is derived from a shared convolutional template consisting of 3×3 convolutions, each followed by BatchNorm and a ReLU activation, and downsampling layers implemented through strided convolutions. It enables independent variation of network depth θ and residual connectivity r across configurations, while maintaining structural and semantic consistency.

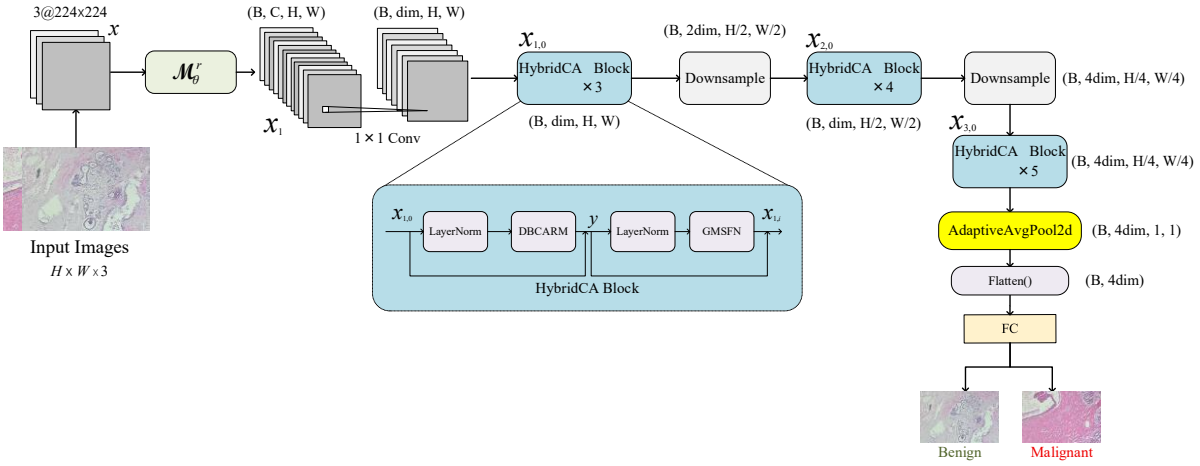


Figure 1. HGCTransformer Architecture

As outlined in Table 1, the HGCtransformer algorithm systematically transforms an input image $x \in \mathbb{R}^{B\times3\times224\times224}$ into classification logits $\in \mathbb{R}^{B\times C}$, where B represents the batch size, 3 denotes RGB channels, 224×224 represents the image resolution, and $C = 2$ denotes the number of classes for

classification. The logits are real-valued scores that are subsequently passed through softmax function to obtain class probabilities.

Table 1. HGCTransformer algorithm.

Algorithm 1: HGCTransformer Structure
Require: Input image $x \in \mathbb{R}^{B\times 3\times 224\times 224}$
Ensure: Classification result $logits \in \mathbb{R}^{B\times C}$, where C means the number of Class.
Stage 1: M_θ^r Ensemble
$x_1 \leftarrow M_\theta^r\ (x)$
$x_{1,0} \leftarrow Conv_{1\times 1}(x_1)$
Stage 2: HybridCA Block × N_1
for i = 1 to N_1 do
$y \leftarrow DBCARM(LayerNorm(x_{1,i-1})) + x_{1,i-1}$
$x_{1,i} \leftarrow GMSFN(LayerNorm(y)) + y$
end for
Stage 3: $x_{2,0} \leftarrow Downsample(x_{1,N_1})$
Stage 4: HybridCA Block × N_2
for i = 1 to N_2 do
$y \leftarrow DBCARM(LayerNorm(x_{2,i-1})) + x_{2,i-1}$
$x_{2,i} \leftarrow GMSFN(LayerNorm(y)) + y$
end for
Stage 5: $x_{3,0} \leftarrow Downsample(x_{2,N_2})$
Stage 6: HybridCA Block × N_3
for i = 1 to N_3 do
$y \leftarrow DBCARM(LayerNorm(x_{3,i-1})) + x_{3,i-1}$
$x_{3,i} \leftarrow GMSFN(LayerNorm(y)) + y$
end for
Stage 7: Classification Head
$logits_1 \leftarrow AdaptiveAvgPool2d(x_{3,N_3}),\ logits_1 \in \mathbb{R}^{B\times 4dim\times 1\times 1}$
$logits_2 \leftarrow Flatten(logits_1),\ logits_2 \in \mathbb{R}^{B\times 4dim}$
$logits \leftarrow Linear(logits_2),\ logits \in \mathbb{R}^{B\times C}$
return $logits \in \mathbb{R}^{B\times C}$

(DBCARM: Dual-Branch Convolution and Attention Residual Module, GMSFN: Gated Multi-Scale Feed-forward Network)

After processing by the Adaptive Backbone Ensemble, the input image is transformed into a feature map tensor of shape $x_1 \in \mathbb{R}^{B\times C\times H\times W}$, where B denotes the batch size, C the number of channels, and $H \times W$ the spatial resolution. This intermediate representation captures rich local and semantic information and serves as the input to the subsequent hierarchical encoder stages.

Subsequently, a 1×1 convolution to project the channel dimension into the model's unified embedding space dim. This operation provides flexibility in channel adjustment and enhances feature compatibility, allowing outputs from different backbone configurations to seamlessly interface with the subsequent hierarchical encoder stages.

The network then proceeds through three encoder stages, containing 3, 4, and 5 HybridCA Blocks respectively, enabling progressive feature refinement from local details to global semantics. Within each stage, downsampling operations gradually reduce spatial resolution while increasing channel depth, forming a hierarchical representation hierarchy. This design facilitates effective aggregation

of multi-granularity information, improving the model's discriminative power for complex pathological patterns.

Finally, the spatial dimension of the output is reduced to 1×1 via the adaptive average pooling, and then flattened and linear layers are carried out to generate the classification $logits \in \mathbb{R}^{B\times C}$.

3.2. HybridCA Block: Hybrid Convolutional and Attention Block

The HybridCA block serves as the cornerstone of the HGCTransformer model, effectively integrating the DBCARM and the GMSFN within a dual pre-layer normalization framework. Each normalization is applied independently at every spatial location, computing the mean and variance over the channel dimension, ensuring consistent feature scaling while preserving spatial structure.

Given an input tensor $x \in \mathbb{R}^{B\times C\times H\times W}$, it is first reshaped into a 3D format $x_{3d} \in \mathbb{R}^{B\times (HW)\times C}$ via the reshape function, where each spatial position (h, w) is treated as an independent C-dimensional vector. Layer normalization is then applied along the channel dimension for each spatial location independently. This design preserves the intrinsic spatial structure of visual features while facilitating robust feature learning.

The LayerNorm operation is defined as Equation (4):

$$\text{LayerNorm}(x) = \gamma\left(\frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}}\right) + \beta \tag{4}$$

Where x denotes the input feature map, μ and σ are the mean and variance computed over the channel dimension, and computed as Equation (5) and Equation (6):

$$\mu = \frac{1}{C}\sum_{c=1}^{C} x_{3d} \tag{5}$$

$$\sigma = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(x-\mu)^2 + \varepsilon} \tag{6}$$

$\varepsilon = 1\times 10^{-5}$ ensures numerical stability, and γ, β are learnable affine parameters that allow the model to scale and shift the normalized values. The normalized features are then reshaped back to the original 4D structure $x_{4d} \in \mathbb{R}^{B\times C\times H\times W}$ via the reshape function.

The processing flow of each HybridCA Block in the k-th stage (k = 1, 2, 3) is as follows: the input feature $x_{k,i-1}$ undergoes LayerNorm and DBCARM, then fuses with the original input via a residual connection to obtain y, as defined in Equation (7):

$$y = \text{DBCARM}(\text{LayerNorm}(x_{k,i-1})) + x_{k,i-1} \tag{7}$$

Subsequently, y is processed by LayerNorm and GMSFN to produce the output feature $x_{k,i}$ (where i denotes the i-th HybridCA Block in the stage, starting from i = 1), which is again fused via a residual connection to generate the final output. The operations are formulated as Equation (8):

$$x_{k,i} = \text{GMSFN}(\text{LayerNorm}(y)) + y \tag{8}$$

Furthermore, the integration of residual connections throughout the architecture allows for seamless gradient propagation and the preservation of fine-grained, low-level features. The overall architecture of the HybridCA block, including the placement of LayerNorm, residual connections, DBCARM, and GMSFN, is illustrated in Figure 1.

3.3. DBCARM: Dual-Branch Convolution and Attention Residual Module

CNNs are highly effective at capturing local textures and morphological patterns, while Transformers leverage self-attention mechanisms to model long-range dependencies and global contextual relationships [21]. However, simply combining these two paradigms in a parallel or sequential manner often leads to suboptimal feature fusion and excessive computational overhead, particularly when applied to high-resolution histopathological images.

To address this challenge, we propose a dual-path architecture designed to achieve synergistic integration of local detail extraction and global context modeling, as illustrated in Figure 2.

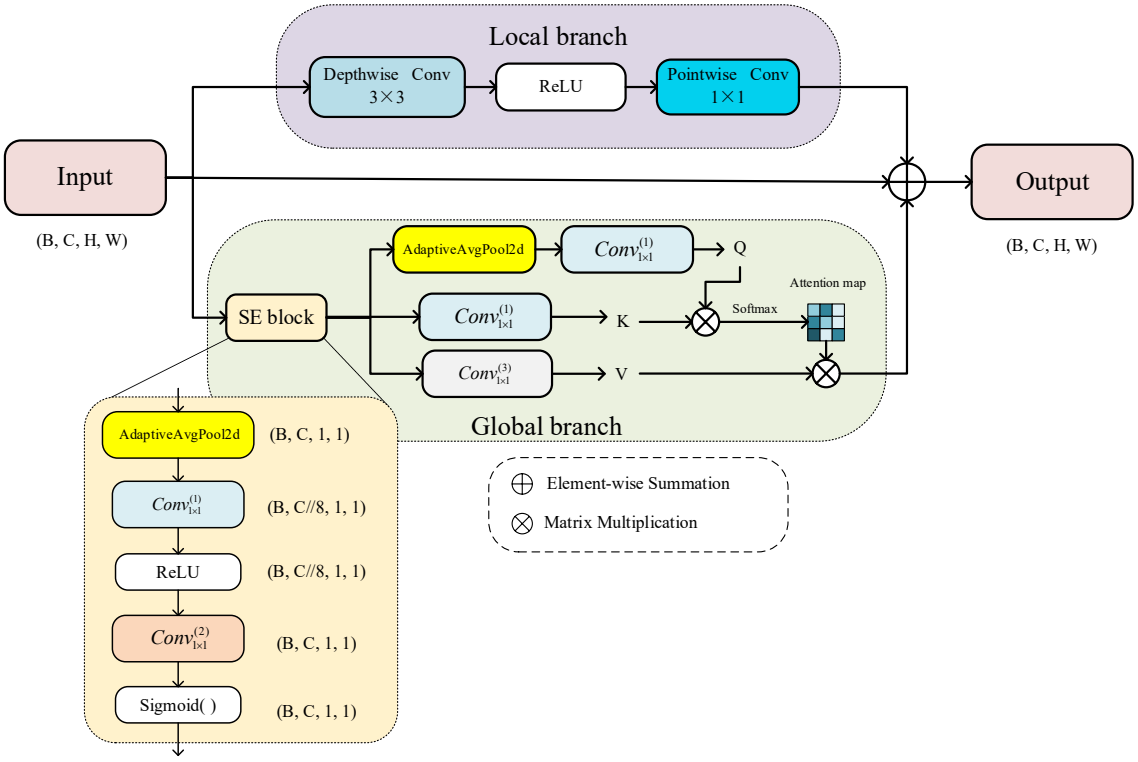


Figure 2. Illustration of Dual-Branch Convolution and Attention Residual Module (DBCARM)

In the global branch, the input feature map $x \in \mathbb{R}^{B\times C\times H\times W}$ is first refined by a Squeeze-and-Excitation (SE) block to enhance the representational power of informative channels through channel-wise attention. The SE block takes the original input x as input and outputs a channel-weighted version $\tilde{x}$. Specifically, the SE block consists of a GAP layer followed by a bottleneck structure: a 1×1 convolution reducing the channel dimension from C to C/8, a ReLU activation, another 1×1 convolution restoring the dimension to C, and a sigmoid activation. The output of the SE block is a channel-wise attention weight vector that is applied to x via element-wise multiplication, yielding the refined feature map $\tilde{x}$, as defined in Equation (9):

$$\tilde{x} = x \odot \sigma(Conv_{1\times1}^{(2)}(ReLU(Conv_{1\times1}^{(1)}(GAP(x))))) \tag{9}$$

Where $GAP(x) \in \mathbb{R}^{B\times C\times 1\times 1}$ denotes global average pooling, $Conv_{1\times 1}^{(1)}: \mathbb{R}^{C} \to \mathbb{R}^{C/8}$ denotes the operation that reduces the number of channels from C to $C/8$, $Conv_{1\times 1}^{(2)}: \mathbb{R}^{C/8} \to \mathbb{R}^{C}$ denotes the operation that restores the number of channels to C. σ is the sigmoid function. $\odot$denotes element-wise multiplication. The Output $\tilde{x} \in \mathbb{R}^{B\times C\times H\times W}$ is a channel-weighted version of x.

Subsequently, $\tilde{x}$ is processed through three parallel pathways. In the top pathway, an AdaptiveAvgPool2d operation compresses the spatial dimensions to 1×1, followed by $Conv_{1\times 1}^{(1)}$ that projects the feature into a global context-aware query vector $Q \in \mathbb{R}^{B\times \lfloor C/8 \rfloor \times 1\times 1}$ with an appropriately reshaped form $\hat{Q} \in \mathbb{R}^{B\times 1\times \lfloor C/8 \rfloor}$.

The middle pathway applies $Conv_{1\times 1}^{(1)}$ to $\tilde{x}$ to generate local key matrix $K \in \mathbb{R}^{B\times \lfloor C/8 \rfloor \times HW}$, which encodes spatially distributed "attention-receptive" features across the feature map. In the bottom pathway, another 1×1 convolution, denoted as $Conv_{1\times 1}^{(3)}$, is applied to $\tilde{x}$ in order to produce the value matrix $V \in \mathbb{R}^{B\times C\times HW}$, preserving full channel expressiveness without dimension. Finally, the attention weights are computed via matrix multiplication between the reshaped query $\hat{Q}$ and the flattened key K, as formulated in Equation (10):

$$energy = softmax(\hat{Q}^{T} \cdot K) \tag{10}$$

The attention weights energy are then used to perform a weighted sum, as defined in Equation (11):

$$Aggregated = energy \cdot V^{T} \tag{11}$$

In the local branch, a 3×3 depthwise separable convolution is first applied to extract local spatial features, followed by a ReLU activation function to introduce non-linearity. Subsequently, a 1×1 convolution to enhance cross-channel interactions while preserving the spatial resolution of the feature map. The local branch is formulated as shown in Equation (12):

$$LocalBranch(x) = Conv_{1\times 1}(ReLU(Conv_{3\times 3}^{depthwise}(x))) \tag{12}$$

Where $Conv_{3\times 3}^{depthwise}$ denotes a 3×3 depthwise separable convolution, and $Conv_{1\times 1}$ represents the standard pointwise convolution. The aggregated local and global features are then fused with the original input via a residual connection, as expressed in Equation (13):

$$x_{out} = x + Aggregated + LocalBranch(x) \tag{13}$$

3.4. GMSFN: Gated Multi-Scale Feedforward Network

To enhance the capability for nonlinear feature transformation, we propose the Gated Multi-Scale Feed-forword Network (GMSFN), which aggregates multi-scale features to enhance nonlinear information transformation for classification.

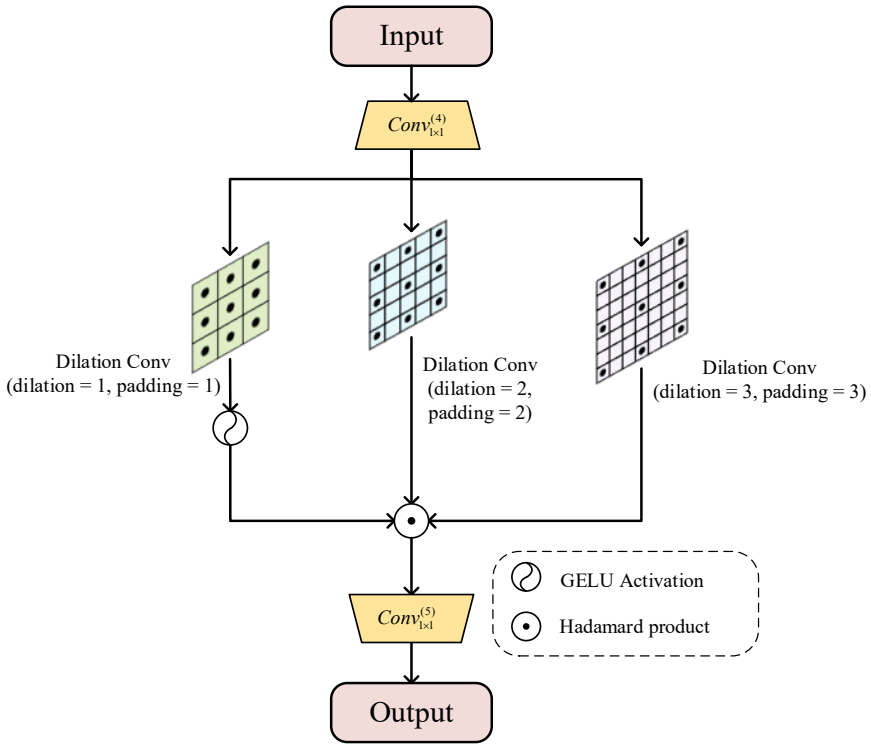


Figure 3. Illustration of GMSFN Architecture

The architecture of GMSFN is illustrated in Figure 3. The input tensor $x \in \mathbb{R}^{B\times C\times H\times W}$ is first projected into a higher-dimensional space via a 1×1 convolution with an expansion ratio of $\gamma = 3$, resulting in a feature map $\tilde{x} \in \mathbb{R}^{B\times 3C\times H\times W}$. This expansion is implemented as Equation (14):

$$\tilde{x} = Conv_{1\times1}^{(4)}(x) \tag{14}$$

Where $Conv_{1\times1}^{(4)}$ denotes a 1×1 convolution that maps C channels to 3C. The expanded feature $\tilde{x}$ is then split along the channel dimension into three equal parts: $x_1, x_2, x_3 \in R^{B\times C\times H\times W}$, corresponding to the three parallel branches.

Each branch is independently processed by a 3×3 depthwise convolution with different dilation rates of 1, 2 and 3, respectively, where the corresponding padding values are set to match the dilation (i.e., 1, 2 and 3, respectively) to maintain spatial resolution across all branches. This design enables multi-scale contextual modeling: the first branch captures local spatial details, the second expands the receptive field while preserving spatial resolution and the third models broader context, these outputs are denoted as f_1, f_2, f_3, respectively, and depicted in Equation (15)-(17).

$$f_1 = GELU(W_{3\times3}^{d=1}(x_1)) \tag{15}$$

$$f_2 = W_{3\times3}^{d=2}(x_2) \tag{16}$$

$$f_3 = W_{3\times3}^{d=3}(x_3) \tag{17}$$

Where $W_{3\times3}^{d=1}$, $W_{3\times3}^{d=2}$ and $W_{3\times3}^{d=3}$ indicate 3×3 dilated convolutions with dilation rates of 1, 2 and 3, respectively.

To introduce non-linearity and enhance feature discrimination, we apply the GELU activation function to the output of the first branch before fusing the three outputs through element-wise

multiplication. This forms an adaptive non-linear gating mechanism, where local details dynamically modulate the fusion of multi-scale features. The fused representation is then projected back to the original channel dimension using another 1×1 convolution.

The computational formula for the GMSFN module is defined in Equation (18):

$$x_{out} = Conv_{1\times1}^{(5)}(f_1 \odot f_2 \odot f_3) \tag{18}$$

Where $\odot$ denotes element-wise multiplication, $Conv_{1\times1}^{(5)}$ reduces the channel dimension from $3C$ to C.

3.5. Experimental Evaluation Metrics

To address the class imbalance and feature complexity present in the dataset, the following six metrics are specifically chosen to evaluate the model's classification performance from different perspectives: performance evaluation metrics include test set Accuracy (ACC), Precision (Pre), Specificity (Spec), Sensitivity (Sen), F1-score and Area under the curve (AUC). In this context, the positive class (labeled as 1) corresponds to malignant samples, while the negative class (0) denotes benign cases. Let TP, FP, TN and FN denote the counts of true positives, false positives, true negatives, and false negatives, respectively. These values form the basis for all subsequent metrics.

Specifically, Accuracy (ACC) represents the proportion of correctly classified images in the test set and is computed as Equation (19):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{19}$$

Where a higher ACC values indicate better classification performance.

Precision (Pre) reflects the model's reliability of identifying malignant samples, with its expression defined in Equation (20):

$$Pre = \frac{TP}{TP + FP} \tag{20}$$

This metric evaluates the model's ability to control misdiagnosis of malignant cases [22].

Sen measures the model's capacity to correctly identify malignant samples and is computed according to Equation (21):

$$Sen = \frac{TP}{TP + FN} \tag{21}$$

It evaluates how effectively the model identifies true positives, thus reducing the risk of missed diagnoses [23].

Spec assesses the model's ability to correctly identify negative samples, as defined in Equation (22):

$$Spec = \frac{TN}{TN + FP} \tag{22}$$

High specificity indicates strong performance in avoiding over-diagnosis [24].

The F1-score is the harmonic mean of Precision and Sensitivity providing a balanced assessment especially under class imbalance, with its expression given in Equation (23):

$$F1 = 2 \times \frac{Pre \times Sen}{Pre + Sen} \tag{23}$$

The AUC represents the area under the Receiver Operating Characteristic (ROC) curve, where the ROC curve is plotted with the False Positive Rate (FPR) on the x-axis and the True Positive Rate (TPR) on the y-axis. TPR, also known as Sensitivity, represents the proportion of positive samples correctly identified by the model. FPR is the proportion of negative samples incorrectly identified as positive, its expression given by Equation (24):

$$FPR = \frac{FP}{FP + TN} \tag{24}$$

The AUC is computed as defined in Equation (25):

$$AUC = \int_0^1 TPR\,(FPR)d(FPR) \tag{25}$$

A higher AUC value (closer to 1.0) indicates superior discriminative ability between malignant and benign samples.

4. EXPERIMENTAL RESULTS

4.1. Dataset

We evaluate the proposed method on two public breast histopathology datasets: BreakHis and the Invasive Ductal Carcinoma (IDC) dataset. Both consist of H&E-stained whole slide images digitized using Aperio scanners, with benign and malignant regions annotated by pathologists. All images are in 3-channel RGB format (PNG).

The BreakHis (97 patients) contains manually extracted patches at 40×, 100×, 200×, and 400× magnifications (700 × 460 pixels) [25]. A total of 7909 patches are selected for this study, forming the source domain for training and validating.

The IDC dataset (162 patients) comprises 50 × 50 pixel patches automatically generated via sliding windows with binary labels ("IDC"/"non-IDC") [26]. To evaluate cross-dataset generalization and serve as an extension of the study. BreakHis serves as the source domain, and IDC as the target domain for external validation. A subset of 7906 patches is randomly sampled from IDC while preserving the original class distribution, ensuring a fair and computationally efficient evaluation.

4.2. Training and Testing Methods

4.2.1. Implementation and Hardware Setup

The model was implemented using the PyTorch framework. All experiments were conducted on a hardware platform equipped with an Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz processor and an NVIDIA GeForce RTX 4090 GPU. During model training and evaluation, the entire model was deployed on the GPU, fully utilizing the GPU's computational capabilities through parallel computing strategies.

4.2.2. Dataset Partitioning and Training Configuration

The BreakHis dataset is divided into training, validation, and test sets with an 80%–10%–10% split, respectively. During training, images are processed in mini-batches of size 64 and fed into the HGCTransformer model. The model parameters are optimized using the Adaptive Moment

Estimation (Adam) [27] algorithm with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}. To adaptively adjust the learning rate during training, a ReduceLROnPlateau [28] scheduler is employed, which reduces the learning rate by a factor of 0.5 whenever the validation loss shows no improvement over a specified number of epochs. Specifically, the scheduler monitors the validation loss in 'min' mode with a patience of 8 epochs before reducing the learning rate. The learning rate is updated after each epoch, training proceeds for a maximum of 200 epochs, with early stopping applied if the validation loss does not improve for 15 consecutive epochs. Additionally, a dropout layer with a rate of 0.1 is incorporated for regularization to mitigate overfitting.

In the default configuration, the Adaptive Backbone Ensemble employs a 34-layer residual network (denoted as M_{32}^{1}) as the feature extractor, the network architecture consists of an initial 7×7 convolutional layer, followed by a max-pooling layer. Subsequently, the network is composed of four stacked residual blocks with repeated 3×3 convolutional layer [29]. The backbone is initialized with weights pre-trained on ImageNet, The backbone is initialized with weights pre-trained on ImageNet [30], leveraging transfer learning to accelerate convergence and improve generalization on the histopathological classification task.

4.3. Data Preprocessing and Augmentation Techniques

4.3.1. Data Augmentation

To enhance the robustness and generalization of the model, a comprehensive set of data augmentation techniques is applied during training, including random resized cropping, horizontal and vertical flipping, random rotation, and affine shearing, which collectively simulate variations in scale, orientation, and spatial deformation commonly observed in histopathological images. The specific techniques and their corresponding parameter values are detailed in Table 2.

Table 2. Data Augmentation Techniques and Parameters

Augmentation	Parameter Value
Random Resized Crop	Size = 224
Random Horizontal Flip	Probability = 0.5
Random Vertical Flip	Probability = 0.5
Random Affine	Shear Range = [−10°, 10°]
Random Rotation	Angle Range = [-5°, 5°]

4.3.2. Data Preprocessing

To standardize input images for model compatibility, deterministic preprocessing operations are applied to the training, validation, and test sets, ensuring alignment with pretrained model requirements. Specific operations and parameters are shown in Table 3.

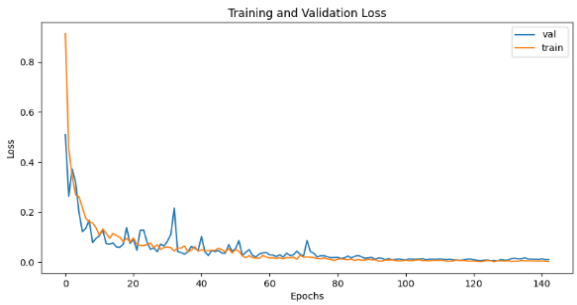
Table 3. Data Preprocessing Techniques and Parameters

Preprocessing	Parameter Value
Resize	Size = 256 × 256
CenterCrop	Size = 224
ToTensor	Range = [0, 1]
Normalize	mean=[0.714, 0.599, 0.775], std=[0.155, 0.198, 0.160]

4.4. Experimental Result

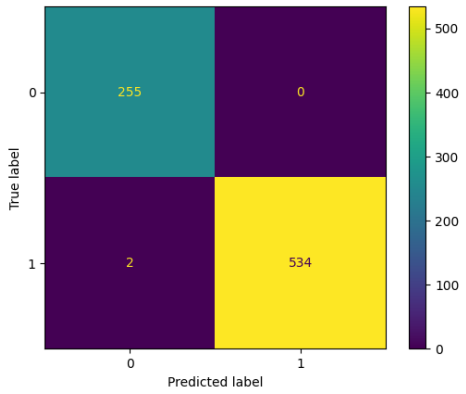
As training epochs increase, both training and validation losses converge stably after approximately 100 epochs with no significant overfitting (Figure 4a). The corresponding accuracies stabilize above 99.6% (maximum gap of less than 1%), indicating strong generalization, as shown in Figure 4b. On

the independent test set (n=791), the model achieves 99.7% ACC (Figure 4c). The Receiver Operating Characteristic (ROC) curve achieves an AUC of 1.00(rounded to two decimal places), as depicted in Figure 4d.

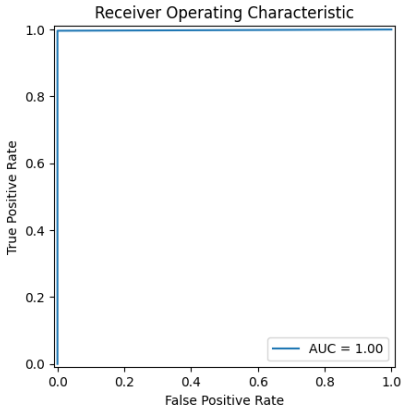


a. Training and Validation Loss Curves

b. Training and Validation Accuracy Curves



c. Confusion Matrix



d. ROC Curve

Figure 4. Model performance evaluation result

Table 4 compares the classification performance of various models on the BreakHis dataset, including ViT [13], CvT [14], MaxViT [15], SepViT [16], MedViT [17], HCANet [18], HCTNet [31], and HyFormer [32]. Under identical training/test splits and evaluation protocols, all baseline methods were re-implemented using their publicly available code. The proposed HGCTransformer achieves superior performance with ACC=99.76%, Pre=0.99, Spec=1.00, F1-score=1.00, and AUC=1.00(all metrics reported to two decimal places).

Table 4. Comparison of Model Classification Metrics (Bold means the best)

Model	ACC	Pre	Sen	Spec	F1-score	AUC
ViT [13]	82.27	0.82	0.82	0.76	0.78	0.76
CvT [14]	86.72	0.88	0.87	0.88	0.86	0.87
MaxViT [15]	91.57	0.89	0.92	0.91	0.90	0.91
SepViT [16]	80.76	0.80	0.80	0.72	0.75	0.72
MedViT [17]	92.37	0.91	0.92	0.92	0.92	0.92
HCANet [18]	89.82	0.89	0.90	0.89	0.89	0.88
HCTNet [31]	94.95	0.93	0.95	0.94	0.94	0.94
HyFormer [32]	97.12	0.96	0.97	0.96	0.97	0.96
HGCTransformer (Ours)	99.76	1.00	1.00	0.99	1.00	1.00

4.5. Ablation Study

To evaluate the effectiveness of each component in the proposed HGCTransformer model, a systematic ablation study was conducted, and the results are summarized in

Table 5. This analysis quantifies the contribution of key modules-DBCARM, GMSFN-as well as the role of established components such as M_{32}^{1} and pre-layer normalization [33].

When integrated into the baseline model (ACC = 93.8%), DBCARM improves ACC to 95.94% and boosts the F1-score to 0.97, demonstrating its effectiveness to refine feature representations through adaptive channel weighting and long-range dependencies modeling. Independently, GMSFN achieves ACC=97.22% and AUC =0.97, highlighting its strength in capturing discriminative multi-scale features via non-linear transformations across varying receptive fields. In contrast, M_{32}^{1} alone yields ACC=97.02%, indicating limited adaptability to the target dataset; however, when combined with either DBCARM (97.98%) or GMSFN (98.32%), performance increases significantly, confirming its role as a stable feature extractor that complements the proposed modules despite not being novel. Further analysis shows that combining DBCARM and GMSFN yields ACC=97.60% — competitive but inferior to configurations incorporating M_{32}^{1} —while the strong performance of GMSFN+M_{32}^{1} underscores the value of hierarchical convolutional features enhanced by multi-scale modeling.

Table 5. Comparison of Classification Metrics in Ablation Experiments. (Bold means the best)

Model	ACC	Pre	Sen	Spec	F1-score	AUC
BaseLine	93.8	0.93	0.0.94	0.93	0.93	0.93
+DBCARM	95.94	0.97	0.97	0.95	0.97	0.95
+GMSFN	97.22	0.97	0.97	0.97	0.97	0.97
+M_{32}^{1}	97.02	0.96	0.97	0.97	0.98	0.97
+DBCARM+GMSFN	97.6	0.97	0.98	0.97	0.97	0.97
+DBCARM+M_{32}^{1}	97.98	0.98	0.98	0.98	0.98	0.98
+GMSFN+M_{32}^{1}	98.32	0.98	0.98	0.98	0.98	0.98
HGCTransformer (Ours)	99.76	1.00	1.00	0.99	1.00	1.00

(DBCARM: Dual-Branch Convolution and Attention Residual Module, GMSFN: Gated Multi-Scale Feedforward Network)

The complete HGCTransformer model, integrating DBCARM, GMSFN, and M_{32}^{1}, achieves ACC=99.76%, an AUC=1.00, an F1-score=1.00, outperforming all ablated variants by up to 1.74 percentage points. This superior performance confirms the necessity of the synergistic design and validates the complementary roles of the components.

In summary, M_{32}^{1} provides foundational feature representations, the proposed DBCARM and GMSFN modules play pivotal roles in context refinement and multi-scale modeling. Their integration within the HGCTransformer framework enables highly accurate and robust decision-making, demonstrating the model's effectiveness in complex visual recognition tasks.

4.6. Cross-Dataset Generalization Evaluation

To assess the generalizability of the proposed HGCTransformer across diverse histopathological imaging domains, we evaluated its performance on a representative subset of the IDCDataset (approximately 277,524 images) under a strict cross-domain setting without fine-tuning—a challenging test of model robustness. Using a consistent 80%-10%-10% train-validation-test split and identical data augmentation (random rotation, flipping) and normalization protocols, all models were compared under controlled conditions.

As shown in Table 6, HGCTransformer achieves an ACC of 93.42%, an AUC of 0.93, significantly surpassing state-of-the-art architectures such as ViT (ACC=87.22%, AUC=0.82), MaxViT (91.8%, 0.86), MedViT (90.51%, 0.92), as well as recently proposed models HCTNet (88.15%, 0.81) and HyFormer (88.78%, 0.82).

The consistent performance gap, particularly against transformer-based models such as ViT, MaxViT, MedViT, HCTNet and HyFormer, suggests that the integration of DBCARM and GMSFN enhances feature discriminability under domain shift. While these models exhibit lower AUC and specificity, likely due to overfitting on local artifacts or noise, HGCTransformer leverages hierarchical convolutional features (M_{32}^{1}) and multi-scale contextual modeling to stabilize learning and improve generalization.

These results demonstrate that HGCTransformer is not only effective on the original BreakHis dataset but also exhibits strong transferability to heterogeneous histopathological data. The model's ability to generalize across different acquisition pipelines underscores its potential for real-world applications in digital pathology.

Table 6. Cross-Dataset Generalization Performance on IDCDataset. (Bold means the best)

Model	ACC	Pre	Sen	Spec	F1-score	AUC
ViT [13]	87.22	0.81	0.82	0.82	0.82	0.82
MaxViT [15]	91.8	0.88	0.87	0.86	0.87	0.86
MedViT [17]	90.51	0.84	0.92	0.92	0.87	0.92
HCTNet [31]	88.15	0.81	0.90	0.90	0.84	0.90
HyFormer [32]	88.78	0.82	0.89	0.89	0.85	0.89
HGCTransformer	93.42	0.88	0.93	0.93	0.90	0.93

5. CONCLUSION

HGCTransformer performs excellently in the classification of breast cancer histopathological images, achieving good results on the BreakHis dataset and maintaining robust results when evaluated on the independent IDCDataset. The integration of DBCARM and GMSFN modules further enhances feature differentiation and cross-scale modeling, which helps improve generalization ability under different imaging conditions. These findings indicate that this technology has great application value in automated pathology workflows, especially in situations where consistent diagnostic support is required.

Nevertheless, there are still several practical challenges to be addressed before its clinical application. The current architecture relies on a deep convolutional backbone, resulting in a relatively large model that may hinder deployment in clinical environments with limited computational resources. Model compression techniques, such as pruning or quantization, can help reduce inference time and memory footprint, making it more suitable for bedside settings or mobile systems. In addition, although the model performs well on a carefully selected dataset, its behavior under real-world variations (such as staining, scanning equipment, or tissue processing differences) needs further investigation through multicenter prospective studies.

In addition to technical improvements, a more meaningful advance would be to combine histopathological features with genomic and clinical data. Combining imaging phenotypes with molecular characteristics, such as ER/PR/HER2 status, gene expression profiles, may enable more biologically informed predictions, including tumor aggressiveness and treatment response. This multimodal approach, if validated in different patient populations, could go beyond binary classification to enable personalized diagnostic support. For any of these steps to succeed, collaboration between computer scientists, pathologists, and clinicians is crucial to ensure that model development aligns with practical clinical needs.

REFERENCES

[1] BRAY F, LAVERSANNE M, SUNG H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. CA: a cancer journal for clinicians, 2024, 74(3): 229-63.

[2] OBEAGU E I, OBEAGU G U. Breast cancer: A review of risk factors and diagnosis [J]. Medicine, 2024, 103(3): e36905.

[3] DE MARCO P, RICCIARDI V, MONTESANO M, et al. Transfer learning classification of suspicious lesions on breast ultrasound: is there room to avoid biopsies of benign lesions? [J]. European Radiology Experimental, 2024, 8(1): 121.

[4] TSENG L-J, MATSUYAMA A, MACDONALD-DICKINSON V. Histology: The gold standard for diagnosis? [J]. The Canadian Veterinary Journal, 2023, 64(4): 389.

[5] SILVERWOOD S M, BACKER G, GALLOWAY A, et al. Assessing the rates of false-positive ovarian cancer screenings and surgical interventions associated with screening tools: a systematic review [J]. BMJ oncology, 2024, 3(1): e000404.

[6] BOMEISL P, GILMORE H. Spectrum of atypical ductal hyperplasia (ADH) and ductal carcinoma in-situ (DCIS): Diagnostic challenges; proceedings of the Seminars in Diagnostic Pathology, F, 2024 [C]. Elsevier.

[7] JIMENEZ-DEL-TORO O, OTáLORA S, ANDERSSON M, et al. Analysis of histopathology images: From traditional machine learning to deep learning [M]. Biomedical texture analysis. Elsevier. 2017: 281-314.

[8] MANKKI J-J, BOCHENINA K. Vision Transformers in Brain Image Segmentation [J]. 2025.

[9] IQBAL N, MUMTAZ R, SHAFI U, et al. Gray level co-occurrence matrix (GLCM) texture based crop classification using low altitude remote sensing platforms [J]. PeerJ Computer Science, 2021, 7: e536.

[10] SERTE S, DEMIREL H. Gabor wavelet-based deep learning for skin lesion classification [J]. Computers in biology and medicine, 2019, 113: 103423.

[11] SHEN D, JI Y, LI P, et al. Ranet: Region attention network for semantic segmentation [J]. Advances in Neural Information Processing Systems, 2020, 33: 13927-38.

[12] ALQAHTANI Y, MANDAWKAR U, SHARMA A, et al. Breast cancer pathological image classification based on the multiscale CNN squeeze model [J]. Computational Intelligence and Neuroscience, 2022, 2022(1): 7075408.

[13] HAN K, WANG Y, CHEN H, et al. A survey on vision transformer [J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 87-110.

[14] WU H, XIAO B, CODELLA N, et al. Cvt: Introducing convolutions to vision transformers; proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, F, 2021 [C].

[15] TU Z, TALEBI H, ZHANG H, et al. Maxvit: Multi-axis vision transformer; proceedings of the European conference on computer vision, F, 2022 [C]. Springer.

[16] LI W, WANG X, XIA X, et al. Sepvit: Separable vision transformer [J]. arXiv preprint arXiv:220315380, 2022.

[17] MANZARI O N, AHMADABADI H, KASHIANI H, et al. MedViT: a robust vision transformer for generalized medical image classification [J]. Computers in biology and medicine, 2023, 157: 106791.

[18] HU S, GAO F, ZHOU X, et al. Hybrid convolutional and attention network for hyperspectral image denoising [J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 1-5.

[19] NEYESTANAK M S, JAHANI H, KHODARAHMI M, et al. A quantitative comparison between focal loss and binary cross-entropy loss in brain tumor auto-segmentation using U-Net [J]. Journal of Biostatistics and Epidemiology, 2025, 11(1): 15-35.

[20] HE J. Gradient reweighting: Towards imbalanced class-incremental learning; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2024 [C].

[21] PEREIRA G A, HUSSAIN M. A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships [J]. arXiv preprint arXiv:240815178, 2024.

[22] MIAO J, ZHU W. Precision–recall curve (PRC) classification trees [J]. Evolutionary intelligence, 2022, 15(3): 1545-69.

[23] RAINIO O, TAMMINEN J, VENäLäINEN M S, et al. Comparison of thresholds for a convolutional neural network classifying medical images [J]. International Journal of Data Science and Analytics, 2024: 1-7.

[24] BIGGERI A, FORNI S, BRAGA M. The risk of over-diagnosis in serological testing. Implications for communications strategies [J]. Epidemiologia e Prevenzione, 2020, 44(56): 184-92.

[25] TOMA T A, BISWAS S, MIAH M S, et al. Breast cancer detection based on simplified deep learning technique with histopathological image using BreaKHis database [J]. Radio Science, 2023, 58(11): 1-18.

[26] GUPTA I, NAYAK S R, GUPTA S, et al. A deep learning based approach to detect IDC in histopathology images [J]. Multimedia Tools and Applications, 2022, 81(25): 36309-30.

[27] YANG J, LONG Q. A modification of adaptive moment estimation (adam) for machine learning [J]. Journal of Industrial and Management Optimization, 2024, 20(7): 2516-40.

[28] THAKUR A, GUPTA M, SINHA D K, et al. Transformative breast Cancer diagnosis using CNNs with optimized ReduceLROnPlateau and Early stopping Enhancements [J]. International Journal of Computational Intelligence Systems, 2024, 17(1): 1-18.

[29] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition; proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, F, 2016 [C].

[30] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database; proceedings of the 2009 IEEE conference on computer vision and pattern recognition, F, 2009 [C]. Ieee.

[31] ZHAO Y, FU C, XU S, et al. HCT-Net: A hybrid CNN-Transformer network for multi-class cervical cell classification [J]. Biomedical Signal Processing and Control, 2026, 112: 108383.

[32] YAN C, FAN X, FAN J, et al. Hyformer: Hybrid transformer and cnn for pixel-level multispectral image land cover classification [J]. International Journal of Environmental Research and Public Health, 2023, 20(4): 3059.

[33] LI P, YIN L, LIU S. Mix-ln: Unleashing the power of deeper layers by combining pre-ln and post-ln [J]. arXiv preprint arXiv:241213795, 2024.