

ORIGINAL RESEARCH ARTICLE

DualPath-CRC: Adaptive Swin Transformer blocks and Attention-Augmented InceptionNeXt for classification of colorectal cancer histopathology

Uddagiri Sirisha^{1*}, Appalaraju Grandhi², Panguluri Padmavathi³, Jyothi Desireddy⁴, and Kannaiah Chattu⁵¹Department of Computer Science and Engineering, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India²Department of Computer Science and Engineering, School of Computing, Aditya University, Kakinada, Andhra Pradesh, India³Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India⁴Department of Computer Science and Engineering, Lakireddy Bali Reddy College of Engineering, Vijayawada, Andhra Pradesh, India⁵Department of Artificial Intelligence and Machine Learning, School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India***Corresponding author:**Uddagiri Sirisha
(sirisha.u@pvpsiddhartha.ac.in)**Citation:** Sirisha U, Grandhi A, Padmavathi P, Desireddy J, Chattu K. DualPath-CRC: Adaptive Swin Transformer blocks and Attention-Augmented InceptionNeXt for classification of colorectal cancer histopathology. *Artif Intell Health*. doi: 10.36922/AIH026260072**Received:** June 27, 2026**Revised:** July 25, 2026**Accepted:** August 3, 2026**Published online:** August 18, 2026**Copyright:** © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.**Publisher's Note:** AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.**Abstract**

Colorectal cancer (CRC) is the third most common cancer and the second leading cause of cancer-related death worldwide, with about 1.85 million cases and 850,000 deaths annually. While a definitive diagnosis depends on the histopathological examination of hematoxylin and eosin-stained tissue slides, considerable variability remains in slide interpretation among pathologists, and manual examination is time-consuming. These limitations have led to the development of computer-aided methods for analyzing colorectal histopathology slides. Recent hybrid Transformer models often lack adaptive channel re-weighting, use fixed window sizes that may restrict the multi-scale receptive field, or rely on traditional feature-fusion techniques that assume simple element-wise addition. To overcome these shortcomings, DualPath-CRC proposes three architectural innovations: (i) Attention-Augmented InceptionNeXt blocks, which apply sigmoid-based channel attention to multi-branch depthwise convolutions; (ii) Adaptive Window Swin Transformer blocks, in which a Window-Size Prediction Network adaptively controls the receptive field of self-attention; and (iii) a Gated Cross-Pathway Fusion module, which fuses the local and global feature streams through learned sigmoid gating. For the classification of histological tissue patches, DualPath-CRC achieves an accuracy of 99.98% on NCT-CRC-HE-100K and 99.34% on Kather-5K. The regions highlighted by Grad-CAM, Grad-CAM++, and LIME analyses provide qualitative information about which parts of the image are important for the model's predictions and typically correspond to morphologically relevant tissue areas.

Keywords: Colorectal cancer; Histopathology; Hybrid deep learning; Squeeze-and-excitation attention; Adaptive Swin Transformer; Gated feature fusion; Medical image classification; Grad-CAM

1. Introduction

1.1. Clinical motivation and epidemiological context

Cancer is a major cause of morbidity and mortality in the world. Colorectal cancer (CRC) is the third most commonly diagnosed cancer and the second leading cause of cancer-related death worldwide.^{1,2} Recent epidemiological data indicate that CRC is responsible for approximately 1.85 million new cases and 850,000 deaths annually, representing 9.8% of all cancer deaths and 9.2% of all new cancer cases globally.³ CRC mainly affects people aged 65–74 years and is slightly more common in women. One of the most important prognostic factors is the stage at diagnosis, with a five-year survival rate of almost 90% for early-stage (localized) CRC compared with only 13% for advanced-stage (metastatic) CRC. This substantial difference in prognosis highlights the need for early, accurate, and automated detection systems.

Histopathological examination of hematoxylin and eosin (H&E)-stained biopsy specimens is therefore the gold standard for diagnosing CRC.^{4–6} In clinical practice, pathologists visually examine the slides under a microscope for characteristic neoplastic features, including glandular irregularity, nuclear atypia, stromal invasion, and mucin production. This process is time-consuming because each whole-slide image (WSI) contains a large tissue area that must be examined, and slide interpretation is subject to considerable inter- and intra-observer variability.⁷ Most CRCs develop through the malignant transformation of colorectal polyps. Colonoscopy is widely used for screening, but it may miss small or flat polyps and is dependent on operator experience.^{8,9}

1.2. Deep learning in medical image analysis

Deep learning has significantly transformed medical image analysis. Convolutional neural networks (CNNs)^{10,11} have demonstrated an excellent ability to automatically extract features and achieve human-level performance in a wide variety of diagnostic imaging applications, such as skin cancer detection¹², cervical cancer screening¹³, lung and colon cancer classification¹⁴, myocardial infarction detection¹⁵, leukemia subtype classification¹⁶, retinal disease diagnosis¹⁷, and brain abnormality segmentation.¹⁸ For CRC, CNN-based computer-aided design (CAD) systems using architectures such as ResNet, DenseNet, EfficientNet, and VGG have consistently improved classification accuracy and advanced the state of the art.^{19–21}

Although CNNs are extremely powerful at capturing local features, they are constrained by an inductive locality bias: they typically process spatial structures within predetermined kernel receptive fields and are less suited

to capturing long-range spatial relationships needed to represent global tissue architecture, such as the relationships between multiple tumor nests and the surrounding stromal components. Dosovitskiy *et al.*²² introduced the Vision Transformer (ViT), which addresses this limitation through self-attention mechanisms that capture relationships among all image tokens. Another efficient architecture is the Swin Transformer²³, which uses local windows to capture local image structure while retaining cross-window communication through a shifted-window mechanism. These benefits have led to the development of models that are effective in CRC classification tasks.^{24–26} However, such models can be computationally expensive and data-intensive, limiting their use in clinical settings where sufficiently large datasets may be unavailable.

1.3. Hybrid architectures and existing gaps

The limitations of pure CNN and pure ViT approaches have led to hybrid models that combine multi-scale local feature extraction with global contextual modeling. Hybrid CNN–Transformer models have demonstrated better accuracy–efficiency tradeoffs in several medical image analysis studies.^{27–30} In CRC histopathology, a hybrid architecture has been proposed that comprised InceptionNeXt blocks, which use multi-branch depthwise convolutions to capture local morphological features; enhanced Swin Transformer blocks, which use shifted-window self-attention to model long-range dependencies; and residual multilayer perceptron (ResMLP) blocks for stable deep-feature refinement. This model outperformed more than 20 state-of-the-art CNN- and ViT-based models, achieving median accuracies of 99.96% and 99.06% on NCT-CRC-HE-100K and Kather-5K, respectively.

However, the baseline leaves three structural challenges unaddressed:

- (i) Absence of channel attention: The multi-branch depthwise convolutions treat all feature channels equally, regardless of their diagnostic relevance. In pathology, diagnostically relevant staining and morphological features, such as nuclear chromatin density and mucin staining, are more informative than background slide artifacts. Squeeze-and-excitation (SE) attention has been demonstrated to improve the discriminative ability of CNNs by explicitly modeling inter-channel dependencies, but it has not yet been adopted in current CRC hybrid architectures.
- (ii) Fixed self-attention window size: The shifted-window attention blocks use a fixed window size of $M = 7$ at every layer and for every input patch. CRC histopathology images, however, contain multi-scale diagnostic patterns ranging from individual nuclei

(~5–10 μm) to glandular lumen structures (~50–200 μm) and stromal organization (~200–500 μm). A single fixed window may not capture this range because a small window may lose tissue context, whereas a large window may increase the computational cost of extracting nuclear-scale details.

- (iii) Non-adaptive feature fusion: The local CNN and global Transformer feature streams are fused by element-wise addition without adaptive fusion weights, thereby assigning equal weights to both streams regardless of spatial position or feature informativeness. When local morphology is ambiguous in certain tissue areas, global contextual features should be prioritized; conversely, for patches containing high-frequency nuclear details, local features should receive greater weight. A learned gating mechanism²⁸ is therefore needed to achieve this context-dependent balance.

1.4. Contributions of this work

To address these gaps, this paper proposes DualPath-CRC, a dual-pathway deep learning framework that incorporates three targeted architectural innovations into the hybrid CRC classification pipeline. The specific contributions are as follows:

- (i) Attention-Augmented InceptionNeXt (AA-INeXt) blocks: We integrate SE channel attention into the multi-branch depthwise convolution branches, enabling data-driven re-weighting of feature channels and enhancing the representation of informative morphological patterns while suppressing less informative features.
- (ii) Adaptive Window Swin Transformer (AW-SwinT) blocks: We introduce a Window-Size Prediction Network (WSPN), implemented as a lightweight two-layer multilayer perceptron, to predict an effective window size, $M_{\text{eff}} \in [M_{\text{min}}, M_{\text{max}}]$, from the global average-pooled feature representation of each input patch. This enables the self-attention receptive field to adapt to the visual characteristics of the input tissue patch.
- (iii) Gated Cross-Pathway Fusion (GCPF) module: We introduce the GCPF module, which fuses the local and global feature streams through an adaptively weighted sum at each spatial point. The weights are generated using a sigmoid activation function, enabling better discrimination between morphologically similar tissue categories while preserving the complementary representations of both streams.
- (iv) Comprehensive evaluation: We compare DualPath-CRC with more than 20 CNN- and ViT-based models for colorectal histopathological tissue-patch classification on the benchmark NCT-CRC-HE-

100K and Kather-5K datasets. Ablation studies are performed to determine the contribution of each proposed component. Furthermore, Grad-CAM, Grad-CAM++, and LIME analysis are used to identify the image regions that contribute to the model's predictions and assess whether these regions correspond to morphologically relevant structures.

1.5. Paper structure

Section 2 reviews the related literature on CNN, ViT, and hybrid architectures for CRC classification. Section 3 provides a detailed mathematical description of DualPath-CRC. Section 4 presents the experiments, ablation studies, and comparative analyses. Section 5 discusses the findings, limitations, and future directions. Section 6 concludes the paper.

2. Literature review

2.1. CNN-based classification of histopathology images of colorectal cancer

Kather *et al.*⁴ introduced the NCT-CRC-HE-100K dataset and provided a baseline texture-based classification method. Later studies increasingly adopted deeper CNN architectures. A VGG19-based model achieved an accuracy of 96.4%,²¹ whereas ensemble CNNs achieved accuracies of 96.16% on NCT-CRC-HE-100K and 92.83% on Kather-5K.¹⁹ Abd El-Ghany *et al.*³¹ reported an accuracy of 99.87% using EfficientNet-B0 with an adaptive dynamic learning-rate schedule. Sharkas and Attallah³² proposed Color-CADx, a triple-CNN architecture that uses the discrete cosine transform for color preprocessing and achieved an accuracy of 99.30% on NCT-CRC-HE-100K. Pan *et al.*³³ reported an accuracy of 99.48% using their proposed EL-CNN. Khalid *et al.*³⁴ developed CCDNet, which combines atrous convolution with a coordinate-attention Transformer and achieved an accuracy of 98.61%. While CNNs can effectively extract local features, their limited receptive fields constrain their ability to model global tissue patterns.

2.2. Vision Transformer methods

The ViT was proposed by Dosovitskiy *et al.*²² as an attention-based model for global image representation and classification. Liu *et al.*²³ proposed the Swin Transformer, which uses shifted-window attention and a hierarchical architecture to reduce computational complexity. For CRC classification, ViT achieved an accuracy of 94.75% on Kather-5K,²⁴ whereas an improved ViT with wavelet positional embedding achieved an accuracy of 95.25% on NCT-CRC-HE-100K.²⁵ Li²⁶ combined Swin Transformer V2 with self-supervised learning and knowledge distillation, achieving an accuracy of 96.0%. IL-MCAM, proposed by

Chen *et al.*³⁵, achieved an average accuracy of 99.78%.

2.3. Hybrid CNN–Transformer models

Hybrid architectures have been developed to combine the strengths of CNN and Transformer. Reis and Turk²⁷ successfully implemented a CNN–Transformer combination for skin cancer detection. Sriwastawa and Arul Jothi²⁸ used ViT variants for breast cancer histopathology classification. Tanveer *et al.*³⁶ developed TransNetV, which achieved an accuracy of 98.54% for CRC classification.

2.4. Channel-attention and feature-fusion techniques

Squeeze-and-excitation networks were first proposed by Hu *et al.*⁵ They explicitly model channel dependencies and recalibrate channel-wise responses using global average pooling, a two-layer multilayer perceptron bottleneck, and sigmoid activation. The convolutional block attention module³⁷ further enhanced this approach by applying channel and spatial attention sequentially. Gated fusion techniques, similar to those used in gated recurrent units, allow the contributions of multi-scale features.³⁸ Channel attention and gated fusion provide the fundamental concepts underlying the AA-INeXt blocks and GCPF module, respectively.

2.5. Nuclei-level segmentation and classification

In addition to tissue-patch classification, nuclei-level analysis has been investigated through benchmark challenges such as MoNuSAC2020³⁹, which focuses on multi-organ nucleus segmentation and classification. Such instance-level methods provide fine-grained morphological information relevant to tumor-microenvironment characterization but operate at a different level of granularity from patch-level tissue classification.^{40,41}

3. Proposed methodology: DualPath-CRC

3.1. Dataset information

3.1.1. NCT-CRC-HE-100K dataset

NCT-CRC-HE-100K is a publicly available CRC histopathology image-patch dataset comprising 100,000 nonoverlapping patches extracted from H&E-stained WSIs of CRC and normal tissues.⁴⁰ The tissue samples were obtained from the National Center for Tumor Diseases Biobank in Heidelberg and the pathology archive of University Medical Center Mannheim, Germany. The images were acquired at 20× magnification with a resolution of 224 × 224 pixels, a resolution of 0.5 μm per pixel. Macenko's color normalization method was applied to the images. The dataset comprises nine tissue classes: adipose tissue (ADI), background (BACK), debris (DEB),

lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), and colorectal adenocarcinoma epithelium (TUM). However, the released image patches do not provide patient- or slide-specific identifiers that would enable patient- or slide-level partitioning. Representative images are shown in [Figure 1](#).

3.1.2. Kather-5K dataset[4]

Kather-5K is a publicly available CRC histopathology dataset comprising 5,000 image tiles of 150 × 150 pixels, acquired at 20× magnification.⁴ The dataset is class-balanced, with 625 images in each of eight classes: adipose tissue, background (Empty), debris, lymphocytes (Lympho), mucosa, stroma, complex stroma, and tumor. The image labels were confirmed by pathologists from University Medical Center Mannheim (Germany). The public release of the image tiles does not specify the number of patients or source slides used to create the images. Representative images are shown in [Figure 2](#).

3.2. Data preprocessing

3.2.1. Data partitioning

To ensure uniform input dimensions, all images were resized to 224 × 224 pixels before training. The NCT-CRC-HE-100K and Kather-5K datasets were partitioned separately into training, validation, and test sets at ratios of 72%, 8%, and 20%, respectively, using stratified sampling by class. Because the released image patches do not provide patient- or slide-level identifiers, neither dataset could be partitioned strictly at the patient or slide level. The proposed model was trained and evaluated separately on each dataset using image-patch-level partitions.

3.2.2. Data augmentation and transfer learning

To improve model robustness and reduce overfitting, online data augmentation was applied only to the training images. The augmentation procedures included horizontal and vertical flipping, random rotation, random cropping, scaling, color jitter, MixUp, CutMix, and the addition of Gaussian noise. No additional stain normalization was applied to NCT-CRC-HE-100K because its images had already been color-normalized using Macenko's technique. To improve multiclass CRC tissue classification performance, transfer learning was performed by fine-tuning ImageNet-pretrained weights.

3.3. Architectural overview

DualPath-CRC represents a hierarchical architecture comprising four stages that incorporate attention, adaptive receptive fields, and gated feature fusion. Given H&E-stained histopathology image patches, the network

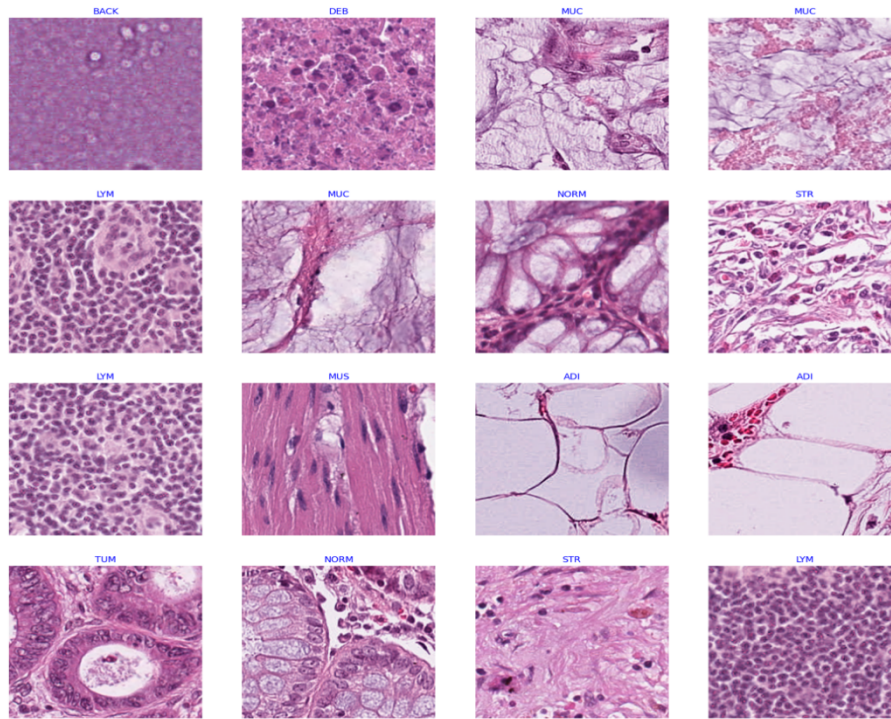


Figure 1. Sample images from the NCT-CRC-HE-100K dataset

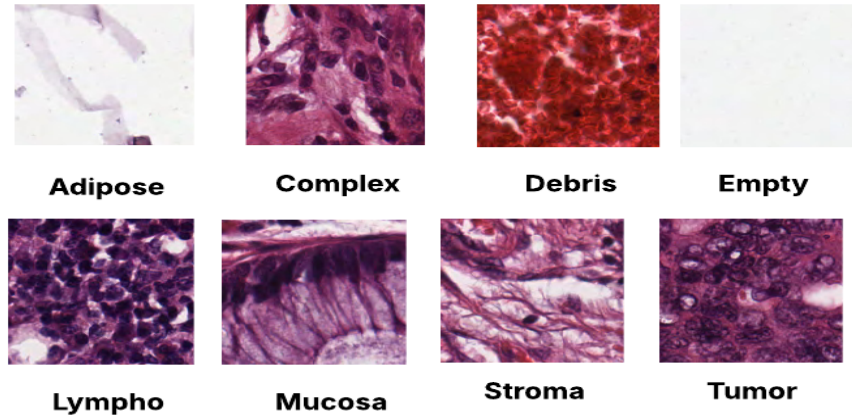


Figure 2. Sample images from the Kather-5K dataset

generates a probability distribution over nine tissue classes for NCT-CRC-HE-100K and eight tissue classes for Kather-5K. The Dualpath-CRC architecture is shown in Figure 3.

3.4. Squeeze-and-excitation channel attention

The SE structure facilitates feature recalibration on a per-channel basis. Given an input feature map $X \in R^{B \times C \times H \times W}$, the SE block first performs a squeeze operation using global average pooling to produce a channel-description

vector $z \in R^{B \times C}$, as illustrated in Equation 1:

$$z_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W X_{c,i,j} \quad (1)$$

where c indexes the channel and (i, j) indexes spatial positions. This squeeze compresses each of the spatial feature maps into a single scalar value, which captures the global channel statistics. The excitation step uses a two-

layer bottleneck multilayer perceptron, as described in **Equation 2**:

$$s = \sigma(W'_j \delta(W'_i z)) \quad (2)$$

where $W'_i \in R^{(C/r) \times C}$ & $W'_j \in R^{C \times (C/r)}$ are learnable weight matrices; $\delta(\cdot)$ denotes the rectified linear unit activation; $\sigma(\cdot)$ is the sigmoid function, and $r = 16$ is the reduction ratio. The scalar gating vector $s \in R^C$ is then used to re-scale the feature map as in **Equation 3**:

$$\tilde{X} = s \otimes X \quad (3)$$

where $\otimes$ denotes channel-wise multiplication (broadcasting s across spatial dimensions $H \times W$). This technique allows the model to suppress less contributing feature channels such as background staining artifacts or out-of-focus regions while amplifying diagnostically significant channels such as nuclear chromatin density responses or glandular edge features.

3.5. Attention-Augmented InceptionNeXt blocks

The AA-INeXt block modifies the InceptionNeXt structure by introducing an SE module into each branch, as illustrated in **Figure 4**. This design enables each branch to undergo channel recalibration independently, providing more specific channel weighting than applying a single SE module after branch fusion.

The four parallel branches operate on the same input $X \in \mathbb{R}^{B \times C \times H \times W}$, where C is the number of input channels:

- Branch 1—pointwise convolution (1×1): Captures cross-channel feature combinations at individual spatial locations.
- Branch 2— 3×3 depthwise convolution: Captures local spatial patterns, such as nuclear boundaries.
- Branch 3— 5×5 depthwise convolution: Captures intermediate-scale structural patterns, such as glandular lumen boundaries.
- Branch 4—decomposed large-kernel convolution ($1 \times K + K \times 1$): Captures coarse spatial patterns, such as stromal organization and tissue-level morphology. The default value of K is 7.

Each branch b generates a feature map via depthwise convolution and pointwise projection along with SE recalibration as follows.

$DW_b(\cdot)$ indicates the depthwise convolution in branch b , while $PW(\cdot)$ is a 1×1 pointwise projection to branch_ch = $C/4$ channels. $SE(\cdot)$ represents the SE recalibration process. The outputs from the four branches are then concatenated & projected to out_channels as described in **Equations 4 and 5**:

$$F_b = SE(PW(DW_b(X))), \quad b \in \{1, 2, 3, 4\} \quad (4)$$

$$F = PW(Concat[F_1, F_2, F_3, F_4]) \quad (5)$$

Layer normalization is applied channel-wise over the projected output, followed by Gaussian error linear unit (GELU) activation. A residual connection adds the identity-mapped input (projected to out_channels if dimensions differ) as in **Equation 6**:

$$Y = GELU(LN(F)) + Res(X) \quad (6)$$

Here, LN denotes layer normalization applied channel-wise, and $Res(\cdot)$ is a 1×1 convolution if in_channels $\neq$ out_channels; otherwise, it is an identity mapping. The re-scaling of the SEs per branch is done such that each scale of feature extraction separately finds and boosts its important channels, thus greatly enhancing discriminability compared to the non-attentive baseline.

3.6. Window-size prediction network

A fundamental limitation of the original Swin Transformer²³ and its adoption in Ref.¹ is the fixed window size $M = 7$. In CRC histopathology, optimal receptive field size varies considerably across tissue types and magnification levels. To address this, DualPath-CRC introduces the WSPN, a lightweight multilayer perceptron that dynamically predicts an effective window size M_{eff} from the aggregate feature statistics of each input. Given an input feature map $X \in \mathbb{R}^{B \times C \times H \times W}$, the WSPN computes as in **Equations 7 and 8**. During the forward pass, the effective window size is obtained as in **Equation 9**:

$$f_{gap} = GlobalAvgPool(X) \in R^{B \times C} \quad (7)$$

$$\alpha = \sigma(W_w \cdot ReLU(W_1 \cdot f_{gap}) + b_w) \quad (8)$$

$$M_{cont} = M_{min} + (\alpha \cdot (M_{max} - M_{min})), \quad M_{eff} = round(M_{cont}) \quad (9)$$

where $W_1 \in \mathbb{R}^{(c/4) \times C}$ and $W_w \in \mathbb{R}^{1 \times (c/4)}$ are learnable parameters, b_w is a bias term, and $M_{min} = 4$, $M_{max} = 12$ are predefined bounds. The predicted M_{eff} is averaged across the batch dimension and rounded to the nearest integer. During training, M_{eff} adapts based on aggregate feature statistics: patches dominated by high-frequency nuclear detail tend to yield smaller M_{eff} (sharper local attention), while patches exhibiting smooth tissue-level patterns yield larger M_{eff} (broader contextual attention). This dynamic receptive field adaptation addresses the fixed-window constraint.²³ To preserve gradient propagation during backpropagation, the rounding operation is implemented using a straight-through estimator, as in **Equation 10**,

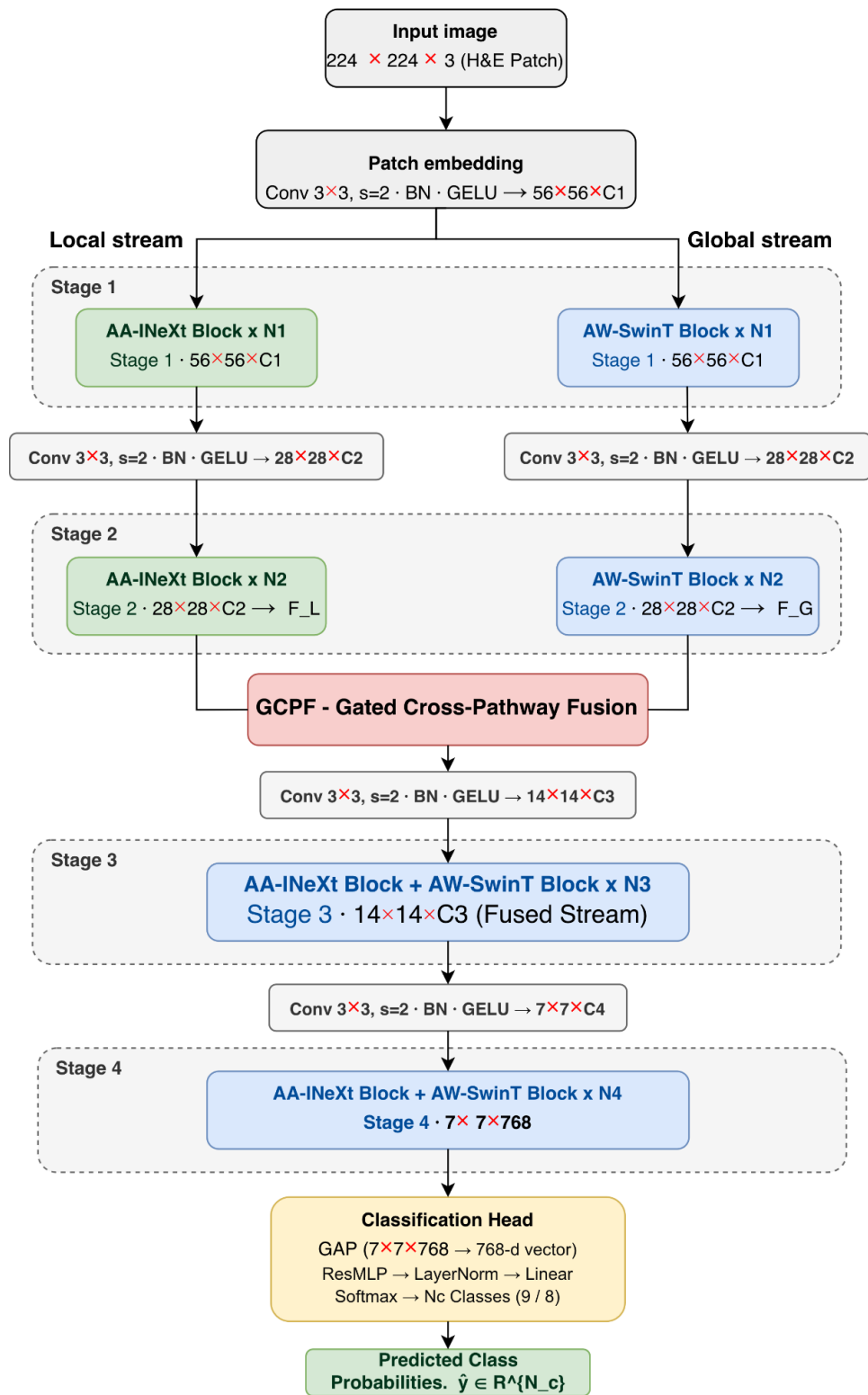


Figure 3. Overall architecture of Dualpath-CRC

Abbreviations: AA-INeXt: Attention-Augmented InceptionNeXt; AW-SwinT: Adaptive-Window Swin Transformer; BN: Batch normalization; Conv: Convolution; GAP: Global average pooling; GCPF: Gated Cross-Pathway Fusion; GELU: Gaussian error linear unit; H&E: Hematoxylin and eosin; LayerNorm: Layer normalization; MLP: Multilayer perceptron; ResMLP: Residual multilayer perceptron.

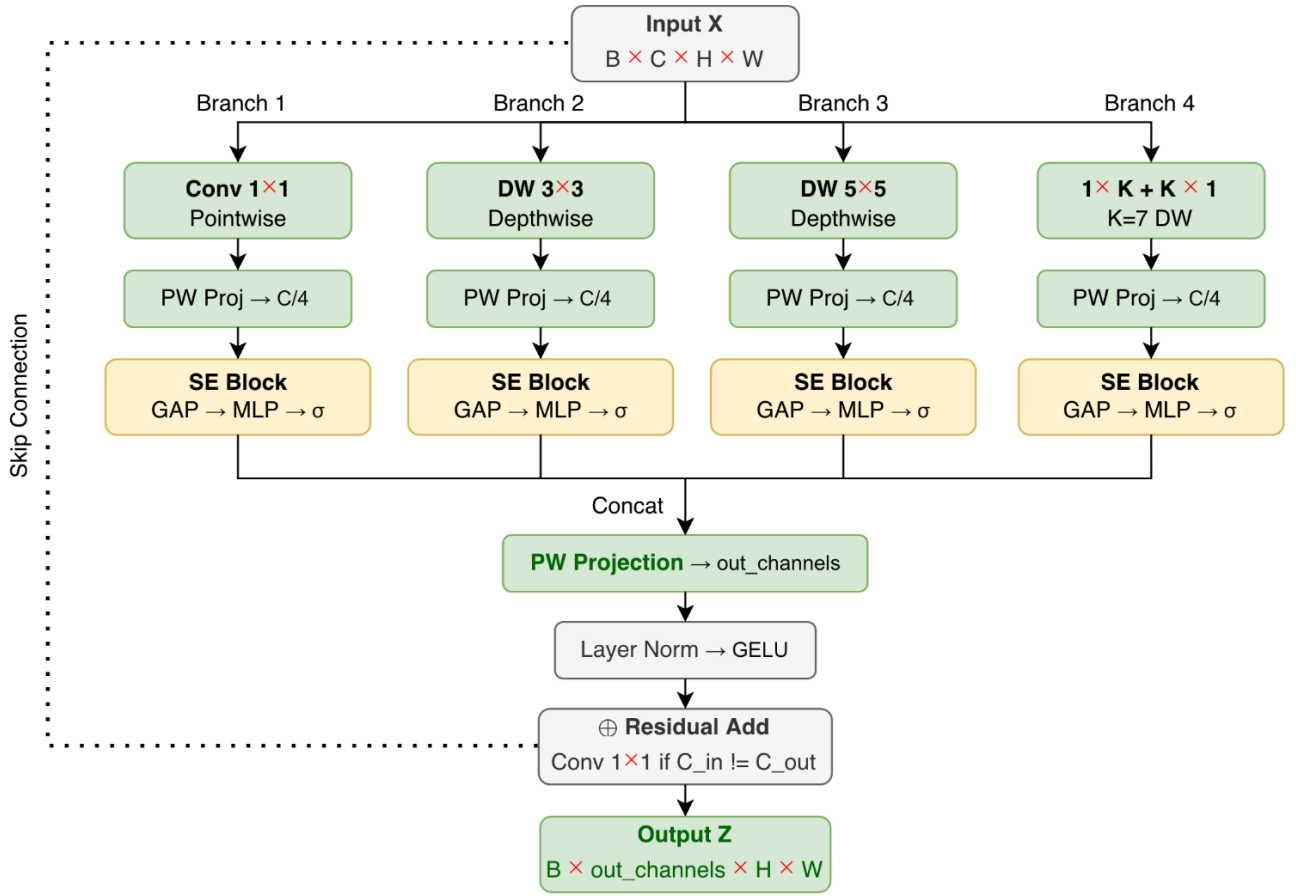


Figure 4. Attention-Augmented InceptionNeXt block

Abbreviations: Conv: Convolution; DW: Depthwise convolution; GAP: Global average pooling; GELU: Gaussian error linear unit; MLP: Multilayer perceptron; Proj: Projection; PW: Pointwise convolution; SE: Squeeze-and-excitation.

allowing gradients from the classification loss to update the WSPN parameters while maintaining an integer window size during inference.

$$\frac{\partial M_{eff}}{\partial M_{cont}} \approx 1 \quad (10)$$

To enable end-to-end optimization, the WSPN first predicts a continuous window size within the predefined range $[M_{min}, M_{max}]$. During the forward pass, the continuous value is discretized to the nearest integer for window partitioning. Since the rounding operation is non-differentiable, a straight-through estimator is employed during backpropagation, allowing gradients to flow through the continuous prediction while preserving discrete window selection during inference. This enables the WSPN to learn adaptive window sizes directly from the classification objective.

3.7. Adaptive Window Swin Transformer block

The AW-SwinT block consists of two consecutive sub-layers (as in **Figure 5**)—a fixed-window MHSA sub-layer (W-MHSA) and a shifted-window MHSA sub-layer (SW-MHSA)—both operating with the dynamically predicted window size M_{eff} from the WSPN. Each sub-layer is followed by a ResMLP component.

Given input $X \in \mathbb{R}^{B \times C \times H \times W}$, it is first reshaped to $X_{hw} \in \mathbb{R}^{B \times H \times W \times C}$ for window operations. Sub-layer 1 (W-MHSA) as in **Equations 11 and 12**:

$$\widehat{T1}^l = \text{W-MHSA}(\text{LN}(T^{l-1})) + T1^{l-1} \quad (11)$$

$$T1^l = \text{ResMLP}(\text{LN}(\widehat{T1}^l)) + \widehat{T1}^l \quad (12)$$

Sub-layer 2 (SW-MHSA), with cyclic shift of $(M_{eff}/2, M_{eff}/2)$, are described in **Equations 13 and 14**:

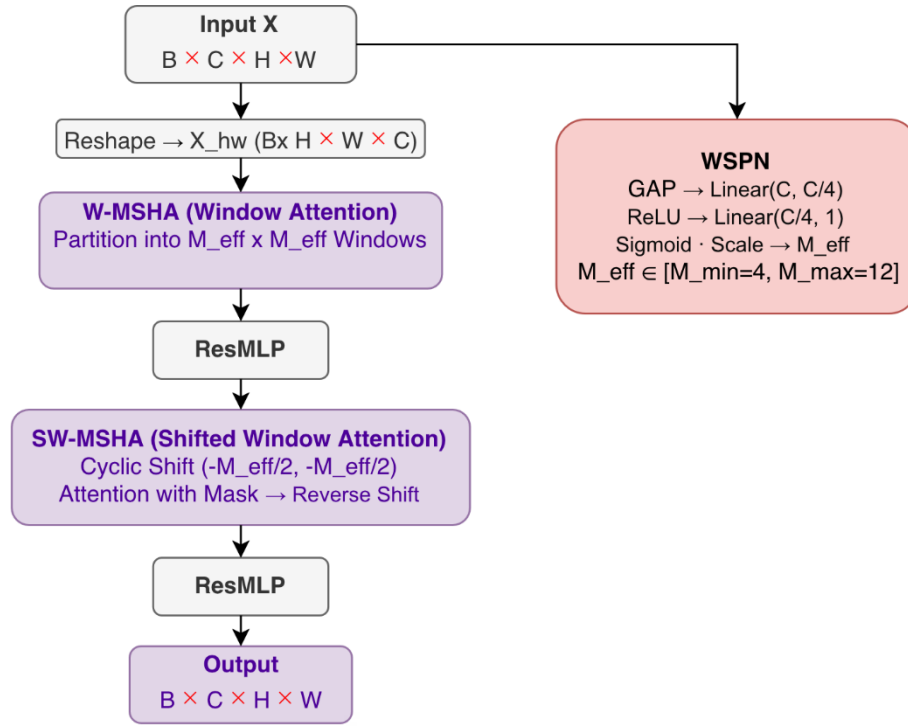


Figure 5. Adaptive Window Swin Transformer block with Window-Size Prediction Network

Abbreviations: GAP: Global average pooling; M_{eff}: Effective window size; ReLU: Rectified linear unit; ResMLP: Residual multilayer perceptron; SW-MHSA: Shifted-window multihead self-attention; W-MHSA: Window-based multihead self-attention; WSPN: Window-Size Prediction Network.

$$\widehat{TI}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(T^l\right)\right) + T^l \quad (13)$$

$$TI^{l+1} = \text{ResMLP}\left(\text{LN}\left(\widehat{TI}^{l+1}\right)\right) + \widehat{z}^{l+1} \quad (14)$$

The multi-head self-attention within each window of $(M_{\text{eff}})^2$ tokens is computed as in **Equation 15**:

$$\text{Attention}(Q', K', V') = \text{softmax}\left(\frac{Q'K'^T}{\sqrt{d_k}} + B\right)V' \quad (15)$$

where $Q' = X_{\text{norm}} \cdot W^{Q'}$, $K' = X_{\text{norm}} \cdot W^{K'}$, $V' = X_{\text{norm}} \cdot W^{V'}$ are the query, key, and value projections; d_k is the key dimension; and B is the relative position bias table.²³ The softmax ensures attention weights sum to 1 across all tokens in the window. The adaptive M_{eff} from the WSPN determines the window partition, enabling the model to implicitly select its attention scope based on input content.

For the SW-MHSA sub-layer, a cyclic shift of $(-M_{\text{eff}}/2, -M_{\text{eff}}/2)$ is applied to the feature map before window partitioning, and the corresponding attention mask is computed to prevent cross-region attention. After attention, the reverse cyclic shift $(+M_{\text{eff}}/2, +M_{\text{eff}}/2)$: restores

spatial alignment. Padding is applied if H or W is not divisible by M_{eff} , and removed after attention.

3.8. Residual multilayer perceptron

The ResMLP component, adopted from Ref.¹, replaces the standard multilayer perceptron in Swin Transformer blocks with a residual-augmented design. Given input $Z_{\text{in}} \in \mathbb{R}^{N \times C}$ (where $N = H \times W$ token sequence length) as described in **Equations 16, 17, and 18**:

$$U = \text{GELU}(Z_{\text{in}} \cdot W_1 + b_1) \quad (16)$$

$$V = UW_2 + b_2 \quad (17)$$

$$Z_{\text{out}} = Z_{\text{in}} + V \quad (18)$$

where $W_1 \in \mathbb{R}^{C \times 4C}$ & " $W_2 \in \mathbb{R}^{4C \times C}$ " are learnable weight matrices. The residual connection $Z_{\text{out}} = Z_{\text{in}} + V$ ensures gradient stability in deep layers and prevents feature degradation, which is particularly important for small datasets like Kather-5K. As demonstrated in the ablation study of Ref.¹, replacing a standard multilayer perceptron with ResMLP improves accuracy from 99.66% to 99.96% on NCT-CRC-HE-100K.

3.9. Gated Cross-Pathway Fusion module

The GCPF module is introduced at the Stage 2–Stage 3 boundary to merge the local feature stream $Z_L \in \mathbb{R}^{B \times C_2 \times H_2 \times W_2}$ (from AA-INeXt Stage 2) with a global feature stream $Z_G \in \mathbb{R}^{B \times C_2 \times H_2 \times W_2}$ (from a preliminary AW-SwinT pass at the same resolution) is illustrated in Figure 6. The key insight is that simple element-wise addition, as used in the baseline, weights both streams equally, which is suboptimal when one stream is substantially more informative for a given spatial region.

The GCPF module computes sigmoid-gated spatial weight maps as illustrated in Equations 19, 20, and 21.

$$G_L = \sigma(\text{convolution}_{1 \times 1}(F_L)) \quad (19)$$

$$G_G = \sigma(\text{conv}_{\text{olution}1 \times 1}(F_G)) \quad (20)$$

$$F_{\text{fused}} = G_L \otimes F_L + G_G \otimes F_G \quad (21)$$

where $G_L, G_G \in \mathbb{R}^{B \times C_2 \times H_2 \times W_2}$ are the gating maps, $\sigma(\cdot)$ is the sigmoid function, and $\otimes$ denotes element-wise (Hadamard) multiplication. Unlike Softmax-based gating, the proposed GCPF module employs independent sigmoid gates, allowing the local and global feature streams to be emphasized or suppressed independently according to the characteristics of the input feature maps. Consequently, the gate values are not constrained to sum to one, enabling the network to simultaneously preserve both local morphological information and global contextual

information when both streams provide complementary evidence, while suppressing noisy or less informative responses from either pathway when necessary. Therefore, the proposed GCPF module performs adaptive feature recalibration by independently learning the relative importance of the local and global feature streams rather than enforcing a competitive relationship between them. The gating parameters are learned end-to-end through backpropagation. Finally, a 1×1 convolution followed by Batch Normalization is applied to the fused feature map to facilitate channel interaction and stabilize feature learning, as expressed in Equation 22:

$$F_{\text{fused}} = \text{BN}(\text{Conv}_{1 \times 1}(F_{\text{sum}})) \quad (22)$$

3.10. Layer normalization and downsampling

Between stages, patch embedding with strided convolution is applied for spatial downsampling as in Equation 23:

$$\text{PatchEmbed}(X) = \text{GELU}(\text{BN}(\text{Conv}_{3 \times 3, s=2}(X))) \quad (23)$$

Layer normalization (LN) is applied in all attention and multilayer perceptron sub-layers over the channel dimension, providing training stability in deep networks. For a feature vector $x \in \mathbb{R}^C$ as in Equation 24:

$$\text{LN}(x) = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \quad (24)$$

where μ and σ are the mean and standard deviation of x

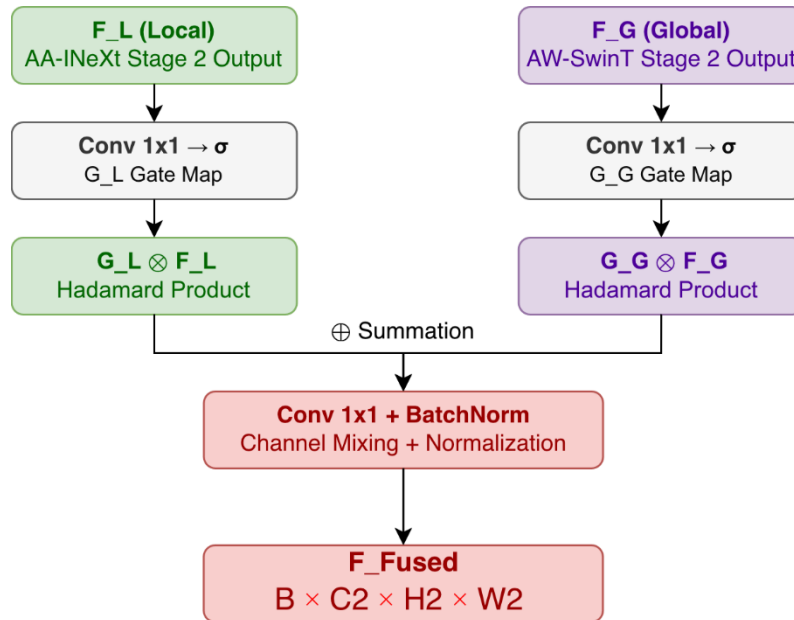


Figure 6. Gated Cross-Pathway Fusion module

across channels, $\varepsilon = 10^{-5}$ is a numerical stability constant, and $\gamma, \beta \in \mathbb{R}^C$ are learnable scale and shift parameters.

3.10. Classification

Following Stage 4, global average pooling aggregates $7 \times 7 \times 768$ feature volumes to create a 768-dimensional vector. A ResMLP block provides final deep non-linear interaction. According to Equation 25, the final linear classifier is normalized before applying the layer normalization:

$$\hat{y} = \text{Softmax}\left(W \cdot \text{ResMLP}\left(\text{LN}\left(\text{GAP}\left(X_{\text{Stage4}}\right)\right)\right) + b\right) \quad (25)$$

where $W \in \mathbb{R}^{N_c \times 768}$ and $b \in \mathbb{R}^{N_c}$ are the classifier weight and bias. The Softmax output $\hat{y} \in \mathbb{R}^{N_c}$ provides the predicted class probability distribution.

4. Results and discussion

4.1. Training protocol

The proposed DualPath-CRC model was evaluated using five-fold stratified cross-validation. Input histopathology images were preprocessed using the EfficientNetB0 preprocess_input() function. Only the training images were augmented using random horizontal and vertical flips, rotation, zoom, and contrast adjustment. The factors for rotation, zoom, and contrast adjustment were 0.15, 0.10, and 0.10, respectively. The network was initialized with ImageNet weights and trained using the Adam optimizer. To improve reproducibility, a fixed random seed was used during training. Training used a batch size of 32 and a TensorFlow tf.data pipeline for data loading. The reported performance values are the means and standard deviations obtained from fivefold cross-validation.

4.2. Performance on the NCT-CRC-HE-100K dataset using cross-validation

Table 1 presents the performance of the proposed model across the five cross-validation folds. The model showed

consistently high performance across all five folds with a mean accuracy of $99.10\% \pm 0.15\%$, macro F1-score of $99.08\% \pm 0.14\%$, macro precision of $99.05\% \pm 0.16\%$, and macro recall of $99.11\% \pm 0.12\%$. The model also achieved a high mean area under the receiver operating characteristic curve (AUC; 99.00%), demonstrating good discriminative ability. The Matthews correlation coefficient and Cohen's kappa values (both $98.98\% \pm 0.16\%$) indicated good overall classification performance and agreement beyond chance. The small standard deviations across the five folds further indicate consistent performance.

The confusion matrix in Figure 7 and the receiver operating characteristic (ROC) curve in Figure 8 illustrate the classification performance of the proposed model across the nine tissue categories (ADI, BACK, DEB, LYM, MUC, MUS, NORM, STR, and TUM). The confusion matrix shows that most samples were correctly classified along the main diagonal and that only a few were misclassified as visually similar tissue classes, indicating high class-wise discriminative capability. The ROC curves in Figure 8 plot the true-positive rate against the false-positive rate at different decision thresholds and show the model's ability to distinguish each tissue class from the remaining classes. Curves closer to the upper-left corner and higher AUC values indicate better discrimination. Overall, the concentration of predictions along the diagonal of the confusion matrix and the high AUC values confirm that the proposed model can accurately classify the nine colorectal histological tissue categories. However, the relatively greater confusion among the STR, MUS, and TUM classes indicates that these morphologically similar tissue classes remain more difficult to distinguish.

Table 2 shows that DualPath-CRC outperformed the evaluated CNN-based and Transformer-based models in terms of accuracy, positive prediction value, sensitivity, and F-measure.

Table 1. K-fold cross-validation

Metric	Mean $\pm$ Std	Fold::1	Fold::2	Fold::3	Fold::4	Fold::5
Accuracy	0.9910 $\pm$ 0.0015	0.9900	0.9910	0.9890	0.9905	0.9909
Macro F1-score	0.9908 $\pm$ 0.0014	0.9898	0.9898	0.9900	0.9908	0.9888
Macro precision	0.9905 $\pm$ 0.0016	0.9895	0.9892	0.9885	0.9897	0.9900
Macro recall	0.9911 $\pm$ 0.0012	0.9891	0.9900	0.9903	0.9910	0.9909
AUC (one-vs-rest)	0.9900 $\pm$ 0.0001	0.9901	0.9988	0.9987	0.9900	0.9989
MCC	0.9898 $\pm$ 0.0016	0.9884	0.9888	0.9890	0.9885	0.9893
Cohen's kappa	0.9898 $\pm$ 0.0016	0.9873	0.9898	0.9895	0.9890	0.9895

Abbreviations: AUC: Area under the receiver operating characteristic curve; MCC: Matthews correlation coefficient; Std: Standard deviation.

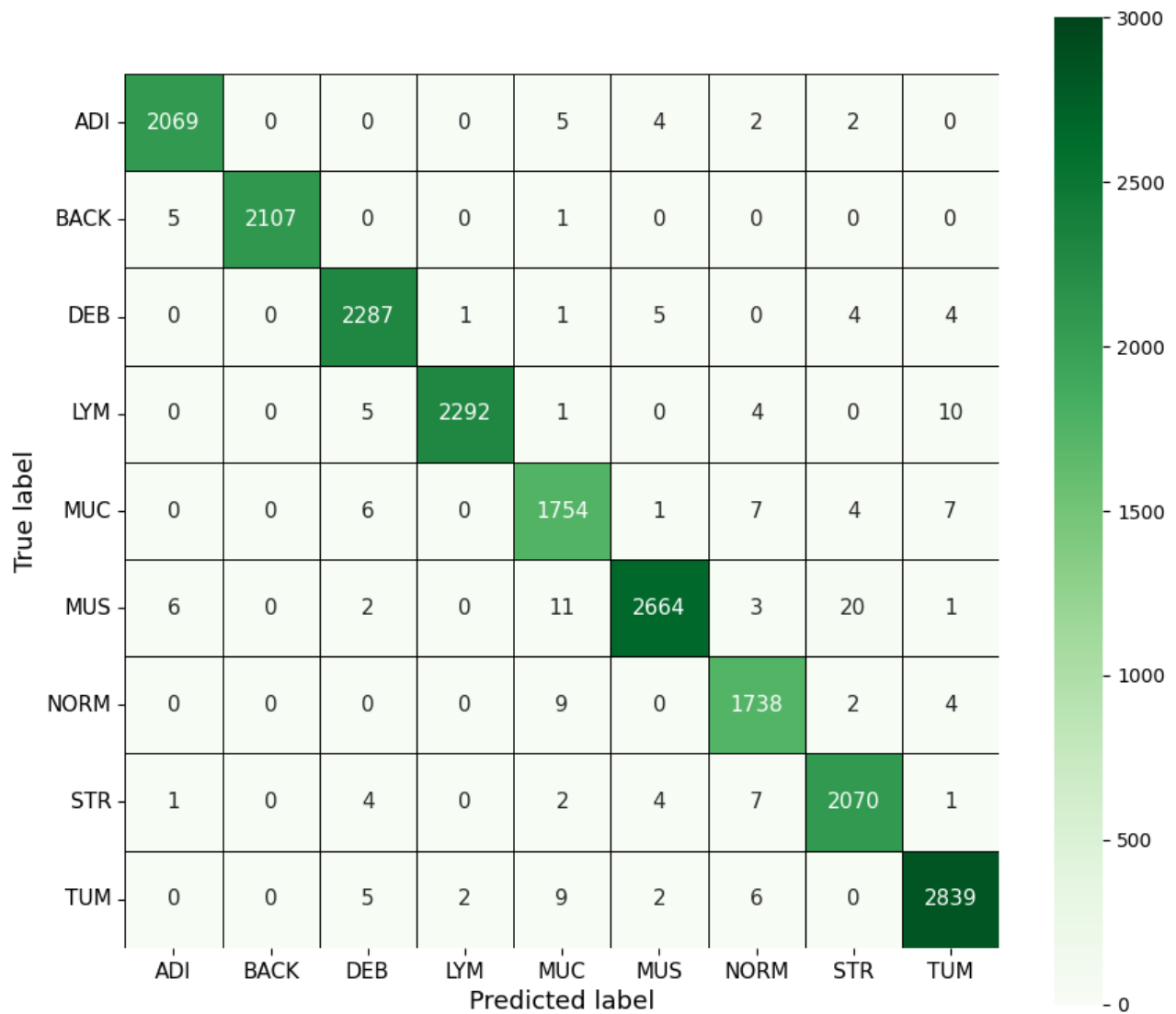


Figure 7. Confusion matrix of the proposed model

Abbreviations: ADI: Adipose tissue; BACK: Background; DEB: Debris; LYM: Lymphocytes; MUC: Mucus; MUS: Smooth muscle; NORM: Normal colon mucosa; STR: Cancer-associated stroma; TUM: Colorectal adenocarcinoma epithelium.

Table 2. Comparative performance on the NCT-CRC-HE-100K dataset using the proposed DualPath CRC

Method	Overall performance	Positive prediction value	Sensitivity	F-measure
ResNet50	0.97	0.96	0.95	0.96
DenseNet121	0.96	0.95	0.94	0.95
EfficientNetv2	0.97	0.97	0.96	0.96
Swin-Base	0.96	0.95	0.94	0.94
InceptionNeXt-Swin Transformer	0.98	0.97	0.98	0.97
DualPath-CRC (proposed)	0.99	0.98	0.98	0.98

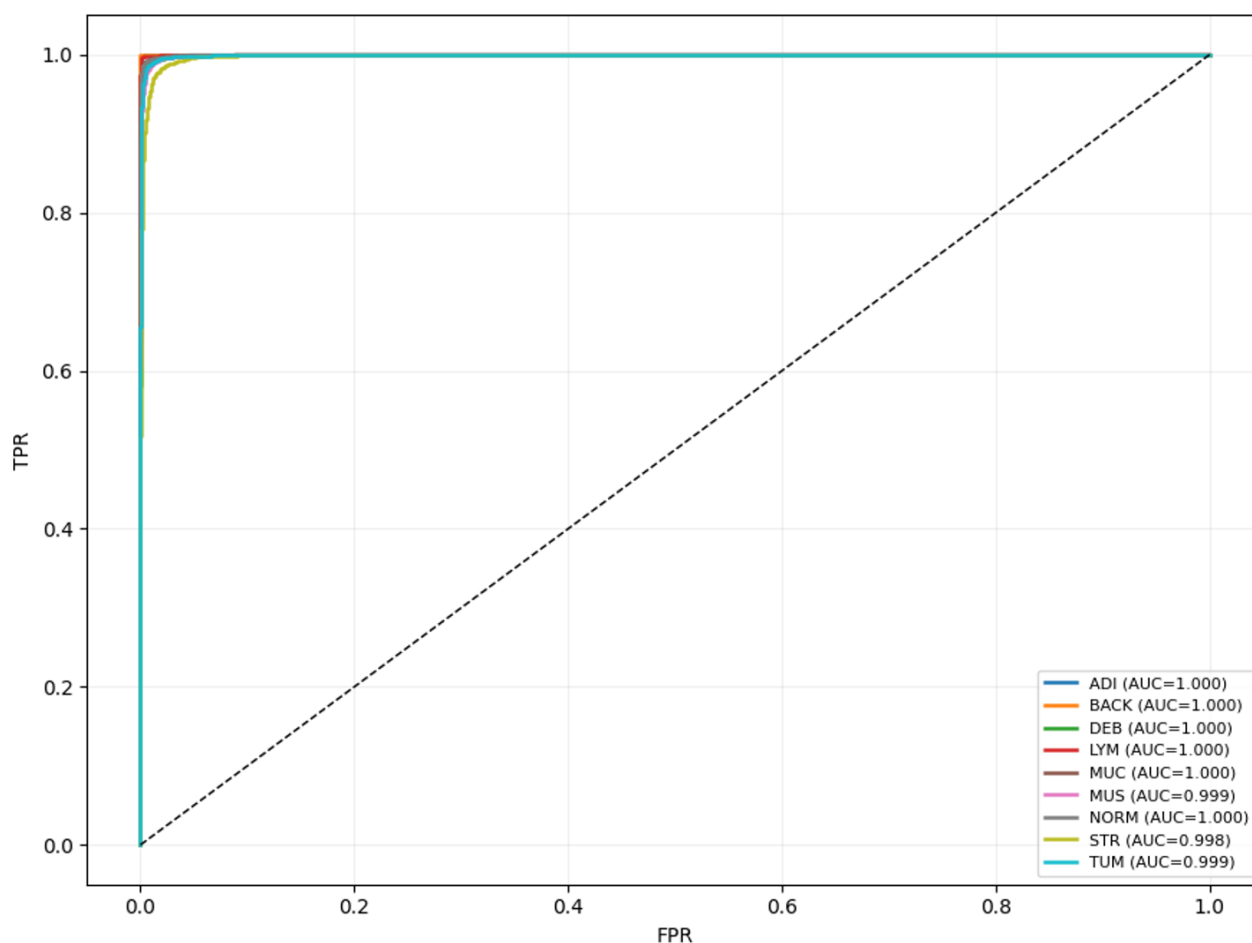


Figure 8. Multiclass ROC curves for the best-performing fold (Fold 5) using the one-vs-rest approach

Abbreviations: ADI: Adipose tissue; AUC: Area under the receiver operating characteristic curve; BACK: Background; DEB: Debris; FPR: False-positive rate; LYM: Lymphocytes; MUC: Mucus; MUS: Smooth muscle; NORM: Normal colon mucosa; OVR: One-vs-rest; ROC: Receiver operating characteristic; STR: Cancer-associated stroma; TPR: True-positive rate; TUM: Colorectal adenocarcinoma epithelium.

4.3. Performance on the Kather-5K dataset using cross-validation

Table 3 presents the classification performance of the proposed model based on fivefold cross-validation. The results indicate high and consistent performance across the folds. There was little variation across the folds, and the mean macro F1-score (0.9795) and macro recall (0.9976) were high. The high mean macro precision (0.9898 ± 0.0027) also suggests that a large proportion of the predicted positive cases were correctly classified. Fold 3 achieved the highest accuracy of 98.60%, whereas Fold 1 had the lowest accuracy of 97.70%. Overall, the small standard deviations indicate stable performance across the folds and support the robustness of the model. The class-wise confusion matrix for the Kather-5K dataset is illustrated in Figure 9.

On the Kather-5K dataset, DualPath-CRC achieved an accuracy of 99.34% (Table 4), which was 0.28 percentage points higher than that of the next-best model, the InceptionNeXt-Swin Transformer.

4.4. Ablation Study

Table 5 presents the results of the ablation study, in which the proposed modules were evaluated individually and in combination. The baseline InceptionNeXt-Swin Transformer achieved an accuracy of 99.96%. Adding AA-INeXt alone increased the accuracy to 99.97%, representing an improvement of 0.01 percentage points over the baseline. However, adding AW-SwinT alone did not improve the baseline accuracy (99.96%) and thus could not be observed when it was introduced independently on the NCT-CRC-HE-100K dataset. The combination of

Table 3. Five-fold cross-validation on the Kather-5K dataset

Metric	Mean $\pm$ Std	Fold::1	Fold:: 2	Fold:: 3	Fold::4	Fold:: 5
Accuracy	98.84% $\pm$ 0.27%	97.70%	98.30%	98.60%	98.00%	98.10%
Macro F1-score	0.9795 $\pm$ 0.0027	0.9769	0.9731	0.9660	0.9701	0.9712
Macro precision	0.9898 $\pm$ 0.0027	0.9670	0.9734	0.9765	0.9703	0.9717
Macro recall	0.9794 $\pm$ 0.0027	0.9770	0.9730	0.9660	0.9700	0.9710

Table 4. Comparative performance of DualPath-CRC and other models on the Kather-5K dataset

Method	Overall performance	Positive prediction value	Sensitivity	<i>F</i> -measure
ResNet50	0.9651	0.9653	0.9651	0.9651
ViT-Patch16-Base	0.9825	0.9827	0.9825	0.9826
DeiT3-Base	0.9839	0.9842	0.9839	0.9840
CCDNet	0.9896	0.9937	0.9880	0.9864
InceptionNeXt-Swin Transformer	0.9906	0.9907	0.9906	0.9908
DualPath-CRC (Proposed)	0.9934	0.9936	0.9934	0.9934

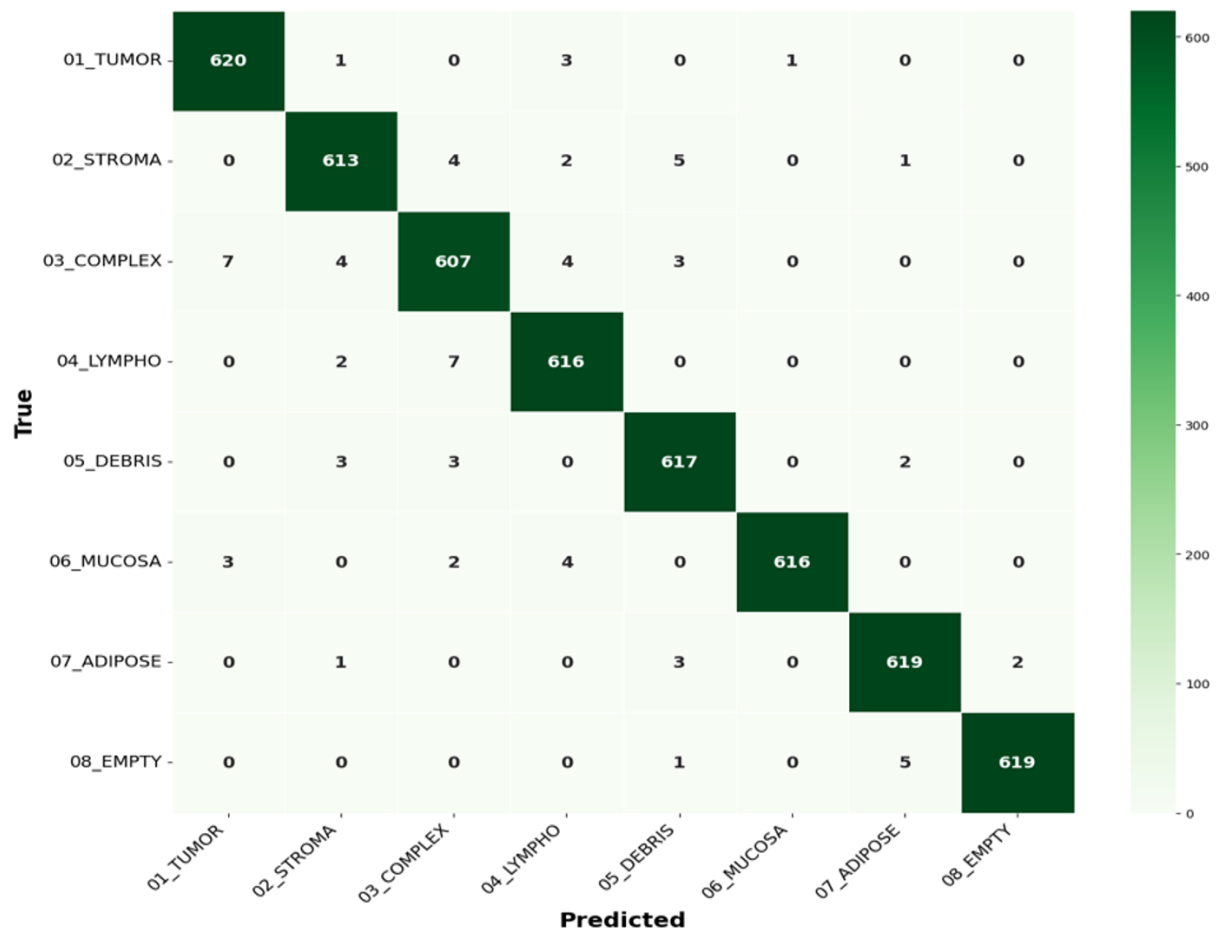


Figure 9. Class-wise confusion matrix for the Kather-5K dataset

Table 5. Ablation study on the NCT-CRC-HE-100K dataset

Configuration	AA-INeXt	AW-SwinT	GCPF	Accuracy (%)
InceptionNeXt-Swin Transformer Baseline	X	X	X	99.96
+ AA-INeXt only	✓	X	X	99.97
+ AW-SwinT only	X	✓	X	99.96
+ AA-INeXt + AW-SwinT	✓	✓	X	99.97
DualPath-CRC (proposed)	✓	✓	✓	99.98

Abbreviations: AA-INeXt: Attention-Augmented InceptionNeXt; AW-SwinT: Adaptive-Window Swin Transformer; GCPF: Gated Cross-Pathway Fusion.

AA-INeXt and AW-SwinT also achieved an accuracy of 99.97%, the same as that achieved using AA-INeXt alone, suggesting that AW-SwinT did not provide an additional improvement in accuracy under this experimental setting. Finally, incorporating the GCPF module increased the accuracy by a further 0.01 percentage point, resulting in the highest accuracy of 99.98% for the complete DualPath-CRC model. These results suggest that the observed improvement in accuracy was primarily associated with the addition of the AA-INeXt and GCPF modules.

4.5. Grad-CAM interpretability

To gain deeper insights into the interpretability of the proposed DualPath-CRC model and qualitatively evaluate whether its predictions were based on histopathologically relevant tissue structures, Grad-CAM⁶, Grad-CAM++, and LIME analyses were performed. The Grad-CAM visualizations in Figure 10 provide qualitative evidence that the DualPath-CRC model focused on regions corresponding to relevant histopathological features. In the colorectal adenocarcinoma epithelium category, regions of high activation, shown in red, appeared to correspond to distorted glands, cribriform morphology, and other abnormal epithelial structures. For the lymphocyte category, high activation was mainly observed around regions of lymphocytic infiltration. For the cancer-associated stroma category, activation was concentrated in fibrous tissue, collagen-rich regions, and other stromal structures. The Grad-CAM++ visualizations produced more localized and fine-grained activation maps, especially when multiple discriminative regions were present within the same tissue patch. Compared with standard Grad-CAM, Grad-CAM++ produced more localized activation patterns around relevant cellular and tissue structures, providing additional qualitative information about the model's responses to visually similar regions and complex tissue patterns. The LIME explanations supported these observations by identifying superpixels that contributed to the model's predictions. The highlighted superpixels were primarily located within tissue regions rather than the background and appear to correspond to morphologically

relevant features, suggesting that the model's predictions were associated with meaningful morphological regions. These visualizations provide qualitative interpretability evidence only and do not independently establish the correctness of the classifications or fully explain the basis of the model's predictions.

The attention maps generated by DualPath-CRC showed focused activation at challenging stroma-tumor boundaries. This finding indicates that the GCPF module helps distinguish subtle morphological differences between visually similar tissue categories. The broadly consistent regions highlighted by Grad-CAM, Grad-CAM++, and LIME provide complementary qualitative evidence that the DualPath-CRC predictions were associated with relevant histopathological features, supporting the interpretability and plausibility of the proposed framework.

4.6. Computational complexity

Table 6 compares the computational complexity, inference latency, and classification accuracy of the evaluated models. ResNet50, Swin-Base, and InceptionNeXt-Swin were selected as baseline models for comparison. ResNet50 had an accuracy of 99.63% with 25.6 M parameters and an inference latency of 2.1 ms, while the InceptionNeXt-Swin model had an accuracy of 99.96% with 26.4 M parameters and an inference latency of 3.0 ms. The proposed DualPath-CRC achieved the highest classification accuracy of 99.98%, with 27.4 million trainable parameters, 4.61 giga floating-point operations, and an inference latency of 3.1 ms, which showed good prediction performance and efficiency. The computational benchmarks were performed on the Kaggle Notebook platform using an NVIDIA Tesla T4 GPU, Python 3.12.13, and TensorFlow (version 2.20) in precision mode. A batch size of 1 was used for inference. Five warm-up iterations were performed before the average inference latency was measured over three timed runs. Input data were loaded into memory, and data-loading and disk input/output operations were excluded from the latency measurements so that the reported values reflected model inference time only. These benchmarks

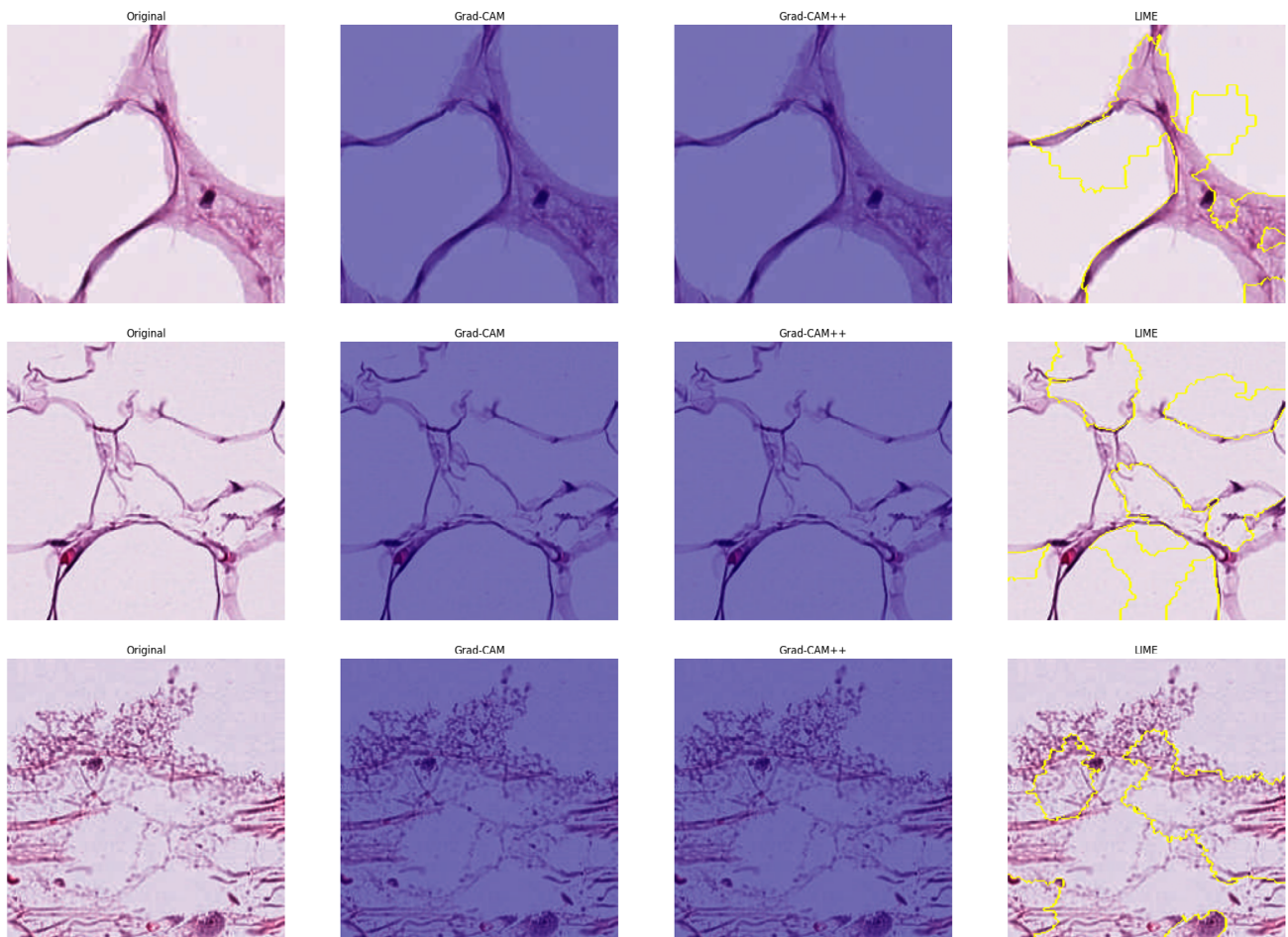


Figure 10. Explainable artificial intelligence analysis of the proposed model

Table 6. Comparison of computational complexity among the baseline architectures and the proposed model

Model	Parameters (M)	FLOPs (G)	Latency (ms)	Accuracy (%)
ResNet50	25.6	4.1	2.1 ± 0.1	99.63
Swin-Base	87.8	15.4	8.2 ± 0.3	99.91
InceptionNeXt-Swin	26.4	4.36	3.0 ± 0.2	99.96
DualPath-CRC (proposed)	27.4	4.61	3.1 ± 0.1	99.98

Abbreviation: FLOPs: Floating-point operations.

were obtained under controlled experimental conditions and should not be interpreted as evidence of clinical utility or readiness for real-world deployment.

4.7. Limitations

Although DualPath-CRC showed potential, several limitations should be acknowledged. First, the near-ceiling performance observed on benchmark datasets such as

NCT-CRC-HE-100K leaves little room for numerical improvement; therefore, small increases in accuracy should be interpreted with caution. Second, the current assessment was performed at the patch level rather than at the WSI or patient level. Additionally, patch-level partitioning may result in slide- or patient-level data leakage if patches from the same source slide or patient are assigned to different subsets. Therefore, slide-disjoint

or patient-disjoint evaluation is necessary to ensure the robustness of the model. Third, inter-site domain shifts resulting from differences in staining protocols, scanners, institutional imaging practices, and tissue preparation may affect model generalization across clinical centers. The absence of external institutional validation further limits conclusions regarding real-world clinical applicability. While Grad-CAM, Grad-CAM++, and LIME offer valuable qualitative insights into regions associated with model predictions, these interpretations do not definitively establish the correctness of the classifications or the basis of the predictions. They would benefit from systematic evaluation by domain experts or quantitative localization methods. Finally, near-perfect performance on curated benchmark datasets does not necessarily demonstrate patient-level clinical effectiveness. To assess the robustness and clinical utility of DualPath-CRC, future studies should consider patient- or slide-disjoint data partitions, multi-institutional external cohorts, WSI-level analysis, diverse staining and scanning conditions, and prospective clinical validation.

5. Conclusion

This paper introduces a novel dual-pathway deep learning architecture, termed DualPath-CRC, for the automated classification of CRC histopathology images. The proposed architecture includes three components: (i) AA-INeXt blocks with a channel attention mechanism; (ii) AW-SwinT blocks with WSPN; and (iii) a GCPF module for integrating local and global feature representations. DualPath-CRC achieved accuracies of 99.98% on NCT-CRC-HE-100K and 99.34% on Kather-5K under the reported evaluation protocols. The ablation study evaluated the individual and combined effects of the proposed architectural components. In addition, the Grad-CAM, Grad-CAM++, and LIME analyses indicated that the model focused on histopathologically relevant regions and provided qualitative evidence supporting its interpretability. Future directions include applying knowledge distillation to develop lightweight model versions suitable for edge devices, aggregating patch-level predictions for WSI-level analysis, adapting the model for tumor grading, staging, and treatment-response prediction, and using self-supervised pre-training on large, unannotated histopathology datasets.

Acknowledgments

None.

Funding

This research received no external funding. The authors did not receive any financial support, grants, or other funding

from public, commercial, or not-for-profit organizations for the conduct of this research or for the publication of this article.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Uddagiri Sirisha

Formal analysis: Panguluri Padmavathi, D. Jyothi

Investigation: D. Jyothi, Kannaiah Chattu

Methodology: Uddagiri Sirisha, Appalaraju Grandhi

Writing-original draft: Uddagiri Sirisha, Panguluri Padmavathi

Writing-review & editing: Uddagiri Sirisha, Kannaiah Chattu

Ethics approval and consent to participate

Not applicable. This study used publicly available colorectal cancer datasets and did not involve human participants, patient recruitment, or identifiable personal data.

Consent for publication

Not applicable.

Availability of data

The datasets used in this study are publicly available and can be obtained from their respective public repositories. The dataset sources are cited in the manuscript, and all data used for training, validation, and testing are available from the corresponding dataset providers.

References

1. Pacal I, Attallah O. Hybrid deep learning model for automated colorectal cancer detection using local and global feature extraction. *Knowledge-Based Syst.* 2025;319:113625. doi: 10.1016/j.knosys.2025.113625
2. 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. *CA Cancer J Clin.* 2024;74(3):229-263. doi: 10.3322/caac.21834
3. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A. Colorectal cancer statistics, 2023. *CA Cancer J Clin.* 2023;73(3):233-254. doi: 10.3322/caac.21772
4. Kather JN, Weis CA, Bianconi F, *et al.* Multi-class texture analysis in colorectal cancer histology. *Sci Rep.* 2016;6:27988. doi: 10.1038/srep27988

5. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proc IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR)*. 2018:7132-7141.
doi: 10.1109/CVPR.2018.00745
6. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE; 2017:618-626.
doi: 10.1109/iccv.2017.74
7. Kather JN, Krisam J, Charoentong P, *et al.* Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med*. 2019;16(1):e1002730.
doi: 10.1371/journal.pmed.1002730
8. Shafi ASM, Molla MMI, Jui JJ, Rahman MM. Detection of colon cancer based on microarray dataset using machine learning. *SN Appl Sci*. 2020;2(7):1243.
doi: 10.1007/s42452-020-3051-2
9. Alqudah AM, Alqudah A. Improving machine learning recognition of colorectal cancer using 3D GLCM applied to different color spaces. *Multimed Tools Appl*. 2022;81(8):10839-10860.
doi: 10.1007/s11042-022-11946-9
10. Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput*. 2017;29(9):2352-2449.
doi: 10.1162/neco_a_00990
11. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. *Evol Intell*. 2022;15(1):1-22.
doi: 10.1007/s12065-020-00540-3
12. Attallah O. Skin-CAD: explainable deep learning classification of skin cancer from dermoscopic images. *Comput Biol Med*. 2024;178:108798.
doi: 10.1016/j.compbimed.2024.108798
13. Attallah O. CerCan-Net: cervical cancer classification model via multi-layer feature ensembles of lightweight CNNs and transfer learning. *Expert Syst Appl*. 2023;229:120624.
doi: 10.1016/j.eswa.2023.120624
14. Attallah O. Lung and colon cancer classification using multiscale deep features integration of compact CNNs. *Technologies*. 2025;13(2):54.
doi: 10.3390/technologies13020054
15. Attallah O, Ragab DA. Auto-MyIn: automatic diagnosis of myocardial infarction via multiple GLCMs, CNNs, and SVMs. *Biomed Signal Process Control*. 2023;80:104273.
doi: 10.1016/j.bspc.2022.104273
16. Attallah O. Acute lymphocytic leukemia detection and subtype classification via extended wavelet pooling based-CNNs. *Image Vis Comput*. 2024;147:105064.
doi: 10.1016/j.imavis.2024.105064
17. Attallah O. GabROP: Gabor wavelets-based CAD for retinopathy of prematurity diagnosis via CNNs. *Diagnostics*. 2023;13(2):171.
doi: 10.3390/diagnostics13020171
18. İnce S, Kunduracioglu I, Bayram B, Pacal I. U-Net-based models for precise brain stroke segmentation. *Chaos Theory Appl*. 2025;7(1):50-60.
doi: 10.51537/chaos.1605529
19. Paladini E, Vantaggiato E, Bougourzi F, *et al.* Two ensemble-CNN approaches for colorectal cancer tissue type classification. *J Imaging*. 2021;7(3):51.
doi: 10.3390/jimaging7030051
20. Zhou P, Cao Y, Li M, *et al.* HCCANet: histopathological image grading of colorectal cancer using CNN. *Sci Rep*. 2022;12(1):15103.
doi: 10.1038/s41598-022-18879-1
21. Anju TE, Vimala S. Finetuned-VGG16 CNN model for tissue classification of colorectal cancer. In: Raj JS, *et al.*, eds. *Intelligent Sustainable Systems*. Lecture Notes in Networks and Systems. Vol 665. Springer; 2023:73-84.
doi: 10.1007/978-981-99-1726-6_7
22. Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. Preprint posted online 2020.
doi: 10.48550/arXiv.2010.11929
23. Liu Z, Lin Y, Cao Y, *et al.* Swin Transformer: hierarchical vision transformer using shifted windows. In: *Proc IEEE/CVF Int Conf Comput Vis (ICCV)*. 2021:9992-10002.
doi: 10.1109/ICCV48922.2021.00986
24. Zeid MAE, El-Bahnasy K, Abo-Youssef SE. Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers. In: *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE; 2021:224-230.
doi: 10.1109/icicis52592.2021.9694125
25. Ding M, Qu A, Zhong H, *et al.* An enhanced vision transformer with wavelet position embedding. *Pattern Recognit*. 2023;140:109532.
doi: 10.1016/j.patcog.2023.109532
26. Li M. Transformer-Based Self-Supervised Learning and Distillation for Medical Image Classification: Improving Colorectal Cancer Detection on NCT-CRC-HE-100K with Swin-T V2. In: *2024 3rd International Conference on Cloud Computing, Big Data Application and Software Engineering*

- (CBASE). IEEE; 2024:644-648.
doi: 10.1109/cbase64041.2024.10824558
27. Reis HC, Turk V. Fusion of transformer attention and CNN features for skin cancer detection. *Appl Soft Comput.* 2024;164:112013.
doi: 10.1016/j.asoc.2024.112013
28. Sriwastawa A, Arul Jothi JA. Vision transformer and its variants for image classification in digital breast cancer histopathology. *Multimed Tools Appl.* 2023;83(13):39731-39753.
doi: 10.1007/s11042-023-16954-x
29. Martínez-Fernandez E, Rojas-Valenzuela I, Valenzuela O, Rojas I. Computer aided classifier of colorectal cancer on histopathological whole slide images analyzing deep learning architecture parameters. *Appl Sci.* 2023;13(7):4594.
doi: 10.3390/app13074594
30. Ghosh S, Bandyopadhyay A, Sahay S, et al. Colorectal histology tumor detection using ensemble deep neural network. *Eng Appl Artif Intell.* 2021;100:104202.
doi: 10.1016/j.engappai.2021.104202
31. Abd El-Ghany S, Mahmood MA, Abd El-Aziz AA. Adaptive Dynamic Learning Rate Optimization Technique for Colorectal Cancer Diagnosis Based on Histopathological Image Using EfficientNet-B0 Deep Learning Model. *Electronics.* 2024;13(16):3126.
doi: 10.3390/electronics13163126
32. Sharkas M, Attallah O. Color-CADx: a deep learning approach for colorectal cancer classification through triple convolutional neural networks and discrete cosine transform. *Sci Rep.* 2024;14(1).
doi: 10.1038/s41598-024-56820-w
33. Pan XL, Hua B, Tong K, et al. EL-CNN: an enhanced lightweight classification for colorectal cancer histopathological images. *Biomed Signal Process Control.* 2025;100:106933.
doi: 10.1016/j.bspc.2024.106933
34. Khalid SVM, Deivasigamani S, Rajendran S. An efficient colorectal cancer detection network using atrous convolution with coordinate attention transformer. *Sci Rep.* 2024;14(1):19109.
doi: 10.1038/s41598-024-70117-y
35. Chen H, Li C, Li X, et al. IL-MCAM: interactive learning and multi-channel attention for colorectal histopathology classification. *Comput Biol Med.* 2022;143:105265.
doi: 10.1016/j.compbiomed.2022.105265
36. Tanveer M, Akram MU, Khan AM. TransNetV: an optimized hybrid model for enhanced colorectal cancer image classification. *Biomed Signal Process Control.* 2024;96:106579.
doi: 10.1016/j.bspc.2024.106579
37. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: *Proc Eur Conf Comput Vis (ECCV).* 2018:3-19.
doi: 10.1007/978-3-030-01234-2_1
38. Cho K, van Merriënboer B, Gulcehre C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics; 2014:1724-1734.
doi: 10.3115/v1/d14-1179
39. Verma R, Kumar N, Patil A, et al. MoNuSAC2020: a multi-organ nuclei segmentation and classification challenge. *IEEE Trans Med Imaging.* 2021;40(12):3413-3423.
doi: 10.1109/TMI.2021.3085712
40. Kather JN, Halama N, Marx A. 100,000 Histological images of human colorectal cancer and healthy tissue. Published online April 7, 2018.
doi: 10.5281/ZENODO.1214456
41. Touvron H, Cord M, Jégou H. DeiT III: revenge of the ViT. In: Avidan S, Brostow G, Cissé M, et al, eds. *Computer Vision – ECCV 2022.* Cham: Springer; 2022:516-533.
doi: 10.1007/978-3-031-20053-3_30