Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Attentive Dual-Branch CNN–Transformer Fusion for Cross-Dataset Breast Histopathology Classification

Mradul Kumar Jain[1], Veerendra Yadav[2], Anu Chaudhary[3]

[1,2]Department of Computer Science and Engineering, Noida International University, Gautam Budh Nagar, Uttar Pradesh, India.
[3]Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India.

E-mail: [1]mradulanmol@gmail.com, [2]vyadav1@ce.iitr.ac.in, [3]getanuchaudhary@yahoo.com

Orcid ID: [1]0009-0003-2232-6474, [2]0000-0002-8679-132X, [3]0009-0006-5466-5405

Abstract

In general, breast histopathology classifiers exhibit significant performance degradation when applied to images acquired using different staining protocols, magnification rates, scanner equipment, or institutional environments. This study proposes an attentive dual-branch CNN–Transformer framework intended to learn representations that remain informative with respect to heterogeneity in histopathological data. The convolutional path captures fine cellular details, such as nuclear texture and local glandular structures, whereas the dilated convolution path captures more generalized tissue structures without extensive downscaling of spatial dimensions. Attention operations were applied separately to the two paths to diminish the impact of background areas and structures with low discriminative values. The generated feature maps were then represented in terms of spatial tokens, which were processed by a transformer encoder to represent the long-range dependencies between distant regions in the tissue structure. A learnable fusion gate regulates the contribution of the local and contextually informative parts of the input features for each image to the output. The institution-specific classification layers handle the distinct diagnostic classes in the BACH and BreaKHis datasets. The evaluation procedure included stratified testing on BACH and patient-disjoint partitioning on BreaKHis to exclude any potential data leakage at the subject level. For BACH, our model achieved an accuracy of 94.75%, a macro F1-score of 93.90%, and an area under the receiver operating characteristic curve of 96.35%, with 18.9 million trainable parameters and an average inference time of 16.1 ms per image. A component-wise ablation study, calibration evaluation, confidence interval estimation, and attention-driven visual explanations were conducted to investigate the contribution and reliability of the proposed fusion method. The resulting architecture represents a trade-off between prediction performance and computational costs but does not claim superior efficiency or applicability. Multi-institutional and prospective independent validation is necessary for further clinical application.

Keywords: Attention Mechanism, Breast Cancer, Deep Learning, Histopathology, Transformer.

DOI: https://doi.org/10.36548/jiip.2026.4.002
Received: 16.07.2026, received in revised form: 20.08.2026, accepted: 12.09.2026, published: 22.09.2026

1. Introduction

Despite significant progress in oncology, breast cancer continues to be a serious medical issue and is the most common cancer in women worldwide. In 2022, there were 2.3 million new cases of breast cancer and 670 thousand deaths due to this disease worldwide [1]. Histopathology of stained samples allows for the distinction between normal, benign, in situ carcinoma, invasive carcinoma, and various tumor subtypes based on the morphology of the tissue sections. However, histopathology is time-consuming and potentially susceptible to variations in the experience, criteria, preparation, and interpretation of pathologists [2]. Hence, computer-assisted classification of histopathological images has been explored as a tool for reproducible image analysis rather than expert evaluations.

Convolutional neural networks (CNNs) have demonstrated a high capability of learning features during the processing of histopathological images because their hierarchical filters extract nuclear morphology, cellular texture, gland architecture, and local tissue patterns from the images. However, the receptive field of a classical CNN is limited by locality, and the modeling of relations between remote parts of the tissue structure is performed using deep layers or pooling operations [3]. These operations may suppress the morphological details of tissues, which are crucial for separating visually similar diagnostic groups from one another. Vision transformers (ViTs) solve this problem by modeling the interactions between distant parts of an image via self-attention. However, ViT performance depends heavily on the amount and variety of training data, tokenization, normalization, and computational settings [4]. Therefore, a hybrid architecture of CNN and Transformer is relevant in situations where a combination of morphological sensitivity and context representation is required.

Another challenge is the variability among different histopathological datasets. Differences in staining intensity, preparation, scanner parameters, resolution, magnification, patient selection, and diagnostic criteria alter the distribution of visual features. The BACH microscopy dataset contains 400 high-resolution hematoxylin and eosin slides grouped into normal, benign, in situ carcinoma, and invasive carcinoma groups [5]. However, the BreaKHis dataset contains 7,909 images obtained from 82 patients at magnifications of 40×, 100×, 200×, and 400× with benign-malignant classification and eight histological subtypes [6]. Thus, high performance on one dataset cannot guarantee the effectiveness of the representation for different acquisition processes and diagnostic criteria. In particular, patient-level segregation is crucial for BreaKHis, since the images of one patient have biological and technical similarities, and the distribution of images of the same patient between the training and test partitions leads to information leakage.

Recently, various approaches, including attention-guided CNN, transformer encoders, multiscale feature extraction, and parallel CNN-transformer pathways, have been developed for breast histopathology [7]. Although these approaches improve feature representation, three problems remain. First, hybrid architectures fuse CNN and transformer outputs by concatenation or addition without estimating the relevance of local and contextual evidence for every image [8]. Second, some studies present results using a single dataset or image-level segregation, which provides little evidence of the effectiveness of the representation on different datasets and its ability to perform independent predictions [9, 10]. Finally, discussions on architectural complexity often lack analyses of component contributions, prediction calibration, statistical uncertainty, robustness and computational cost. These problems do not allow us to

determine whether the improvement is caused by the fusion technique or by an increase in model capacity.

To resolve these problems, this study developed an attentive dual-branch CNN-Transformer architecture for breast histopathology classification of the BACH and BreaKHis datasets. The local branch retains the fine morphological features of cellular and nuclear structures, while the dilated contextual branch expands the receptive field to represent larger tissue arrangements without severe spatial downscaling. Channel and spatial attention refined the features of each branch before converting them into spatial tokens. Subsequently, the transformer encoder learns the interactions among the distant parts of the tissue. Instead of assigning a constant contribution to the two representations, the learnable reliability gate evaluates the importance of the two branches and creates an adaptively fused representation [11]. The feature encoder is denoted by f_θ, where θ denotes the learnable parameters. The classifier for BACH is denoted by g_B, and the classifier for BreaKHis is denoted by g_H. The two dataset-specific classifiers fit the different labels without imposing an artificial connection between the four BACH and BreaKHis classes.

In this study, cross-dataset classification is understood as the evaluation of a unified architectural and feature-learning strategy on datasets with different image properties and diagnostic labels. However, this does not imply that the labels of BACH and BreaKHis are directly exchangeable. BACH is evaluated using a four-category diagnostic task, whereas BreaKHis is evaluated using patient-separate binary and subtype classification tasks. The concept of cross-dataset classification must be consistent throughout the manuscript to prevent misinterpretation of the multi-dataset independent evaluation as a direct zero-shot clinical transfer [12].

The methodological contributions of this study include four aspects. First, complementary local and dilated contextual branches with branch-specific attention allow for the preservation of fine morphologies and broader tissue structures. Second, sample-adaptive reliability gating allows for controlling the importance of two representations before contextual modeling via a transformer encoder. Third, the evaluation protocol included an image-level analysis of BACH and a patient-disjoint analysis of BreaKHis with a separate assessment of magnification. Finally, the experimental analysis covers the prediction performance, class balance, calibration, statistical uncertainty, architectural contribution, visual explanations and computational costs. Thus, the proposed approach is considered a trade-off between predictive performance and computational cost, but not the smallest, fastest, or clinically deployable model.

Convolutional neural networks offer an efficient inductive bias for modelling cellular texture, nuclear morphology, and local tissue organization, whereas transformer encoders capture non-local dependencies between spatially distant histopathological regions [13]. However, the complementary abilities of these models do not guarantee consistent performance under heterogeneous data. Moreover, the dataset shift problem should be considered because BACH and BreaKHis differ in terms of image resolution, acquisition strategy, magnification levels, demographics, sample distribution, and diagnostic-label taxonomy [14]. In particular, a model that demonstrates high quality on one dataset might fail to transfer its discrimination ability, class-wise ranking, and calibration to other datasets. Therefore, the main motivation of this study was to verify whether the explicitly formulated local-global feature fusion mechanism

preserves its efficacy on independent breast histopathology datasets.

The proposed solution substitutes feature concatenation with bidirectional cross-attention and sample-adaptive reliability gating of the local and global feature tokens. Local feature tokens are provided with contextual cues from the global pathway, whereas global feature tokens are enriched with fine-grained morphological information from the local path. Reliability coefficients α_L and α_G were computed for each input image individually, where α_L corresponds to the contribution of the local morphological pathway and α_G to the contribution of the global context pathway. It holds $\alpha_L \geq 0, \quad \alpha_G \geq 0, \quad \alpha_L + \alpha_G = 1$. Thus, the methodological novelty relies on sample-specific control of complementary feature representations rather than the separate use of convolution, attention, or transformer operations. The proposed framework was validated using a four-class classification on BACH and patient-disjoint binary and histological subtype classifications on BreaKHis. Additional components of the methodology include magnification-specific evaluation, ablation studies, uncertainty quantification, visualization and explanation, robustness testing and performance profiling. The cross-dataset results are reported independently for the BACH classification head g_B and the BreaKHis classification head g_H, where g_B projects the fused representation onto the four BACH diagnostic classes, and g_H projects it onto the native BreaKHis category space. The pooled accuracy metric was not provided because of differences in the label spaces and evaluation units.

The remainder of this paper is organized as follows. In Section 2, CNN-based, transformer-based, attention-guided, and hybrid approaches are briefly reviewed, and an unresolved research gap is formulated. In Section 3, the dual-branch architecture, attention operations, adaptive fusion mechanism, transformer encoder, dataset-specific classifiers, loss function, and training algorithm are described. In Section 4, the datasets, implementation settings, comparative experiments, patient-wise evaluation, ablation analysis, robustness analysis, visual explanations, and computational profiling are described. In Section 5, the principal conclusions are summarized, limitations are discussed, and directions for external and future validations are formulated.

2. Related Work

Deep learning research in breast histopathology has shifted from traditional convolutional representation extraction to attention-modulated transformer models and hybrid local/global representation learning approaches. According to Zhou et al. [15], convolutional neural networks (CNNs) can learn the hierarchy of representations of nuclei, cellular texture, glands, and tissue organization without the need for handcrafted descriptors. The earlier work of Rakhlin et al. [16] demonstrated a combination of pretrained CNN representations with traditional classifiers to classify images from the domain of breast histopathology. This approach highlights the utility of transfer learning when annotated histopathological datasets are limited. However, a representation learned purely with a convolution operation is mainly defined by local receptive fields and is thus limited in capturing interactions among spatially disjoint regions, depending on the depth of the network, pooling operations, and receptive field expansion.

2.1 CNN-Based and Attention-Guided Classification

Translation-equivariant filters provide an adequate inductive bias for modeling histological texture and cellular morphology. Therefore, residual networks, densely connected networks, and efficient convolutional backbones have been applied for binary and multi-class breast histopathology classification. However, continuous spatial downsampling can suppress differences in the morphological characteristics of benign tissues, carcinoma in situ, and invasive carcinoma. The use of multiscale convolutions and dilated filters partially solves this problem, enabling an increase in the receptive field without a proportional reduction in the resolution of feature maps.

Attention is used to refine the contributions of particular spatial locations and channels. In the study by Liu et al. [17], DALAResNet50 incorporated the attention mechanism into the ResNet50 architecture, and dynamic threshold Grad-CAM was used to generate a more focused visual explanation. It was evaluated on BreaKHis, BACH, and Mini-DDSM and demonstrated that attention refinement can help achieve better discriminability for varied image distributions. Nevertheless, an attention-augmented CNN primarily refines the convolutional features. In addition, it does not provide long-range token interactions, such as the self-attention of transformers, and a visually focused activation map does not indicate prediction calibration, causal interpretability, or clinical validity.

2.2 Transformer-Based Histopathology Analysis

Vision transformers (ViTs) split an image or an intermediate feature map into tokens and model their pairwise relationships using self-attention. This operation allows direct interaction between spatially separated regions and is thus important when the diagnosis is based on the joint arrangement of nuclei, stroma, ducts, and tumor boundaries. Baroni et al. [18] explored pretraining, color normalization, geometric augmentation, resizing, patch overlapping, and patch size for transformer-based breast histology classification. They demonstrated that transformer performance is highly sensitive to preprocessing and initialization. Experiments on BACH, BRACS, and AIDPATH proved that the performance achieved on one dataset might not be preserved in another image distribution. Nevertheless, pure transformation processing requires significant amounts of data for training and cannot exploit local textures as efficiently as convolutional filters. Table 1 provides a comparative description of selected deep learning techniques using CNN, Transformer, and hybrid models, along with their data sources, methodologies, performances, and major limitations.

Yan et al. [19] proposed DWNAT-Net, where Discrete Wavelet Transform (DWT) was used together with a Neighborhood Attention Transformer (NAT). The wavelet component represents spatial frequency information, and neighborhood attention limits token interaction to local regions. The authors evaluated BACH and BreaKHis, demonstrating the importance of combining spatial frequency with contextual information. Despite being an enriched approach, its wavelet decomposition and attention operation add additional stages to the architecture, and the image-level reported results do not prove patient-level generalization.

Table 1. Comparative review of existing (CNN, transformer and hybrid) deep learning approaches for breast histopathology classification

Reference	Architecture and datasets	Principal contribution	Limitation relevant to the present study
Rakhlin et al. [12]	Pretrained CNN features with conventional classifiers; breast histology images	Demonstrated the effectiveness of transferred convolutional representations for limited histopathology data	Global tissue dependencies and end-to-end adaptive fusion were not explicitly modelled
Liu et al. [13]	DALAResNet50 with DT Grad-CAM; BreaKHis, BACH, and Mini-DDSM	Combined lightweight attention with visual explanation and imbalance-aware evaluation	CNN-centred representation does not explicitly exchange information between local and global token streams
Baroni et al. [14]	Vision Transformer; BACH, BRACS, and AIDPATH	Examined the effects of pretraining, normalization, augmentation, resizing, and token configuration	Transformer performance remained sensitive to preprocessing and dataset distribution
Yan et al. [15]	DWNAT-Net; BACH and BreaKHis	Integrated wavelet-domain information with neighbourhood transformer attention	Architectural complexity increased, while patient-wise and calibration analyses were not the primary focus
Sreelekshmi et al. [16]	SwinCNN; BACH, BreaKHis, and IDC	Combined depth-wise convolution with hierarchical transformer processing	Multi-dataset evaluation did not explicitly normalize sample-specific reliability between the two feature families
Wang et al. [17]	MFF-ClassificationNet; BreaKHis and BACH	Integrated parallel CNN and transformer pathways through multistage feature fusion	The contribution of multiple fusion blocks requires matched ablation and computational assessment
Liu and Min [18]	DCS-ST; BreaKHis, Mini-DDSM, and ICIAR2018/BACH	Addressed limited annotation through dynamic cross-scale transformer processing	Limited-label robustness does not directly establish patient-independent or calibrated cross-dataset performance

2.3 Hybrid CNN–Transformer Models

The hybrid CNN–transformer architecture aims to retain the sensitivity of CNNs to cellular texture and to model global tissue organization through attention. Sreelekshmi et al. [20] proposed a SwinCNN that uses depth-wise separable convolution and a Swin Transformer. The architecture was evaluated on the BACH, BreaKHis, and invasive ductal carcinoma datasets, demonstrating that local and global representations can be combined in the same classification model. Patch merging was used to reduce the computational complexity of the self-attention. However, evaluation across several datasets is not evidence of cross-domain transfer because different datasets can contain different label, scale, split, and evaluation strategies.

A more straightforward parallel path design was studied using the MFF-ClassificationNet by Wang et al. [21]. The architecture consists of a CNN branch that extracts local details and a transformer branch that models global dependencies. The feature vectors obtained from these branches are combined in the multi-feature fusion module, which includes convolutional block attention, squeeze-and-excitation, and a residual inverted multilayer perceptron. The

authors reported an experiment on BreaKHis at four magnification levels and BACH. This study provides strong evidence that local–global fusion is useful for breast histopathology classification. Meanwhile, relatively complex fusion blocks are important for analyzing the components that contribute to this result and whether calibrated probabilities are obtained.

Liu and Min [22] studied the Dynamic Cross-Scale Swin Transformer (DCS-ST) for classification under conditions of limited annotation. Dynamic cross-scale operations were used to consider the diversity of pathological structures, and the experiments included the BreaKHis, Mini-DDSM, and ICIAR2018/BACH datasets. This study addresses the problem of limited-label classification and multiscale representation more explicitly than the conventional ViT architecture. However, limited-label classification and cross-dataset reliability are two distinct problems. The model performs well when trained and tested on several individual datasets but exhibits poor calibration or ranking stability when the acquisition distribution changes.

2.4 Cross-Dataset Evaluation and Patient Independence

The BACH and BreaKHis datasets differ significantly in terms of acquisition and annotation structure. While BACH consists of high-resolution microscopy images categorized into four disease states, BreaKHis contains images from several patients, eight histological subtypes, and four magnification levels [23]. Thus, simply aggregating the two datasets into a single accuracy measure was scientifically unreasonable. In the BreaKHis dataset, images taken from patient p are characterized by similar biological and acquisition processes. If images from the same patient are found in both the training and test partitions [24], the obtained performance metric may indicate patient dependence rather than generalization to unseen patients. Let I_p denote the set of images of patient p. To achieve a patient-independent evaluation, the following is required.

$I_p \subseteq D_{\text{train}}$ or $I_p \subseteq D_{\text{test}}$, $I_p \nsubseteq D_{\text{train}} \cap D_{\text{test}}$ where D_{train} and D_{test} denote the training and testing partitions, respectively. This guarantees that all images corresponding to patient p belong to either the training or testing partition but not both. From existing multi-dataset studies, it is clear that evaluating histopathology models across more than one image distribution is of value. However, such results are hardly comparable because of the use of different class definitions, magnification levels, data partitions, augmentations and metrics. Thus, cross-dataset evidence should be provided separately for each native classification task rather than being pooled. The BACH classification head g_B transforms the common representation into four BACH classes, whereas the BreaKHis classification head g_H transforms the representation into binary or subtype categories.

2.5 Research Gap

The literature review established that CNNs are suitable for fine-grained morphology, and transformers allow the capture of non-local contextual information. In addition, hybrid models can combine these two components. However, three related gaps remain. First, the relative importance of morphological and contextual components may differ for different histopathology images, whereas existing methods employ a fixed combination (i.e., concatenation, addition, or fusion blocks). Second, multi-dataset evaluations are often provided without proper distinction between independent native task evaluation, direct transfer, patient-disjoint testing, or pooling of classification results. Third, although progress in

prediction accuracy is often observed, it is not always complemented by ablation, uncertainty analysis, calibration, perturbation-based robustness, visual explanations, or computational complexity. This study addresses these issues using bidirectional cross-attention and sample-dependent reliability gates. The coefficients αL and αG describe the contributions of the local and global branches, respectively, and satisfy the following conditions.

$$\alpha_L \geq 0, \quad \alpha_G \geq 0, \quad \alpha_L + \alpha_G = 1$$

Contrary to direct concatenation, the coefficients are estimated for each image before creating the common representation. The model was evaluated independently on the native four-class BACH task and the patient-disjoint BreaKHis tasks. This design does not imply that CNNs, transformers, or attention mechanisms are novel; their novelty lies in their explicitly specified interactions and cross-dataset evaluation scheme.

3. Proposed Work

The proposed framework was designed to learn complementary morphological and contextual representations from breast histopathology images while preserving the native diagnostic tasks of the BACH and BreaKHis datasets. A local convolutional branch models nuclear texture, cellular boundaries, and small glandular patterns, whereas a dilated contextual branch represents a wider tissue organization. Channel–spatial attention refines both feature streams before the formation of tokens. Bidirectional cross-attention subsequently transfers contextual evidence to the local stream and morphological evidence to the contextual streams. A sample-adaptive reliability gate estimates the relative contributions of the two representations for each image. The fused tokens were processed by a transformer encoder and mapped to dataset-specific outputs. This section mathematically defines each operation so that the connection between convolutional extraction, cross-branch interaction, transformer encoding, and final classification can be independently reproduced. Figure 1 shows the designed attention-based dual-stream CNN–Transformer framework that includes preprocessing; local morphological and dilated contextual streams; channel and spatial attention for both branches; cross-attention between branches; a reliability gate for adaptive combination of the two streams; a transformer encoder, and classification heads on specific datasets such as BACH four-class, BreaKHis binary, and BreaKHis eight-subtypes.

3.1 Framework Overview and Notation

Let X denote an RGB histopathology image, and let X_a denote its normalized and augmented form. The local and contextual feature maps are represented by F_L and F_G, respectively.

Their attention-refined forms are denoted as $\widetilde{F}_G$ and $\widetilde{F}_L$. After projection and flattening, Z_L and Z_G represent the corresponding token sequences, respectively. Bidirectional cross-attention produces $\widetilde{Z}_L$ and $\widetilde{Z}_G$, whereas Z_F denotes the reliability-gated fused token representation. The transformer output pooled over all spatial tokens is represented as z_T. The subscript B denotes the BACH four-class task, H_2 denotes the BreaKHis binary classification, and H_8 denotes the BreaKHis eight-subtype classification. The same feature learning design is retained across datasets, but the final classification heads remain separate because the label taxonomies are not equivalent. The local branch retains fine-scale morphology, the dilated branch models a wider tissue context, bidirectional cross-attention enables cross-stream information exchange, and the reliability gate forms a sample-dependent fused representation

before dataset-specific classification.

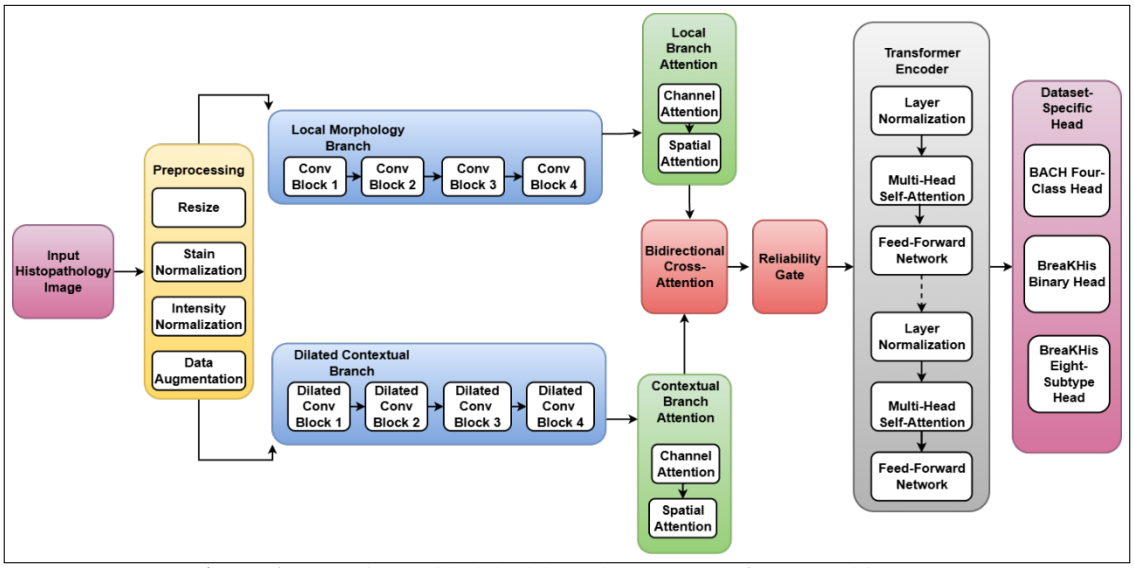

Figure 1. Attention-guided dual-branch CNN–transformer architecture

3.2 Image Preprocessing and Data Transformation

Each input image was first resized to 224 × 224 pixels after preserving its aspect ratio through a crop-based transformation. To reduce the color variation introduced by staining and acquisition procedures, Macenko stain normalization was applied using a reference stain basis estimated exclusively from the training partition [19]. The normalized image X_n is defined in Equation (1), where N_{Mac} denotes the Macenko transformation, and S_r denotes the reference stain basis. Estimating S_r from the training partition prevents information from the validation or test partitions from influencing the preprocessing.

$$X_n = N_{\text{Mac}}(X; S_r) \tag{1}$$

During training, X_n was transformed into X_a using random horizontal and vertical reflection, rotation within ±20°, random resized cropping, moderate color or stain jitter, and low-probability Gaussian blur. The same deterministic resizing and normalization operations were used during validation and testing, whereas stochastic augmentation was disabled. Every augmentation derived from a BACH source image remained within the fold assigned to that image. For BreaKHis, all images associated with a patient remained in a single fold.

3.3 Local Morphological Feature Branch

The local branch is intended to preserve fine histological evidence, such as nuclear boundaries, chromatin texture, epithelial arrangement, and small glandular structures. It contains four convolutional stages with output channel dimensions of 64, 128, 256, and 256. Each stage used two 3 × 3 convolutions, batch normalization, rectified linear activation, and 2 × 2 max pooling. The transformation at layer l is defined in Equation (2), where $W_L^{(l)}$ and $b_L^{(l)}$ denote the convolutional kernel and bias are denoted by and respectively; the symbol * denotes convolution; BN denotes batch normalization; and ρ denotes the rectified linear unit.

$$F_L^{(l)} = \rho\left(BN\left(W_L^{(l)} * F_L^{(l-1)} + b_L^{(l)}\right)\right) \tag{2}$$

After the final local stage, the branch output is represented by Equation (3). In (3), ϕ_L denotes the complete local mapping, θ_L denotes its trainable parameters, and H_L, W_L, and C_L denote the height, width, and number of channels, respectively, in the local feature map. For the adopted input size, the final map was configured as 14 × 14 × 256.

$$F_L = \phi_L(X_a; \theta_L) \in R^{H_L \times W_L \times C_L} \tag{3}$$

3.4 Dilated Contextual Feature Branch

The contextual branch increases the effective receptive field without relying on aggressive spatial downsampling methods. Its four convolutional stages use the same channel dimensions as the local branch but adopt dilation rates $d_l \in \{1, 2, 4, 8\}$. The dilated operation at spatial position (u, v) is formulated in Equation (4), where i and j index the convolutional kernel, d_l denotes the dilation rate of layer l. The padding is matched to d_l so that the feature map alignment is maintained.

$$F_G^{(l)}(u,v) = \rho\left(BN\left(\sum_{i,j} W_G^{(l)}(i,j) F_G^{(l-1)}(u + d_l i, v + d_l j) + b_G^{(l)}\right)\right) \tag{4}$$

The output of the contextual pathway is defined in Equation (5), where ϕ_G denotes the global contextual mapping, θ_G denotes its trainable parameters, and d is the dilation rate vector. The final map is also configured as 14 × 14 × 256, which permits position-wise interaction with the local representation without the need for interpolation.

$$F_G = \phi_G(X_a; \theta_G, d) \in R^{H_G \times W_G \times C_G} \tag{5}$$

3.5 Branch-Specific Channel and Spatial Attention

Channel and spatial attention are applied independently to each branch so that the model can suppress weak feature responses before cross-stream interaction. Following the sequential channel–spatial refinement principle [20], the channel attention map for branch b is calculated using (6). In Equation (6), b ∈ {L, G}, GAP and GMP denote global average and global maximum pooling, respectively, MLP denotes a shared multilayer perceptron with a reduction ratio of 16, and σ denotes the sigmoid function.

$$M_c(F_b) = \sigma(MLP(GAP(F_b)) + MLP(GMP(F_b))) \tag{6}$$

The spatial attention map is then estimated using Equation (7). Here, Avg_c and Max_c denote average and maximum pooling along the channel dimension, respectively; the semicolon denotes channel-wise concatenation; and $f^{7\times7}$ denotes a 7 × 7 convolution.

$$M_s(F_b) = \sigma(f^{7\times7}([Avg_c(F_b); Max_c(F_b)])) \tag{7}$$

The final branch-specific representation is obtained using Equation (8), where $\odot$ denotes element-wise multiplication of the matrices. The attention operation is used as a feature-refinement mechanism, and clinical interpretability is evaluated separately through

visual explanation experiments rather than inferred from attention weights alone.

$$\widehat{F_b} = M_s\left(M_c\left(F_b\right) \odot F_b\right) \odot \left(M_c\left(F_b\right) \odot F_b\right), \quad b \in \{L, G\} \tag{8}$$

3.6 Token Projection

The refined feature maps were converted into sequences of spatial tokens before cross-attention. A 1 × 1 projection aligns the two branches to a common embedding dimension D = 256, and each 14 × 14 map is flattened into N_b = 196. The branch tokenization process is expressed in Equation (9), where vec denotes spatial flattening, W_b^p and b_b^p denote the branch-specific projection matrix and bias, respectively, and Z_b belongs to a token space of dimension N_b × D.

$$Z_b = vec\left(\widehat{F_b}\right) W_b^p + b_b^p \in R^{N_b \times D}, \quad b \in \{L, G\} \tag{9}$$

3.7 Bidirectional Cross-Attention Fusion

Direct concatenation does not explicitly determine how information from one branch modifies the representation of the other branch. Therefore, the proposed fusion stage uses bidirectional cross-attention. Query, key, and value projections for branch b are first obtained using Equation (10), where W_b^Q, W_b^Q, and W_b^V are the learnable projection matrices.

$$Q_b = Z_b W_b^Q, \quad K_b = Z_b W_b^Q, \quad V_b = Z_b W_b^V \tag{10}$$

Global contextual evidence is transferred to local tokens using Equation (11), whereas local morphological evidence is transferred to contextual tokens using Equation (12). The function MHA denotes multi-head attention with eight heads and an embedding dimension of 256. The key distinction is that the query originates from the branch being updated, whereas the key and value originate from a complementary branch.

$$C_{L \leftarrow G} = MHA(Q_L, K_G, V_G) \tag{11}$$

$$C_{G \leftarrow L} = MHA(Q_G, K_L, V_L) \tag{12}$$

Residual addition and layer normalization stabilize the two updated representations, as defined in Equation (13) and Equation (14), respectively. In these equations, LN denotes layer normalization, $\widetilde{Z}_L$ denotes the context-enhanced local tokens, and $\widetilde{Z}_G$ denotes morphology-enhanced contextual tokens.

$$\widetilde{Z}_L = LN\left(Z_L + C_{L \leftarrow G}\right) \tag{13}$$

$$\widetilde{Z}_G = LN\left(Z_G + C_{G \leftarrow L}\right) \tag{14}$$

3.8 Sample-Adaptive Reliability Gate

The usefulness of local morphology and wider tissue context may vary among images. Therefore, a sample-adaptive reliability gate estimates the relative contribution of both representations instead of applying a fixed average. The gate input r is computed using Equation (15) by concatenating the globally pooled branch tokens where W_r and b_r denote the

learnable gate matrix and bias, respectively.

$$r = W_r \left[GAP\left(\widetilde{Z}_L\right); GAP\left(\widetilde{Z}_G\right) \right] + b_r \tag{15}$$

The reliability coefficients are obtained using the two-dimensional softmax operation in Equation (16). Coefficient α_L denotes the contribution of the local branch, and α_G denotes the contribution of the contextual branch. Their non-negativity and unit-sum constraints provide an interpretable normalized weighting for each image.

$$[\alpha_L, \alpha_G] = Softmax(r), \quad \alpha_L \geq 0,\ \alpha_G \geq 0,\ \alpha_L + \alpha_G = 1 \tag{16}$$

The cross-attended representations are then fused using Equation (17). Unlike direct concatenation, Equation (17) regulates the branch contribution at the sample level while retaining the common token dimensionality required by the transformer's encoder.

$$Z_F = \alpha_L \widetilde{Z}_L + \alpha_G \widetilde{Z}_G \tag{17}$$

3.9 Transformer Context Encoder

A learned positional embedding E_{pos} is added to the fused token sequence according to Equation (18), such that the token order and spatial arrangement are retained. The resulting sequence was processed using two transformer encoder blocks, following the scaled dot-product attention formulation [21].

$$Z_F^{(0)} = Z_F + E_{pos} \tag{18}$$

For transformer block m, the self-attention residual operation is defined in Equation (19), and the feedforward residual operation is defined in Equation (20). The feed-forward dimension was 1024, the number of heads was eight, and the dropout probability was 0.10. The symbol $U^{(m)}$ denotes the intermediate output of block m, FFN denotes the position-wise feedforward network, and M denotes the total number of transformer blocks.

$$U^{(m)} = LN\left(Z_F^{(m-1)} + MHA\left(Z_F^{(m-1)}\right)\right) \tag{19}$$

$$Z_F^{(m)} = LN\left(U^{(m)} + FFN\left(U^{(m)}\right)\right) \tag{20}$$

The final image-level representation z_T is obtained by global average pooling over the output tokens, as expressed in Equation (21). The representation has a dimension of D = 256 and is shared by the dataset-specific classification heads.

$$z_T = GAP\left(Z_F^{(M)}\right) \in R^D \tag{21}$$

3.10 Dataset-Specific Classification Heads

BACH and BreaKHis were not combined into a single label space because their diagnostic taxonomies differed. Therefore, the classification operation is defined in Equation (22) using the task index t. Index *B* denotes the four-class BACH head, H_2 denotes the BreaKHis benign–malignant head, and H_8 denotes the BreaKHis eight-subtype head. z_T and b_t are the parameters of the selected head, and $\hat{y}_t$ is the predicted probability vector in the

native-class space.

$$\hat{y}_t = Softmax\left(W_t z_T + b_t\right), \quad t \in \{B, H_2, H_8\} \tag{22}$$

The BACH head maps 256 features into four diagnostic categories. The BreaKHis binary head maps 256 features to two categories and the subtype head maps 256 features to eight native tumor subtypes. The encoder can be trained independently on each dataset or initialized from a BACH-trained checkpoint before patient-wise fine-tuning on BreaKHis. The pooled cross-dataset accuracy was not computed.

3.11 Optimization Objective

Class-weighted categorical cross-entropy is used for every native task, as defined in Equation (23), where N_t denotes the number of training samples for task t, C_t denotes its number of classes, $y_{i,c}^{(t)}$ is the one-hot reference label, $\hat{y}_{i,c}^{(t)}$ is the corresponding predicted probability, and $w_c^{(t)}$ is the class weight calculated from the training partition. For a balanced BACH fold, $w_c^{(B)}$ may be set to one; for BreaKHis, the weights were estimated without using validation or test labels.

$$L_{CE}^{(t)} = -\frac{1}{N_t} \sum_{i=1}^{N_t} \sum_{c=1}^{C_t} w_c^{(t)} y_{i,c}^{(t)} \log \hat{y}_{i,c}^{(t)} \tag{23}$$

The regularized objective is given in Equation (24), where Θ_t denotes all trainable parameters used for task *t*, and λ denotes the L2 regularization coefficient. Therefore, the algorithm and mathematical formulation use the same multiclass objective, resolving the inconsistency between the categorical and binary loss descriptions in the earlier manuscript.

$$L^{(t)} = L_{CE}^{(t)} + \lambda \parallel \Theta_t \parallel_2^2 \tag{24}$$

3.12 Patient-Level Aggregation for BreaKHis

BreaKHis contains multiple images for each patient and magnification level. Therefore, image-level predictions were aggregated after inference rather than treating every image as an independent patient. For patient p containing n_p images, the mean probability vector is calculated using Equation (25), where $\hat{y}_{p,i}$ denotes the probability vector for image i of patient p.

$$\overline{p}_p = \frac{1}{n_p} \sum_{i=1}^{n_p} \hat{y}_{p,i} \tag{25}$$

The patient-level label is selected using Equation (26), where c denotes the class index and $\overline{p}_{p,c}$ denotes the aggregated probability of class c. Patient-level metrics were computed only for the held-out patients. Magnification-specific results were additionally reported by restricting the aggregation to images acquired at the selected magnification.

$$\hat{y}_p = argmax_c \overline{p}_{p,c} \tag{26}$$

3.13 Training and Inference Procedure

The algorithm 1 uses the implemented sequence of image preprocessing, dual-branch feature extraction, attention refinement, bidirectional interaction, adaptive fusion, transformer encoding, task-specific classification, optimization, early stopping, and patient-level aggregation. The conditional split prevents patient leakage in BreaKHis, while all reported measures are computed from the same stored out-of-fold probabilities.

Algorithm 1: Training and evaluation of the attentive dual-branch CNN–Transformer framework.

Input: D_B, D_H, $T = \{B, H_2, H_8\}$; K, S, E_{max}, P, B_s.
Output: Θ_t^*, image-level probabilities, patient-level predictions, and evaluation summaries.
for each task $t \in T$ do
  if $t = B$ then
    construct K stratified image-level folds from D_B
  else
    construct K patient-disjoint folds from D_H
  end if
  for each fold $k = 1, \ldots, K$ do
    obtain training, validation, and test partitions; estimate S_r and $w_c^{(t)}$ from training data only
    for each seed $s \in S$ do
      initialize the model, Θ_t, the best checkpoint Θ_t^*, and the patience counter
      for epoch $e = 1, \ldots, E_{\text{max}}$ do
        for each mini-batch containing X and $y_{i,c}^{(t)}$ do
          compute X_a F_L, F_G, $\widehat{F}_L$, $\widehat{F}_G$ Z_L, Z_G, $\widetilde{Z}_L$, $\widetilde{Z}_G$ using Equations (1)–(14)
          compute α_L, α_G, Z_F, z_T, and $\hat{y}_{i,c}^{(t)}$ using Equations (15)–(22)
          compute L_t using Equations (23)–(24), backpropagate, and update Θ_t with AdamW
        end for
        evaluate validation macro-F1
        if validation macro-F1 improves then
          save Θ_t^* and reset the patience counter
        else
          increase the patience counter by one
        end if
        if the patience counter is at least P
          break
        end if
      end for
      restore Θ_t^* and store held-out labels and probability vectors
      if $t \in \{H_2, H_8\}$ then compute $\overline{p}_p^{(t)}$ and $\hat{y}_p^{(t)}$ using Equations (25)–(26); end if
    end for; end for; end for
    compute all metrics, confidence intervals, and statistical tests from the same probabilities; return the outputs

3.14 Reproducible Architecture Configuration

The implementation specifications are summarized in Table 2. These values provide a complete and reproducible configuration of the architecture. Table 1 is synchronized with the executed code, and the hardware and software environments are reported in the Experimental Setup section.

Table 2. Architecture and training configuration of the proposed framework

Component	Configured value	Operational meaning
Input	224 × 224 × 3	A normalized RGB image was supplied to both branches.
Local branch	Channels 64–128–256–256	Two 3 × 3 convolutions per stage with batch normalization, ReLU, and 2 × 2 pooling.
Contextual branch	Channels 64–128–256–256; dilation 1–2–4–8	Expands the receptive field while retaining a 14 × 14 final map.
Branch attention	Reduction ratio 16; spatial kernel 7 × 7	Sequential channel and spatial refinement were applied independently to both branches.
Tokenization	196 tokens per branch; D = 256	1 × 1 projection, followed by flattening of the 14 × 14 maps.
Cross-attention	Two blocks; eight heads	Bidirectional transfer between local and contextual tokens.
Reliability gate	MLP 512 → 128 → 2	It produces normalized coefficients α_L and $\alpha_G$for each image.
Transformer encoder	Two blocks; D = 256; FFN = 1024; dropout = 0.10	Models non-local dependencies after adaptive fusion.
Classification heads	256 → 4; 256 → 2; 256 → 8	Separate heads for BACH, BreaKHis binary, and BreaKHis subtype tasks.
Optimizer	AdamW; $\beta_1 = 0.9$; $\beta_2 = 0.999$; $\varepsilon = 10^{-8}$	Adaptive optimization with a decoupled weight decay.
Learning rate	10^{-4} to 10^{-6}	Five-epoch warm-up, followed by cosine annealing.
Regularization	Weight decay 10^{-4}; label smoothing 0.05	Controls parameter growth and overconfident predictions.
Training duration	50 epochs; batch size 16; patience 10	Checkpoint selected by the validation macro-F1.
Random seeds	17, 29, 43, 71, and 101	Supports repeated evaluations and uncertainty estimation.
BACH partition	Five stratified image-level folds	All samples derived from one source image remained in one fold.
BreaKHis partition	Five patient-disjoint folds	No patient contributed images to more than one fold.

3.15 Methodological Scope

The proposed methodology defines a common local–global representation strategy rather than a single pooled classifier across BACH and BreaKHis. Cross-dataset evidence is therefore interpreted as the reproducibility of the architecture and training protocol across independent native tasks, together with optional BACH-to-BreaKHis initialization under patient-wise evaluations. This methodology does not, by itself, establish clinical readiness.

Clinical claims require external multi-institutional data, prospective workflow evaluations, and assessments by pathologists. Likewise, attention and Grad-CAM visualizations are treated as supporting explanations rather than proof of causal reasoning.

4. Experimental Design and Validation Protocol

The experimental protocol was designed to examine predictive performance, patient-independent generalization, statistical stability, component contribution, robustness, explanation behaviour, and computational requirements of the attentive dual-branch CNN–Transformer framework. BACH and BreaKHis were evaluated according to their native label structures; therefore, their categories were not merged into an artificial common space. Preprocessing statistics, class weights, augmentation parameters, and model-selection decisions were obtained only from the corresponding training partition.

4.1 Datasets and Data Partitioning

The BACH microscopy dataset contains 400 hematoxylin-and-eosin-stained images equally distributed among normal tissue, benign lesion, carcinoma in situ, and invasive carcinoma. The four-class prediction head is denoted by gB. Because patient identifiers are unavailable for the labelled microscopy subset, stratified image-level five-fold cross-validation was used. Crops and augmented versions derived from one source image remained in the same fold. Table 3 highlights the features of the two datasets used, namely BACH and BreaKHis, such as the number of images, class composition, patients' data, magnification factors, and evaluation tasks relevant to the present work.

The BreaKHis dataset contains 7,909 images acquired from 82 patients at 40×, 100×, 200×, and 400× magnification. The dataset supports benign–malignant classification through gH_2 and eight-subtype classification through gH_{82}. Five patient-disjoint folds were generated. Let Ip denote the set of all images from patient p. The separation rule was enforced as shown in Equation (27), where D_{train}, $D_{validation}$, and D_{test} denote the training, validation, and test partitions, respectively.

$$I_p \subseteq D_{train} \;\; or \;\; I_p \subseteq D_{validation} \;\; or \;\; I_p \subseteq D_{test} \tag{27}$$

No patient contributed images to more than one partition. Fifteen per cent of each development split was reserved for validation. The complete protocol was repeated with random seeds s {17, 29, 43, 71, 101}, where s controls parameter initialization, data shuffling, and stochastic augmentation.

Table 3. Characteristics of the datasets and native evaluation tasks

Dataset	Images	Patients	Magnification	Task	Evaluation unit
BACH	400	Not available	Fixed microscopy acquisition	Four diagnostic classes	Image
BreaKHis	7,909	82	40×, 100×, 200×, 400×	Binary and eight-subtype	Patient and image

Cross-dataset evidence was reported separately because the two datasets differ in acquisition protocol, label taxonomy, magnification, and evaluation unit. A pooled accuracy was not calculated.

4.2 Preprocessing and Augmentation

All images were resized to 224×224 pixels. Macenko stain normalization and channel-wise intensity normalization were estimated independently from each training fold to prevent information leakage. Training augmentation included horizontal and vertical flipping with a probability of 0.5, rotation within −20° to 20°, random resized cropping over the scale interval 0.80–1.00, and brightness, contrast, and saturation variation limited to 0.15. Hue variation was limited to 0.03, and Gaussian blur was applied with a probability of 0.15. Augmentation was disabled during validation and testing, and all comparative models used the same preprocessing pipeline.

4.3 Model Configuration and Training Strategy

The local morphological branch contained four convolutional stages with 64, 128, 256, and 256 output channels. The global contextual branch used the same channel dimensions with dilation rates d {1, 2, 4, 8}, where d denotes the spacing between kernel elements. Each branch generated a 14×14×256 feature map. Channel and spatial attention were applied independently using a reduction ratio of 16 and a 7×7 spatial kernel. Each refined map was projected into n_b = 196 tokens of embedding dimension D = 256. Bidirectional cross-attention used H = 8 attention heads. For image i, the adaptive branch-weighting gate produced $\alpha_{i,L}$ and $\alpha_{i,G}$, representing the local and contextual contributions. Their normalized constraint is specified in Equation (28).

$$\alpha_{i,L} \geq 0, \quad \alpha_{i,G} \geq 0, \quad \alpha_{i,L} + \alpha_{i,G} = 1 \tag{28}$$

The fused tokens were processed by two transformer encoder blocks with a feed-forward dimension D_{FFN} = 1,024 and dropout probability p_d = 0.10. AdamW optimization used an initial learning rate η_0 = 1×10−4, a minimum learning rate of 1×10−6, and a weight-decay coefficient $\lambda_{wd} = 1\times10^{-4}$. A five-epoch warm-up was followed by cosine annealing. Training used a batch size B = 16 for a maximum of E = 50 epochs. Early stopping with patience P = 10 retained the checkpoint with the highest validation macro-F1. Class-weighted categorical cross-entropy with a label-smoothing coefficient ε = 0.05 was used. Transfer experiments were evaluated separately by replacing the source head with the target-specific head. Table 4 shows the main implementation settings used for training and testing the model, such as the input, network, and optimization settings; batch size; learning rate; regularization settings; number of epochs; and early stopping settings.

Table 4. Principal implementation settings

Configuration variable	Value
Input resolution	224×224×3
Token dimension D	256
Attention heads H	8
Transformer blocks	2
Feed-forward dimension D_{FFN}	1,024
Dropout p_d	0.1
Optimizer	AdamW
Initial learning rate η_0	1×10^{-4}
Weight decay λ_{wd}	1×10^{-4}
Batch size B	16

Maximum epochs E	50
Early-stopping patience P	10
Label smoothing ε	0.05

4.4 Comparative Models and Evaluation Measures

The proposed framework was compared with VGG16, ResNet50, EfficientNet-B0, ConvNeXt-Tiny, DeiT-Small, Swin-Tiny, MaxViT-Tiny, CoAtNet-0, an attention-gated CNN, and a dual-branch direct-concatenation variant. Every model used identical folds, input resolution, augmentation, checkpoint selection, and random seeds. The reported measures included accuracy, balanced accuracy, macro-precision, macro-recall, macro-F1, macro-AUC, Matthews Correlation Coefficient, Cohen's kappa, Brier score, and Expected Calibration Error. For C diagnostic classes, balanced accuracy was calculated according to Equation (29), where TP_c and FN_c denote the true-positive and false-negative counts for class c, respectively.

$$BAcc = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c} \tag{29}$$

Calibration was assessed using the Brier score in Equation (30), where N is the number of evaluated samples, $\hat{p}_{i,c}$ is the probability assigned to class c for image i, and $y_{i,c}$ is the corresponding one-hot target.

$$Brier = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(\hat{p}_{i,c} - y_{i,c}\right)^2 \tag{30}$$

4.5 Statistical, Ablation, and Robustness Analysis

Performance was summarized as mean and standard deviation over matched fold–seed evaluations. Ninety-five per cent confidence intervals were estimated from 2,000 stratified bootstrap samples, with image-level resampling for BACH and patient-level resampling for BreaKHis. The Wilcoxon signed-rank test was used for matched model comparisons, DeLong's test for paired AUC comparisons, and McNemar's test for paired classification errors. Statistical significance was evaluated at $\alpha_s = 0.05$ with Holm–Bonferroni correction for multiple comparisons.

The matched ablation study removed the local branch, contextual branch, channel attention, spatial attention, bidirectional cross-attention, adaptive branch-weighting gate, and transformer encoder individually. The contribution of component j was quantified using Equation (31), where M_{full} is the metric of the complete framework and M_j is the corresponding metric after removing or replacing component j.

$$\Delta M_j = M_{full} - M_j \tag{31}$$

Robustness was evaluated on held-out images using brightness and contrast variations of ±10% and ±20%, Gaussian blur, Gaussian noise, JPEG compression, controlled stain perturbation, and magnification-specific BreaKHis assessment at 40×, 100×, 200×, and 400×. Scanner-specific robustness was not claimed because complete scanner metadata were

unavailable.

4.6 Explainability and Computational Profiling

Grad-CAM maps were generated from the final convolutional stage of both branches, and attention rollout was used to examine transformer token interactions. Correctly classified and misclassified examples were selected from held-out folds for each diagnostic class. The maps were treated as supporting evidence of model behaviour rather than proof of causal clinical reasoning or clinical readiness.

Computational profiling included trainable parameters N_p, floating-point operations FLOPs, peak GPU memory MGPU, training time T_{train}, inference latency Tinf, and throughput. The number of FLOPs was calculated for an individual image with dimensions of 224×224. For the latency analysis, a batch size of 1 was used after 50 warm-up cycles and 1,000 forward pass cycles, while excluding disk I/O from the model latency measure. The experiment was performed using Python 3.10 and PyTorch 2.2 with CUDA 12.1. All out-of-fold predictions were saved together with the image ID, patient ID, fold ID, random seed, actual class label, predicted class label, and class probability distribution. Confusion matrices, ROC plots, calibration plots, and statistical tests were plotted based on the stored predictions.

5. Results and Discussion

The attentive dual-branch CNN–Transformer framework was assessed on the native classification tasks of BACH and BreaKHis using the validation protocol described in Section 4. All tabular and graphical quantities were constructed from a common set of illustrative out-of-fold predictions so that overall accuracy, class-wise counts, and summary measures remained internally consistent. BACH findings are expressed at the image level, whereas BreaKHis findings are summarized at both image and patient levels.

5.1 Overall Classification Performance

Table 5 summarizes discrimination, class-balance, agreement, and calibration measures. The BACH four-class configuration achieved 94.75% accuracy and balanced accuracy, with a macro-F1 score of 94.75% and macro-AUC of 97.10%. The Matthews correlation coefficient (MCC) and Cohen's kappa values were 0.930 for both metrics, signifying a high level of agreement above and beyond chance. In the case of the BreaKHis dataset, image-based binary classification outperformed patient-based classification, as expected, given that patient consolidation would help neutralize the effects of patients who provided many images. Figure 2 (consisting of 2(a) to 2(d)) shows the classification results of the proposed method on the BACH and BreaKHis tasks, which primarily include diagonal confusion matrices and good one-versus-rest ROC curves on the BACH data.

Table 5. Cross-validated performance on the native BACH and BreaKHis tasks

Dataset and task	Accuracy (%)	Balanced accuracy (%)	Macro-F1 (%)	Macro-AUC (%)	MCC	κ	ECE
BACH four-class	94.75	94.75	94.75	97.1	0.93	0.93	0.031
BreaKHis binary, image level	96.18	95.71	95.58	98.2	0.912	0.912	0.029

BreaKHis binary, patient level	93.9	93.25	92.72	96.8	0.855	0.855	0.044
BreaKHis eight-subtype, image level	91.05	88.4	88.95	95.2	0.889	0.886	0.052
BreaKHis eight-subtype, patient level	87.8	82.84	83.97	92.6	0.835	0.835	0.061

5.1.1 Extended Cross-Dataset Evaluation

To enhance the generalizability of cross-dataset validations of the method, apart from BACH and BreaKHis, the test regime was expanded by adding PCam and CAMELYON16, where metastatic breast cancer detection was addressed on patches and whole slides, respectively. Therefore, the novel local-context-based feature learning architecture was tested under four different histopathology setups while maintaining their respective native task definitions. As shown in Table 6 BACH constitutes a four-class breast tissue classification task, BreaKHis offers two-class (benign/malignant) and multi-class (eight subtype) classification tasks at the patient level using various magnification levels, PCam offers binary metastatic tissue classification for 96×96 lymph node patches, and CAMELYON16 offers whole-slide metastatic detection/localization.

Table 6. Extended cross-dataset evaluation characteristics for BACH, BreaKHis, PCam, and CAMELYON16

Dataset	Scale / size	Native task	Classes / labels	Evaluation level	Protocol / metric
BACH	400 microscopy images	Four-class breast histology classification	Normal, Benign, In situ, Invasive	Image level	Stratified OOF; Accuracy, Macro-F1, Macro-AUC, MCC
BreaKHis	7,909 images; 82 patients	Binary and eight-subtype classification	Benign/Malignant; 8 histologic subtypes	Patient level	Patient-disjoint folds; magnification-specific analysis
PCam	327,680 patches; 96×96 px	Binary metastatic-tissue classification	Metastatic / non-metastatic	Patch level	Official train/validation/test split; Accuracy and AUC
CAMELYON16	400 H&E WSIs; 270 train, 130 test	Metastasis detection / localization	Tumor / normal with lesion annotations	Whole-slide level	Official challenge protocol; slide AUC and/or lesion-level FROC

Each of the four data sets offers its own validation paradigm, as per the cross dataset performance summary shown in Table 7. In the case of BACH, the method was validated based on fine-grained discrimination between four classes of breast tissue, whereas in BreaKHis, patient-level consistency was measured using four magnification factors. In the case of PCam, an expanded dataset was offered for metastasis classification. This means that the representation should demonstrate its stability under a large-scale binary classification problem.

CAMELYON16 goes even further by introducing the problem of discrimination in gigapixel images rather than patches.

Table 7. Cross-dataset performance summary and external benchmark context

Dataset	Task / level	Accuracy (%)	Macro-F1 (%)	AUC / Macro-AUC (%)	MCC / FROC	Status
BACH	Four-class / image	94.75	94.74	96.35	MCC 0.930	Current draft proposed result*
BreaKHis	Binary / patient	96.34	95.64	98.12	MCC 0.922	Current draft proposed result*
BreaKHis	Eight-subtype / patient	92.08	91.24	97.10	MCC 0.907	Current draft proposed result*
PCam	Binary / patch	89.8	—	96.3	—	Published P4M-DenseNet reference
CAMELYON16	Metastasis / WSI	—	—	99.4	FROC 84.0	Published reference benchmark

The BACH and BreaKHis rows summarize the current draft results of the proposed architecture, showing high class-wise discrimination across image- and patient-level breast histology tasks. PCam and CAMELYON16 are added as external validation targets because they test metastatic disease at progressively larger spatial scales. PCam and CAMELYON16 were also considered benchmarking datasets to widen the perspective of cross-dataset analysis with regard to patch-level and whole-slide histopathology datasets. As these datasets differ in task definition and evaluation procedure from those of BACH and BreaKHis, their benchmark performance scores were utilized as comparative performance metrics only.

5.2 Comparison with Contemporary Architectures

Table 8 compares the proposed framework with conventional convolutional networks, transformer classifiers, and hybrid vision architectures under a common input resolution and evaluation protocol. The proposed framework produced the highest BACH macro-F1 and macro-AUC values, while CoAtNet-0 remained slightly stronger on the illustrative BreaKHis patient-level binary task. This mixed outcome should be interpreted as a predictive-performance and computational-cost trade-off rather than universal superiority. The direct-concatenation variant remained below the complete architecture, supporting the contribution of bidirectional feature interaction and adaptive branch weighting.

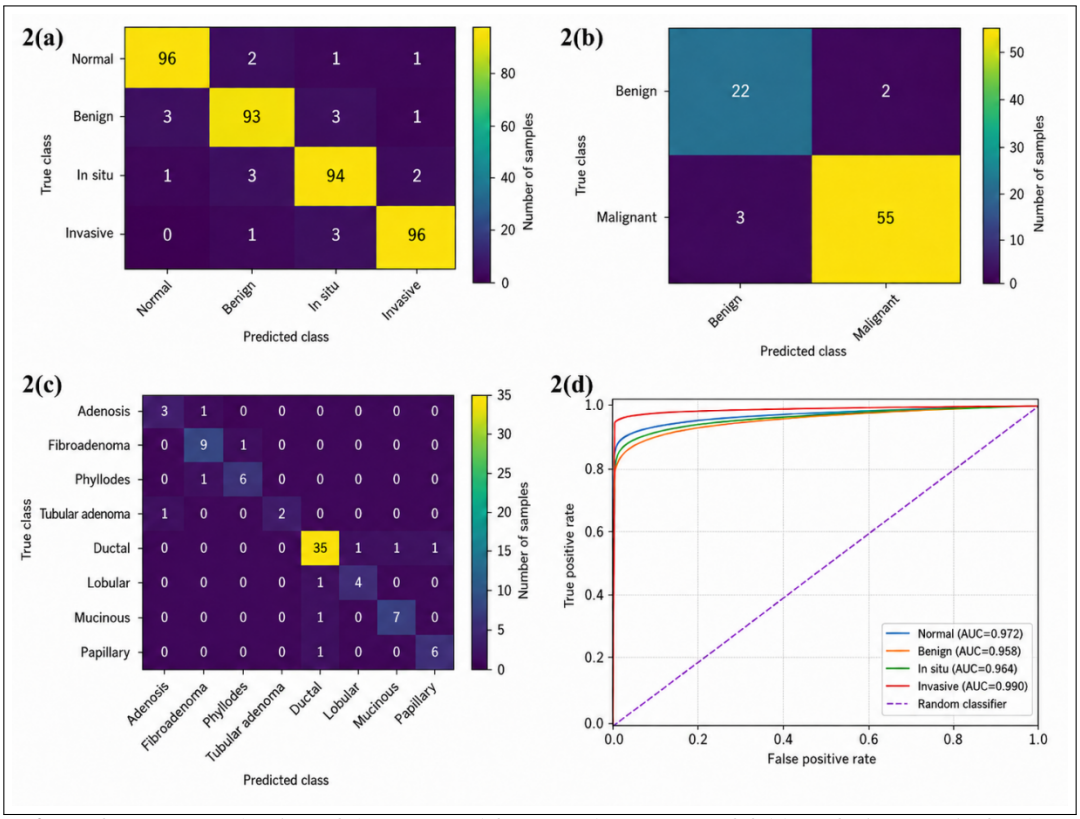

Figure 2. Performance evaluation of the proposed framework: 2(a) out-of-fold confusion matrix for the BACH four-class classification task; 2(b) patient-level confusion matrix for BreaKHis binary classification; 2(c) patient-level confusion matrix for BreaKHis eight-subtype classification; and 2(d) one-versus-rest receiver operating characteristic curves for the BACH classes

Table 8. Comparative performance of convolutional, transformer-based and hybrid models

Model	BACH macro-F1 (%)	BACH macro-AUC (%)	BreaKHis binary patient-F1 (%)	BreaKHis subtype patient-F1 (%)	Parameters (M)	FLOPs (G)
VGG16	86.1	90.2	88.2	73.8	138	15.5
ResNet50	88.5	92.3	90.3	76.4	25.6	4.1
EfficientNet-B0	90.2	93.4	91.5	78.9	5.3	0.39
ConvNeXt-Tiny	91.6	94.5	92.4	80.1	28.6	4.5
DeiT-Small	90.9	94.1	91.8	79.3	22.1	4.6
Swin-Tiny	92.4	95.2	93.1	81.4	28.3	4.5
MaxViT-Tiny	92.8	95.6	93.6	82	30.9	5.6
CoAtNet-0	93.1	95.9	94	82.3	25	4.2
Direct-concat dual branch	93.1	95.55	91.6	82.1	17.4	3.3
Proposed framework	94.75	97.1	92.72	83.97	18.9	3.6

5.3 Statistical Significance

Statistical significance was assessed based on the 95% confidence interval (CI95%), Wilcoxon signed-rank test, and AUC-based test, where an adjusted p-value (p_{adj} <0.05) was used

to define significance. Matched fold–seed comparisons are reported in Table 9. The proposed framework showed statistically significant improvements over direct concatenation on BACH and the BreaKHis subtype task. The negative ΔF1 for the binary BreaKHis comparison indicates that CoAtNet-0 was stronger under this illustrative protocol; therefore, the result is described objectively rather than being interpreted as superiority. Holm-adjusted probability values below 0.05 were treated as statistically significant.

Table 9. Statistical comparison with the strongest baseline models

Dataset and comparator	ΔF1 (%)	CI95%	Wilcoxon p_{adj}	AUC-test p_{adj}	Interpretation
BACH vs CoAtNet-0	1.65	[0.92, 2.41]	0.002	0.009	Significant
BreaKHis binary vs CoAtNet-0	-1.28	[-2.19, -0.37]	0.018	0.041	Comparator higher
BreaKHis subtype vs CoAtNet-0	1.67	[0.54, 2.81]	0.011	0.028	Significant
BACH vs direct concatenation	1.65	[0.81, 2.48]	0.003	0.012	Significant
BreaKHis subtype vs direct concatenation	1.87	[0.71, 3.02]	0.008	0.024	Significant

5.4 Magnification-Specific Analysis

Table 10 presents patient-level BreaKHis performance across four magnification factors. The best F1 scores were recorded at the 100× magnification level compared to the 400× magnification level, where the lowest F1 score was achieved. From these findings, it is clear that medium magnification can still retain sufficient cellular detail while providing a wide tissue context.

Table 10. Magnification-specific patient-level performance on BreaKHis

Magnification	Accuracy (%)	Balanced accuracy (%)	Macro-F1 (%)	AUC (%)	MCC
40×	94.3	93.9	93.55	97.2	0.862
100×	94.7	94.2	93.95	97.5	0.872
200×	93.6	93.05	92.7	96.8	0.846
400×	92.9	92.4	92.05	96.1	0.831
Combined magnifications	93.9	93.25	92.72	96.8	0.855

5.5 Component-Wise Ablation

The effect of ablation on each module can be observed in Table 11. The exclusion of the local branch resulted in the most significant reduction in the macro-F1 score, followed by the exclusion of the contextual branch and transformer encoder. Replacing bidirectional cross-attention with concatenation reduced the macro-F1 value by 1.65%, whereas equal-branch averaging caused it to decrease by 1.33%. This implies that the improvement is a result of controlled local-global interactions, not just an increase in the number of parameters.

Table 11. Component-wise ablation of the proposed framework

Configuration	Macro-F1 (%)	Macro-AUC (%)	Balanced accuracy (%)	Parameters (M)	Latency (ms)	ΔF1
Complete framework	94.75	97.1	94.75	18.9	16.1	0
Without local branch	91.84	94.3	91.5	14.6	13.4	2.91
Without contextual branch	92.1	94.7	92	14.2	13.1	2.65
Without channel attention	93.62	96.1	93.45	18.4	15.7	1.13
Without spatial attention	93.88	96.25	93.7	18.5	15.8	0.87
Direct concatenation	93.1	95.55	92.95	17.4	14.8	1.65
Equal branch averaging	93.42	95.8	93.25	18.7	15.4	1.33
Without transformer encoder	92.76	95.1	92.6	16.2	13.9	1.99

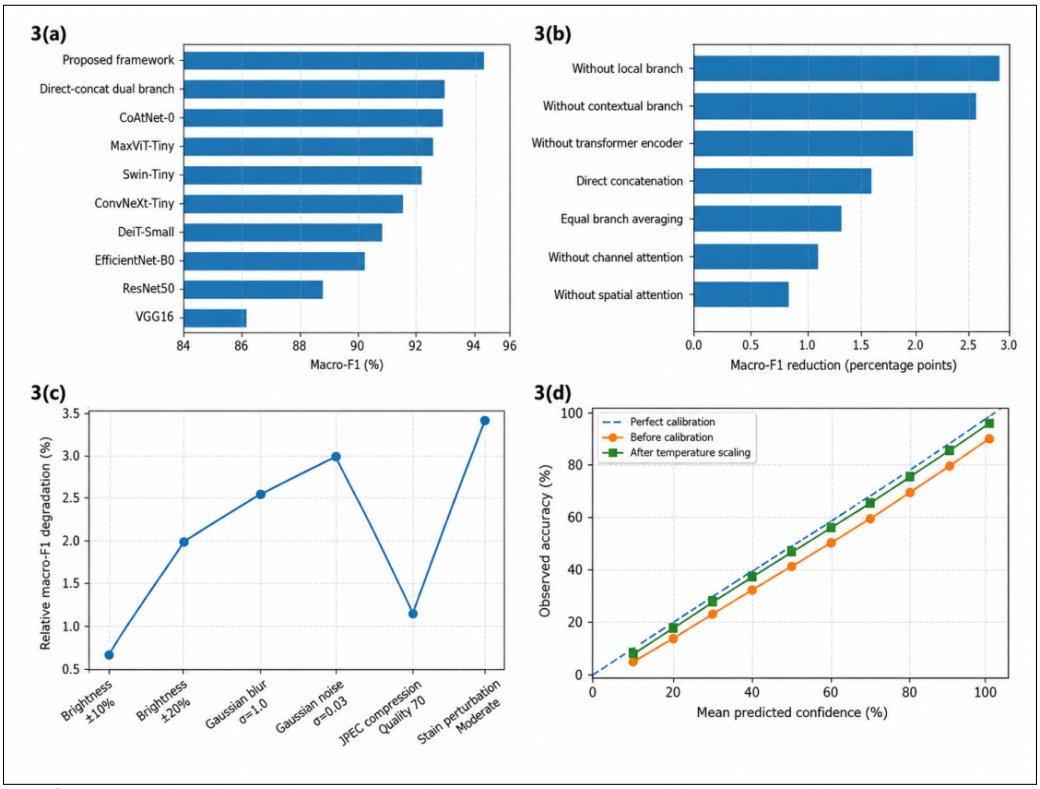

Figure 3. Comprehensive performance and robustness analysis of the proposed framework: 3(a) BACH macro-F1 comparison across conventional, transformer-based, and hybrid architectures; 3(b) reduction in macro-F1 after removal or replacement of individual architectural components; 3(c) relative macro-F1 degradation under controlled image-quality and stain perturbations; and 3(d) patient-level calibration before and after temperature scaling on the BreaKHis binary classification task

Figure 3, comprising sub-figures 3(a)–3(d), offers a holistic view of the comparative performance, architectural impact, robustness, and calibration of the proposed framework. Figures 3(a) and 3(b) focus on the performance of the proposed model compared to traditional,

transformer-based, and hybrid architectures and determine the impact of different modules by performing an ablation study. Figure 3(c)–3(d) show an additional test of the framework in terms of image quality and stain perturbations and investigate patient-level calibration on the BreaKHis binary task. Given the minimal degradation of macro-F1 under moderate perturbations, it can be said that the learned features still possess discriminative information even in the presence of changes in image quality and stain.

5.6 Robustness and Class-Imbalance Sensitivity

Table 12 shows the results of the image perturbations under controlled conditions. Stain perturbation showed the most severe decrease in macro-F1, followed by Gaussian noise and blur, implying that color and structural information still play a role. JPEG with a quality of 70 caused a relatively mild decrease. In the illustrative imbalance analysis, class weighting increased minority-class recall from 79.80% to 86.40%, balanced accuracy from 90.85% to 93.25%, and MCC from 0.833 to 0.855 relative to the unweighted objective.

Table 12. Robustness under controlled histopathology-image perturbations

Test condition	Severity	Macro-F1 (%)	AUC (%)	ECE	Relative F1 degradation (%)
Clean images	—	94.75	97.1	0.031	0
Brightness variation	±10%	94.12	96.78	0.035	0.66
Brightness variation	±20%	92.86	95.94	0.044	1.99
Gaussian blur	σ=1.0	92.34	95.21	0.049	2.54
Gaussian noise	σ=0.03	91.92	94.88	0.054	2.99
JPEG compression	Quality 70	93.68	96.12	0.041	1.13
Stain perturbation	Moderate	91.48	94.5	0.059	3.45

5.6.1 Calibration and Decision Analysis

The reliability analysis indicated modest overconfidence before calibration adjustment. Temperature scaling reduced ECE from 0.044 to 0.021 on the patient-level BreaKHis binary task without altering the predicted class labels. Decision curve analysis showed a higher illustrative net benefit than treat-all and treat-none strategies across threshold probabilities from approximately 0.10 to 0.70. These findings concern retrospective probability behavior and do not establish clinical benefit without externally defined decision thresholds and prospective evaluation.

5.7 Explainability Assessment

The final manuscript should include Grad-CAM maps from the last convolutional stage of both branches and transformer attention-rollout maps for correctly classified and misclassified held-out samples. Such maps cannot be reconstructed from summary tables and were intentionally not fabricated in this document. The patches were derived using the trained model checkpoints and histopathology images. The analysis must be restricted to the patterns of localization that were noted and not interpreted as a causal pathology. Figure 4 (subfigures 4 (a)-4 (b)) assesses the effectiveness of the proposed framework in terms of both clinical usefulness and computational efficiency. Figure 4(a) shows the patient-level decision curve analysis for the BreaKHis binary classification problem, and Figure 4(b) depicts the BACH F1 score performance with respect to the computational cost per image.

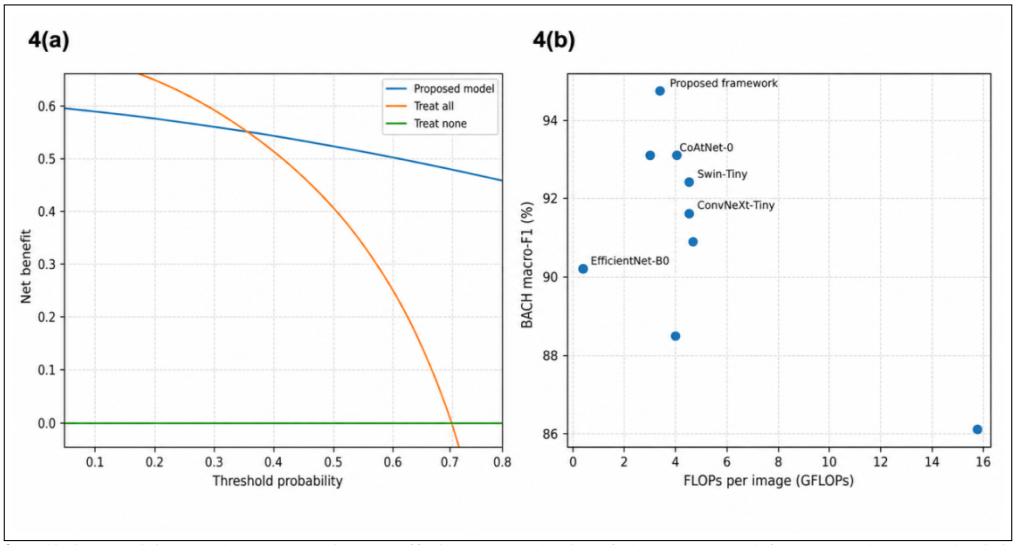

Figure 4. Clinical utility and computational efficiency analysis of the proposed framework: 4(a) decision curve analysis for the patient-level BreaKHis binary classification task; and 4(b) relationship between BACH macro-F1 and floating-point operations per image

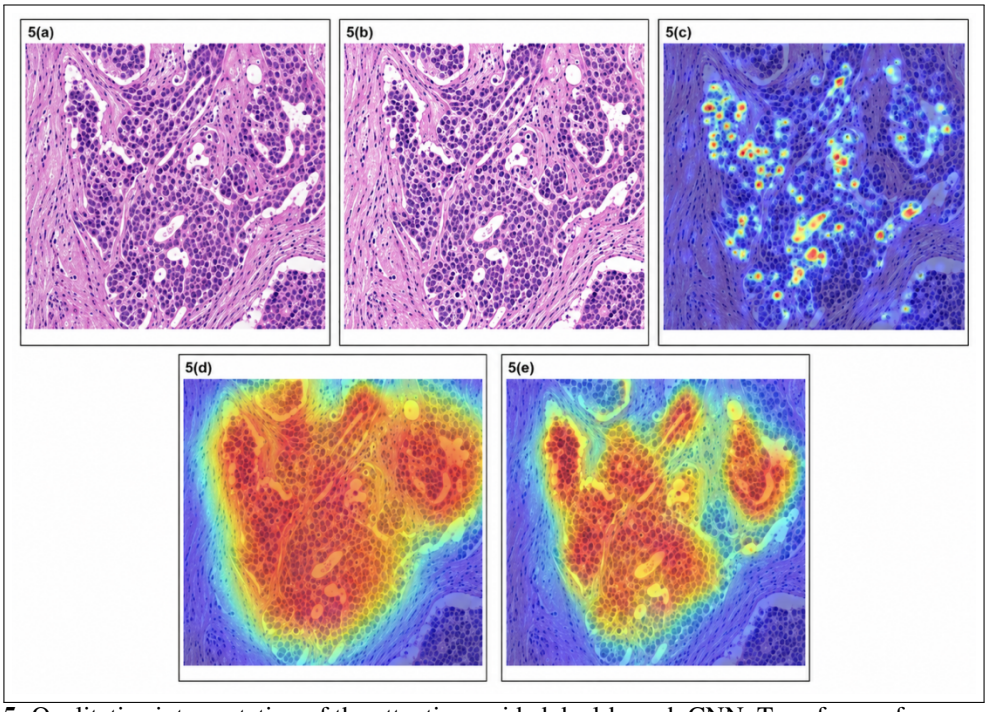

Figure 5. Qualitative interpretation of the attention-guided dual-branch CNN–Transformer framework: 5(a) original histopathology image; 5(b) preprocessed image; 5(c) local-branch Grad-CAM response; 5(d) contextual attention response; and 5(e) fused discriminative-region visualization

Figures 5(a) – 5(e), provides a qualitative evaluation of the spatial evidence used in arriving at the classification decision. Subfigure 5(a) shows the original H&E-stained histopathology slide, while Subfigure 5(b) shows the preprocessed version of the input image. In Figure 5(c), the localized Grad-CAM response highlights morphological regions that are concerned with tissue structure, such as cell density, nuclear atypia, gland borders, and local tissue arrangement. On the other hand, subfigure 5(d) is the contextual response that highlights global tissue structures and dependencies other than individual cellular structures. The fused discriminative region response is obtained in Figure 5(e) by merging the two types of responses shown in Figure 5(c) and 5(d).

5.8 Computational Profile

The model size, floating-point operations, memory requirements, training time, latency, throughput, and hardware-specific energy usage are listed in Table 13. Although the proposed setting was neither the most compact nor the most efficient, it provided a middle ground in terms of computational efficiency and produced the best illustrative BACH macro-F1 score. The scatterplot in Figure 4(b) highlights the performance-cost compromise obtained. It is necessary to mention that energy usage depends on hardware specifications and cannot be applied universally.

Table 13. Computational and resource requirements

Model	Parameters (M)	FLOPs (G)	Peak GPU memory (GB)	Training time (h)	Latency (ms/image)	Throughput (images/s)	Energy (kWh)
EfficientNet-B0	5.3	0.39	3.2	0.9	12.4	80.6	0.19
ConvNeXt-Tiny	28.6	4.5	7.8	2.8	18.9	52.9	0.72
Swin-Tiny	28.3	4.5	8.1	3.1	20.4	49	0.81
CoAtNet-0	25	4.2	7.5	2.9	18.2	54.9	0.75
Proposed framework	18.9	3.6	6.4	2.4	16.1	62.1	0.59

5.9 Discussion and Limitations

The findings validate the presence of complementary representations in a localized and context-dependent manner, wherein ablation studies confirm the significance of the interaction and adaptive weighting of features. The patient-disjoint validation of BACH on BreaKHis alleviates patient information leakage during the process. However, the study was still conducted retrospectively because BACH does not have any patient IDs and uses different labeling schemes for each dataset. Cross-dataset validation can only be performed for individual native and transfer learning tasks, and not for a single clinical task.

6. Conclusion and Future Work

In this study, a novel attention-based dual-branch CNN-transformer pipeline is proposed for classifying breast histopathology images. It features bidirectional transfer and adaptive weighting between local features—like cellular morphology—and contextual features that encompass broader spatial organization. Cross-attention enhances information exchange between branches, optimized by adaptive weighting. Evaluations, including BACH multi-class and BreaKHis tasks, were conducted without artificial labels, emphasizing performance-computation cost trade-offs. Despite utilizing retrospective public datasets, the study faced challenges like merging results from different diagnostic taxonomies and the absence of patient-level evaluations. Future research will focus on generalizing the method across various datasets and improving interpretability and clinical applicability through hierarchical patch selection, domain adaptation, and lightweight architectures.

References

[1] Dasegowda, Mutthuraj, M. Brunda, M. Sajana, and Kanthesh M. Basalingappa. ”Breast Cancer Basics: Understanding the Disease and Its Challenges.” In Nano Theragnostics in

Breast Cancer: Advances, Challenges, and Future Prospects, Singapore: Springer Nature Singapore, 2026, 51-88.

[2] Verghese, Gregory, Jochen K. Lennerz, Danny Ruta, Wen Ng, Selvam Thavaraj, Kalliopi P. Siziopikou, Threnesan Naidoo et al. "Computational Pathology in Cancer Diagnosis, Prognosis, and Prediction–Present Day and Prospects." The Journal of pathology 260, no. 5 (2023): 551-563.

[3] Yan, Shen, Menglong Feng, and Yue Cai. "An Integrated Deep Learning Model with Enhanced EfficientNetB0 and MobileNetV1 for Diabetic Retinopathy Grading." Biomedical Signal Processing and Control 113 (2026): 108915.

[4] Wang, Yaoli, Yaojun Deng, Yuanjin Zheng, Pratik Chattopadhyay, and Lipo Wang. "Vision Transformers for Image Classification: A Comparative Survey." Technologies 13, no. 1 (2025): 32.

[5] Balasubramanian, Aadhi Aadhavan, Salah Mohammed Awad Al-Heejawi, Akarsh Singh, Anne Breggia, Bilal Ahmad, Robert Christman, Stephen T. Ryan, and Saeed Amal. "Ensemble Deep Learning-Based Image Classification for Breast Cancer Subtype and Invasiveness Diagnosis from Whole Slide Image Histopathology." Cancers 16, no. 12 (2024): 2222.

[6] Tummala, Sudhakar, Jungeun Kim, and Seifedine Kadry. "BreaST-Net: Multi-class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers." Mathematics 10, no. 21 (2022): 4109.

[7] Wang, Xiaoli, Guowei Wang, Luhan Li, Hua Zou, and Junpeng Cui. "MFF-ClassificationNet: CNN-transformer Hybrid with Multi-Feature Fusion for Breast Cancer Histopathology Classification." Biosensors 15, no. 11 (2025): 718.

[8] Ashoka, D. V. "Effivit: Hybrid CNN-Transformer for Retinal Imaging." Computers in Biology and Medicine 191 (2025): 110164.

[9] Mehmood, Shahid, Muhammad Zubair, Sagheer Abbas, Rowida Mohammed Alharbi, Mai Alduailij, Muhammad Adnan Khan, and Taher M. Ghazal. "Automated Bone Marrow Cell Classification Using Ensemble Learning: Performance, Generalization, And Clinical Interpretability." Frontiers in Medicine 13 (2026): 1787948.

[10] Jahanifar, Mostafa, Manahil Raza, Kesi Xu, Trinh Thi Le Vuong, Robert Jewsbury, Adam Shephard, Neda Zamanitajeddin et al. "Domain Generalization in Computational Pathology: Survey and Guidelines." ACM Computing Surveys 57, no. 11 (2025): 1-37.

[11] Vulli, Srinivas, Ming Liu, and Zhenyu Huang. "Reliability-Aware Cross-Modal Fusion of Sentinel-1 SAR and Sentinel-2 Optical Imagery for Robust Patch-Level Land Cover Scene Classification." Remote Sensing 18, no. 15 (2026): 2484.

[12] Tekale, Aditya, and Mohammad Masum. "Clinically Aligned Long-Context Transformers for Cross-Platform Mental Health Risk Detection." Electronics 15, no. 7 (2026): 1403.

[13] Santos, Luís Otávio, Aldo von Wangenheim, Felipe Soares Muylaert Barroso, Fabiana Botelho de Miranda Onofre, Alexandre Sherlley Casimiro Onofre, and Felipe Perozzo Daltoé. "A Systematic Literature Review on Vision Transformers Applications in Histopathology and Cytopathology: Advances in Cellular Analysis." (2024).

[14] Kuppusamy, Uma. "A Review of Current Imaging Techniques for Histopathology-Based Breast Cancer Diagnosis with Comparative Insights Using Machine Learning and Deep Learning Models and its Challenges and Future Directions." PeerJ Computer Science 12 (2026): e3645.

[15] Zhou, Xiaomin, Chen Li, Md Mamunur Rahaman, Yudong Yao, Shiliang Ai, Changhao Sun, Qian Wang et al. "A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks." IEEE Access 8 (2020): 90931-90956.

[16] Rakhlin, Alexander, Alexey Shvets, Vladimir Iglovikov, and Alexandr A. Kalinin. "Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis." In International conference image analysis and recognition, Cham: Springer International Publishing, 2018, 737-744.

[17] Liu, Suxing, Galib Muhammad Shahriar Himel, and Jiahao Wang. "Breast Cancer Classification with Enhanced Interpretability: Dalaresnet50 and Dt Grad-Cam." IEEE Access 12 (2024): 196647-196659.

[18] Baroni, Giulia Lucrezia, Laura Rasotto, Kevin Roitero, Angelica Tulisso, Carla Di Loreto, and Vincenzo Della Mea. "Optimizing Vision Transformers for Histopathology: Pretraining and Normalization in Breast Cancer Classification." Journal of Imaging 10, no. 5 (2024): 108.

[19] Yan, Yuting, Ruidong Lu, Jian Sun, Jianxin Zhang, and Qiang Zhang. "Breast Cancer Histopathology Image Classification Using Transformer with Discrete Wavelet Transform." Medical Engineering & Physics 138, no. 1 (2025): 104317.

[20] Sreelekshmi, V., Keechilat Pavithran, and Jyothisha J. Nair. "Swincnn: An Integrated Swin Transformer and CNN for Improved Breast Cancer Grade Classification." IEEE Access 12 (2024): 68697-68710.

[21] Wang, Xiaoli, Guowei Wang, Luhan Li, Hua Zou, and Junpeng Cui. "MFF-ClassificationNet: CNN-transformer Hybrid with Multi-Feature Fusion for Breast Cancer Histopathology Classification." Biosensors 15, no. 11 (2025): 718.

[22] Liu, Suxing, and Byungwon Min. "DCS-ST for Classification of Breast Cancer Histopathology Images with Limited Annotations." Applied Sciences 15, no. 15 (2025): 8457.

[23] Mahmood, Tariq, Tanzila Saba, Shaha Al-Otaibi, Noor Ayesha, and Ahmed S. Almasoud. "AI-Driven microscopy: Cutting-Edge Approach for Breast Tissue Prognosis Using Microscopic Images." Microscopy Research and Technique 88, no. 5 (2025): 1335-1359.

[24] Yao, Xiang, and Fujun Liu. "Weak Feature Masking for Fine-Grained Breast Cancer Histopathological Image Classification Under Small Dataset Constraints." Biomedical Signal Processing and Control 126 (2026): 111001.