Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

RFPN-Net: A Multi-Layer Attention and Cross-Scale Fusion Method for Breast Histopathological Image Classification

Shuai Li
Changchun University
Xiaoli Wang
wangxl@ccu.edu.cn
Changchun University
Guowei Wang
Changchun University
Luhan Li
Tianjin University
Junpeng Cui
Changchun University
Qingxing Su
Changchun University

Research Article

Keywords:

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11717956/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Article type: Research paper

RFPN-Net: A Multi-Layer Attention and Cross-Scale Fusion Method for Breast Histopathological Image Classification

Shuai Li[1], Xiaoli Wang[1,*], Guowei Wang[1], Luhan Li[2], Junpeng Cui[1], Qingxing Su[1]

1 Electronic Information Engineering College, Changchun University, Changchun 130022, China

2 School of Electrical Automation and Information Engineering, Tianjin University, Tianjin 300072, China

Corresponding author: Xiaoli Wang, wangxl@ccu.edu.cn

Abstract

Accurate breast cancer subtype classification is crucial for personalized treatment and prognosis, yet subtle morphological differences and complex multi-scale features in histopathological images remain challenging. To address this, we propose RFPN-Net, a multi-layer attention and cross-scale fusion network. RFPN-Net integrates a Multi-layer Feature Extraction Module (MFEM) and a Cross-Scale Feature Fusion Module (CSFM). MFEM employs an improved ResNet50 with embedded Convolutional Block Attention Modules (CBAM) to hierarchically extract multi-scale features while enhancing critical region representation. CSFM utilizes a Feature Pyramid Network to propagate high-level semantics top-down and incorporates a Swin Transformer for cross-scale global modeling, effectively capturing long-range dependencies. Focal Loss is applied to handle class imbalance. Experiments on the BreakHis dataset with five-fold cross-validation show that RFPN-Net achieves superior performance: average accuracies reach 98.68% at 40× and 98.58% at 400×, with all core metrics above 98% and extremely small standard deviations across all four magnifications. The model effectively balances local details and global semantics, exhibiting strong multi-scale fusion, robustness, and generalization, thereby providing reliable technical support for computer-aided precise diagnosis of breast cancer subtypes.

Keywords: breast cancer subtypes; multi-layer attention; cross-scale fusion; Swin Transformer

1 Introduction

As the global cancer burden continues to escalate, early and precise diagnosis of breast cancer—the second most common cancer worldwide—is crucial for reducing mortality [1]. Histopathological analysis serves as the gold standard for breast cancer diagnosis [2]; however, manual slide review is inherently time-consuming and constrained by subjective experience and diagnostic workload. In recent years, the application of computer-aided diagnosis (CAD) systems has significantly improved diagnostic accuracy and alleviated the burden on pathologists [3, 4]. Histopathological images enable the fine-grained resolution of complex cellular architectures and microenvironmental interactions, offering a critical foundation for disease research and clinical diagnosis [3]. Nevertheless, the inherent complexity of tissue morphology, coupled with imaging variables such as staining density and section flatness, continues to pose substantial challenges to high-quality clinical pathological analysis [5]. Manual pathological diagnosis is inherently limited by subjective experience and visual fatigue, making it challenging to guarantee diagnostic efficiency and consistency. Consequently, there is an urgent need to leverage deep learning for the automated and efficient analysis of histopathological images to overcome existing diagnostic bottlenecks. However, as the cornerstone of CAD systems, the annotation granularity of a dataset directly determines the upper bound of model performance. Constrained by limited annotations and model architecture designs, existing research has predominantly focused on binary classification between benign and malignant tumors; yet, in clinical practice, the precise identification of specific cancer subtypes and benign lesion variants is crucial for guiding personalized treatment [7]. Furthermore, because whole-slide images (WSIs) are massive in scale and extremely computationally exhaustive, conventional patch-based processing methods frequently suffer from severe misclassification due to the presence of localized benign tissue mixtures within malignant slides [8, 9]. To address this, Spanhol et al. [6] introduced the BreakHis dataset, which overcomes the limitation of monolithic annotations by providing eight fine-grained subtype labels across multiple magnification factors. Benefiting from its more accurate and comprehensive annotation information, BreakHis has emerged as a highly promising benchmark dataset for advancing multi-class classification research in breast histopathology.

Since the release of this dataset, feature extraction techniques in this field have undergone a profound evolution. In the early stages, research primarily relied on hand-crafted features [10–12, 23], which were labor-intensive and required extensive prior domain knowledge. Representative methods included image classification based on KAZE keypoint mapping and Support Vector Machines (SVM) [26], as well as adaptive sparse SVM feature recognition approaches incorporating weighted L_1 regularization [27]. However, due to significant textural variations across different magnification levels, the manual design and selection of cross-magnification features are not only cumbersome but also often yield suboptimal generalization performance. To overcome this bottleneck, researchers have increasingly turned to Convolutional Neural Networks (CNNs), which offer stronger semantic representational capabilities and adaptive, data-driven parameter optimization. Currently, various paradigms—including transfer learning via pre-trained models [8, 14], ensemble voting mechanisms based on multiple CNNs [19], BiCNN models that leverage classification label priors to guide feature distance learning [29], FE-BkCapsNet methods integrating the strengths of capsule networks with CNNs [15], and customized lightweight architectures [16]—have all demonstrated superior classification accuracy. Nevertheless, most of these advanced methods remain restricted to binary classification between benign and malignant tumors, failing to provide the specific disease subtype diagnoses that multi-class classification methods [24] can offer.

Driven by the clinical significance of subtype multi-class classification, extensive and in-depth research has been conducted based on the BreakHis dataset, leading to the maturation of various technical frameworks. Early attempts at multi-class classification encompassed experiments involving the fine-tuning of the final residual blocks in ResNet-18 [8], comparative analyses between custom-built CNNs and ensemble machine learning models [11], and the application of autoencoder technologies that combine transfer learning from ImageNet [22] pre-trained InceptionResNetV2 with K-Means clustering [21]. In recent years, research efforts have increasingly shifted toward innovative architectures that emphasize feature fusion, relational guidance, and network lightweighting. Notable examples include the lightweight DBLCNN, which utilizes dependency relationships to guide subtype feature learning while significantly reducing parameter counts; the improved EfficientNetV2 model, which integrates fused mobile inverted bottleneck structures with dual attention mechanisms [20]; hybrid classification methods that synergize deep semantic features with three-channel GLCM texture features [25]; and end-to-end CSDCNNs [6] that incorporate various class-specific priors and multi-layered deep structures (partially inspired by GoogLeNet [28]). These pioneering explorations have collectively advanced the field, resulting in significant improvements in classification accuracy, model robustness, and computational efficiency for breast histopathological multi-class classification.

Although existing multi-class classification networks have demonstrated state-of-the-art recognition performance, further enhancing multi-scale contextual representation and capturing global fine-grained dependencies in cross-magnification and cross-scale pathological slices remains a critical bottleneck to overcome. In recent years, studies in object detection and semantic segmentation [17, 18] have consistently demonstrated that multi-scale feature representation plays a decisive role in enhancing model robustness and discriminative power. Inspired by these insights, this paper proposes a hierarchical multi-scale feature fusion framework that synergistically integrates and improves ResNet50, the Feature Pyramid Network (FPN), and the Swin Transformer. This architecture not only enhances cross-level contextual representations through FPN-based fusion, but also leverages the window-based self-attention mechanism of the Swin Transformer to computationally efficiently model both local and global fine-grained structural dependencies. Ultimately, by cascading the outputs of multi-

layer Transformer branches into a classification head, the framework fully exploits the complementary advantages of features across different scales. The proposed model achieves outstanding multi-class classification performance on the public BreakHis [6] dataset. Furthermore, extensive experiments and evaluations on the BACH dataset [34] robustly validate its superior recognition accuracy and generalization capability.

The main contributions of this paper are summarized as follows:

(1) To address the issue that conventional ResNet50 pays insufficient attention to critical regions and extracts inadequate discriminative information from breast histopathological images, we incorporate a channel and spatial dual attention mechanism (CBAM) into the Bottleneck Block. This enhancement adaptively emphasizes key lesion areas while suppressing redundant background features, thereby significantly improving the precision of feature extraction and the discriminative power of the network.

(2) To overcome the difficulty of single-scale models in simultaneously capturing local details and global structural information, we construct a bottom-up multi-scale feature pyramid structure for efficiently fusing features from different depths. This design allows the network to fully leverage cross-scale complementary information, thereby enhancing the model's generalization ability and robustness on multi-magnification histopathological images.

(3) Given the inherent limitations of CNNs in modeling long-range dependencies and global structural relationships, we introduce Swin Transformer modules after multi-scale feature fusion. Its sliding-window-based self-attention mechanism efficiently captures both local and global contextual associations, improving structural pattern recognition and consequently boosting the overall classification performance.

2 Methods

2.1 Overall architecture

RFPN-Net is divided into two main components: the Multi-level Feature Extraction Module (MFEM) and the Cross-Scale Feature Fusion Module (CSFM). The MFEM adopts ResNet50 embedded with the CBAM attention mechanism as its backbone, which comprises four primary feature stages (Stage1–Stage4), corresponding to the Conv2x, Conv3x, Conv4x, and Conv5x outputs of the backbone network. This module focuses on progressively enhancing the feature abstraction capability, thereby providing rich and discriminative feature representations for the subsequent multi-scale fusion.The Cross-Scale Feature Fusion Module is constructed based on the Feature Pyramid Network (FPN). Through a top-down feature propagation pathway and lateral connections, it effectively compensates low-level detail features with high-level semantic information, resulting in semantically consistent and scale-aligned multi-scale feature representations. The multi-scale features produced by the FPN are then fed into the Swin Transformer module. This module consists of layer normalization (LN), window-based multi-head self-attention (W-MSA), and a multi-layer perceptron (MLP). While retaining the strengths of local window modeling, it enables cross-window information interaction, thus significantly enhancing the model's ability to capture global context and long-range dependencies.

By integrating the multi-scale feature fusion capability of CNNs with the global dependency modeling capability of Transformers, RFPN-Net achieves cross-scale semantic consistency modeling, further strengthening the holistic perception of critical lesion regions, and ultimately forming a unified deep feature representation with stronger discriminative power. The fused semantic features are subsequently passed to the classifier to perform subtype classification of breast cancer histopathological images. During training, the Focal Loss function is adopted to alleviate the class imbalance issue and improve the overall robustness and classification performance of the model. The architecture of RFPN-Net is illustrated in Fig. 1.

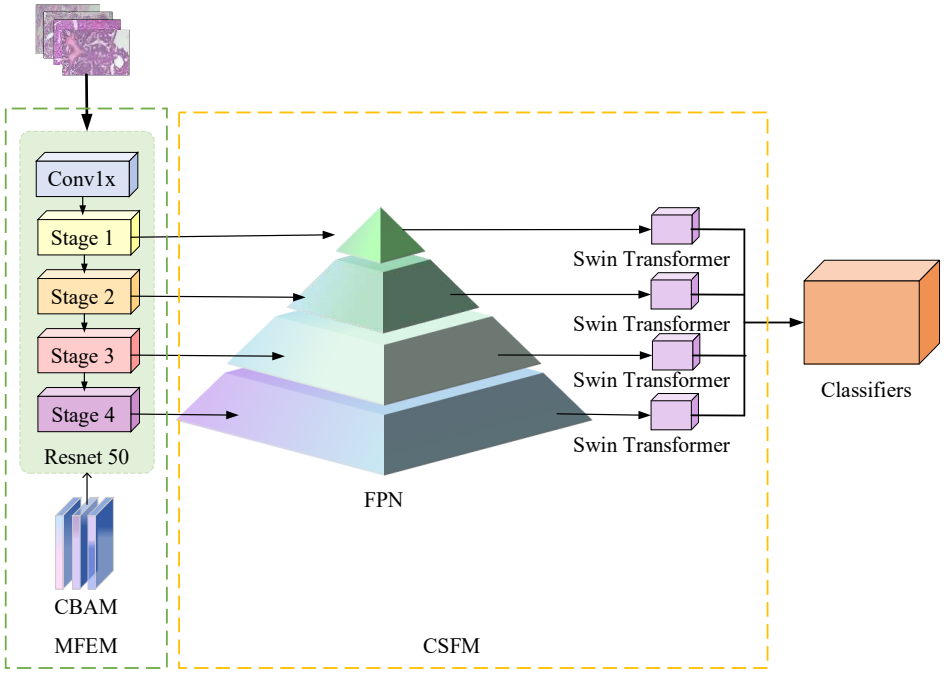


Fig. 1 Architecture of RFPN-Net

2.2 Multi-level feature extraction module

The multi-level feature extraction module adopts ResNet-50 as the core component, stacking residual structures with progressive downsampling to extract multi-scale pathological features in a fine-to-coarse manner from shallow to deep layers. This module comprises four main feature stages (Stage 1 to Stage 4). During feature propagation, each stage progressively expands the receptive field through stacked residual blocks and strided convolutions, enabling the network to gradually abstract low-level visual information into high-level semantic representations at different hierarchies. As a result, the module provides hierarchically rich feature representations for subsequent cross-scale feature fusion and classification.

Stage 1 extracts basic visual features from the input image, preserving fine-grained information such as local textures, cell edges, and glandular contours in breast cancer pathological images, which can be used to characterize the low-level structural properties of lesion regions. Stage 2 and Stage 3 enlarge the receptive field to integrate larger local contextual information while maintaining a certain spatial resolution, thereby capturing mid-level structures in breast cancer, including cell arrangement patterns, stromal distribution, and local morphological variations of lesions, and providing crucial structural information for breast cancer subtype recognition. Stage 4, situated in the deeper layers of the network, employs multi-layer residual learning to perform high-level semantic abstraction of lesion regions, yielding highly discriminative high-level semantic features that reflect the global attributes of lesions and the intrinsic characteristics of categories, and serving as the primary basis for achieving high-accuracy classification of breast cancer pathological images. The architecture of the multi-level feature extraction module is illustrated in Fig. 2.

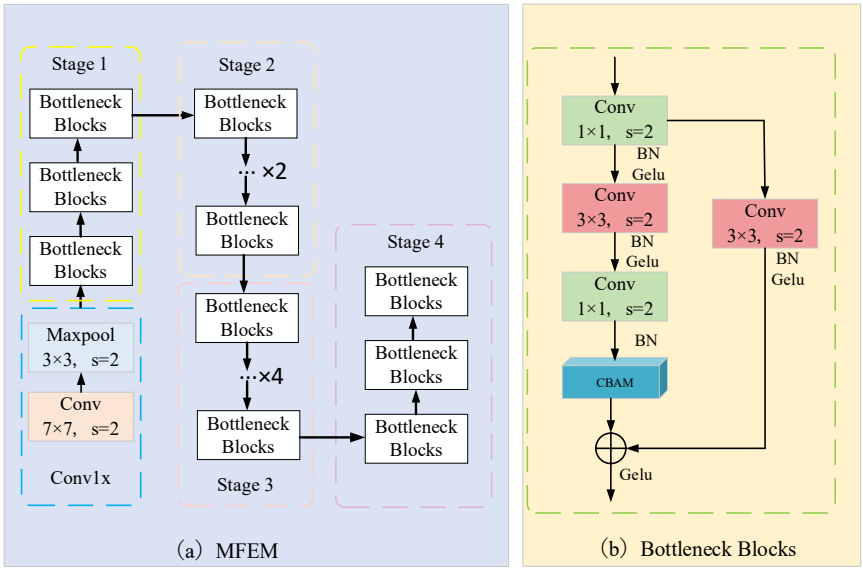


Fig. 2 Architecture of the multi-level feature extraction module

2.2.1 CBAM attention mechanism

CBAM is a lightweight and general-purpose plug-and-play attention module, as well as an attention mechanism designed to enhance the modeling capability of convolutional neural networks (CNNs). It combines the Channel Attention Module (CAM) and the Spatial Attention Module (SAM) in a sequential manner, enabling the network to selectively emphasize task-relevant features along both channel and spatial dimensions, thereby improving feature extraction effectiveness and optimizing the feature space. The CBAM attention module is illustrated in Fig. 3.

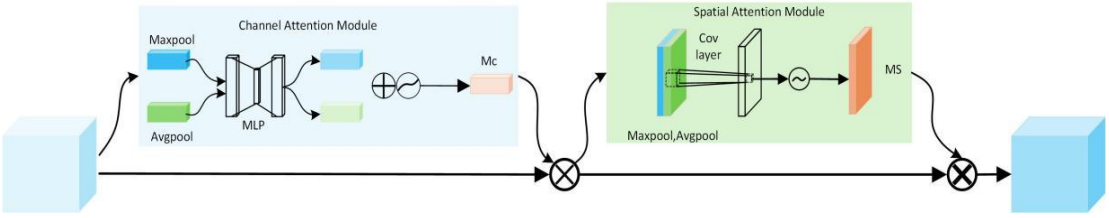


Fig. 3 Architecture of the convolutional block attention module (CBAM)

CAM is primarily designed to capture global inter-channel relationships and dynamically adjust the importance of different channels, thereby enhancing feature representational capacity. It consists of a global average pooling layer, a global max pooling layer, and a multi-layer perceptron (MLP) composed of two fully connected layers. Specifically, the input feature map first undergoes max pooling and average pooling operations separately. The pooled results are then processed by the MLP with two fully connected layers to learn a weight for each channel. Finally, the obtained attention weights are multiplied with the original feature map to selectively enhance or suppress features along the channel dimension. The computation process of CAM is as follows. By dynamically recalibrating channel-wise importance, CAM enables the network to focus more on key features within the image, thus improving the representational power of image features.

$$M_C(F) = \sigma(MLP(Avgpool(F)) + MLP(Maxpool(F))) \tag{1}$$

$$F' = M_C(F) \times F \tag{2}$$

Here, F denotes the input feature map; Avpool and Maxpool represent global average pooling and global max pooling, respectively; MLP refers to a multi-layer perceptron composed of two fully connected layers; σis the Sigmoid activation function; M_c represents the channel attention module (or channel attention weights); anddenotes the feature map obtained after the channel attention module.

The Spatial Attention Module (SAM) is primarily designed to capture correlations among different spatial positions, enabling the network to better focus on critical regions within the image. SAM comprises a 7×7 convolutional layer, a global average pooling layer, a global max pooling layer, and an MLP structure consisting of two fully connected layers. First, the input feature map is processed by global average pooling and global max pooling. The outputs of the two pooling operations are then concatenated along the channel dimension, after which a convolution operation is applied to reduce the channel dimension to 1, thereby extracting spatial information. The result is subsequently processed by the MLP with two fully connected layers. Finally, the obtained spatial attention weights are multiplied with the original feature map to produce the weighted feature representation. The computational process of SAM is defined by Equations (3) and (4). By dynamically adjusting the importance of different spatial locations, SAM enables the network to attend more effectively to informative regions within the image, thereby improving the representational capacity of image features.

$$M_S(F') = \sigma(f^{7\times7}([Avgpool(F'); Maxpool(F')])) \tag{3}$$

$$F'' = M_S(F') \times F' \tag{4}$$

denotes the spatial attention module;denotes the 7×7 convolution operation; anddenotes the feature map obtained after the spatial attention module.

The CBAM attention mechanism integrates channel attention and spatial attention, enabling comprehensive capture of both global and local relationships within image features. By dynamically adjusting the importance of different channels and spatial positions, CBAM enhances the extraction and utilization of key features in images, thereby strengthening the network's representational capacity and generalization ability. Furthermore, CBAM is highly compatible with existing CNN architectures and can be easily integrated into various network structures, offering a simple yet effective means to improve the performance of deep learning models.

2.3 Cross-scale feature fusion module

To address the pronounced scale variation of lesion regions and the insufficient relational modeling among features at different semantic levels in pathological images, this paper proposes a cross-scale fusion network. The network comprises a multi-scale feature pyramid module and a self-attention-based feature modeling module, with the overall architecture illustrated in Fig. 4. It is designed to fuse and model features across different spatial scales and semantic levels, thereby enabling effective cross-scale information interaction. The feature pyramid module takes the outputs from the four stages of the backbone residual network as input to construct multi-scale feature representations. Specifically, it receives feature maps from different stages of the encoder network and progressively propagates high-level semantic information to lower-level feature maps through stepwise upsampling and lateral connections. This process constructs a semantically consistent multi-scale feature set, achieving fused representation of multi-scale information.

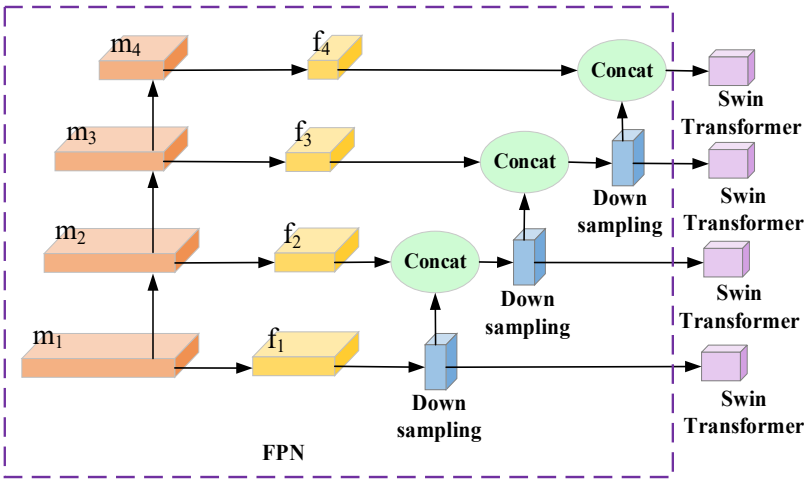


Fig. 4 Architecture of the cross-scale feature fusion module (CSFM)

Based on the outputs of the feature pyramid module, a feature modeling module leveraging window-based self-attention is introduced to further model the fused multi-scale features. Each scale of features output by the FPN is processed through an independent Swin Transformer branch to fully capture long-range dependencies within the features at different scales. A progressive cross-scale fusion strategy is adopted, where features from adjacent scales are downsampled and concatenated (Concat) to achieve hierarchical cross-scale information interaction, thereby enhancing the semantic consistency among multi-scale features.

2.3.1 Feature pyramid network

To address the significant variations in spatial scale, morphological structure, and tissue distribution of lesion regions in pathological images—where single-scale features often struggle to capture both local details and high-level semantic representations simultaneously, leading to insufficient recognition of small lesions or complex structures—an FPN module is introduced into the cross-scale feature extraction network. Its core role is to receive feature maps from the different stages of the Multi-level Feature Extraction Module (MFEM). Through a top-down multi-scale fusion strategy, high-level semantic information is propagated backward to lower-level features, while lateral connections are employed to preserve fine-grained details from low-level features. This design ultimately constructs a multi-scale feature pyramid with semantic consistency and complementary resolutions, effectively bridging the semantic gap between different feature levels in traditional CNNs. The specific implementation of this pyramid structure is illustrated in Fig. 5.

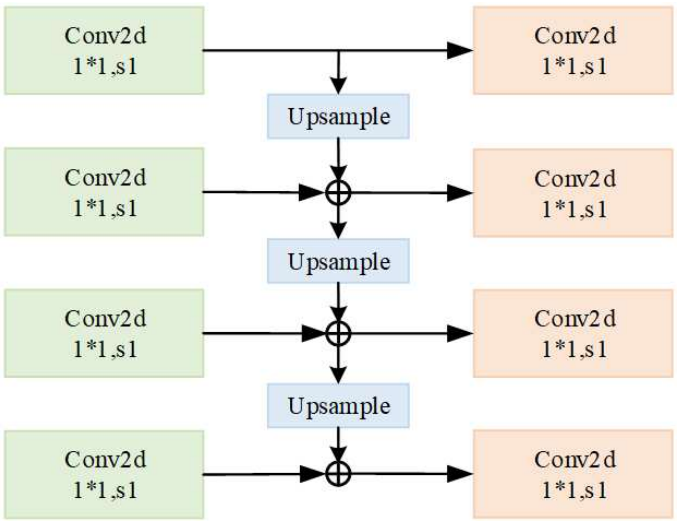


Fig. 5 Detailed architecture of the feature pyramid sub-module (FPN)

Specifically, the FPN receives the feature mapsoutput from Stage1 to Stage4 of the MFEM. Deep features (e.g., M4) contain rich semantic information but suffer from low spatial resolution, making it difficult to capture subtle lesion details and local structures. In contrast, shallow features (e.g., M1) possess high spatial resolution and abundant local texture and edge information, yet have limited semantic discriminative capability. To leverage their complementary strengths, the FPN adopts a fusion strategy of "top-down upsampling + lateral connection alignment." First, in the top-down pathway, starting from the deepest feature map (M4), upsampling is progressively performed to propagate high-level semantic information to shallower layers. Meanwhile, a 1×1 convolution is applied to the original feature map of each stage for channel alignment, ensuring that the channel dimensions match those of the upsampled deep features. The aligned shallow feature maps are then summed element-wise with the upsampled deep features, thereby effectively fusing fine-grained shallow details with high-level semantic representations. Through this mechanism, the fused feature maps not only retain fine details from shallow layers but also incorporate rich semantic information from deep layers, providing high-quality, multi-scale input features for the subsequent Swin Transformer-based cross-scale modeling. This enables RFPN-Net to simultaneously capture local lesion details and global contextual information, ultimately improving the multi-subtype classification performance of breast cancer histopathological images.

Letdenote the original feature map of thestage, andthe corresponding fused pyramid feature obtained after FPN fusion. The specific computation process can be expressed as follows:

$$P_l = conv(M_l) + Upsample(P_{l+1}) \tag{5}$$

Here,conv(·)denotes the 1×1 convolution operation, which primarily serves to unify the channel dimensions of the original feature maps from different levels to the same dimension, thereby avoiding fusion errors caused by channel mismatch.Upsample(·)denotes the bilinear upsampling operation, which upsamples the pyramid featureat levelto match the resolution of the output feature map at level, ensuring effective element-wise fusion between the two. Through this fusion strategy, the FPN ultimately produces a set of multi-scale pyramid features, each corresponding to pathological information at different resolutions. Among them,focuses on microscopic cellular details,focuses on macroscopic tissue semantics, and the intermediate-level features realize a gradual transition between local details and high-level semantics, thereby providing a hierarchical feature foundation for subsequent global modeling.

2.3.2 Swin Transformer architecture

After the multi-scale feature pyramid completes the fusion of features from different levels, a cross-scale feature modeling module based on Window-based Multi-Head Self-Attention (W-MSA) is further introduced to enhance the network's capacity for modeling long-range dependencies and global contextual information. Traditional convolution operations are constrained by local receptive fields, which limits their ability to capture cross-region and cross-scale structural correlations in pathological images. In contrast, the Transformer architecture based on self-attention mechanisms can effectively address this limitation. The fused feature representations output by the feature pyramid module are denoted as:

$$Z^{l-1} \in \mathbb{R}^{H \times W \times C} \tag{6}$$

This feature modeling module adopts the basic structure of the Swin Transformer Block, whose core consists of a Window-based Multi-Head Self-Attention (W-MSA) module and a feed-forward network (MLP). Layer Normalization (LN) is applied before each sub-module, and residual connections are employed to ensure stable feature propagation. The architecture of the Swin Transformer sub-module is illustrated in Fig. 6, and its computation process can be expressed as:

$$Z^{l} = Z^{l-1} + W - MSA(LN(Z^{l-1})) \tag{7}$$

$$Z^{l+1} = Z^{l} + MLP(LN(Z^{l})) \tag{8}$$

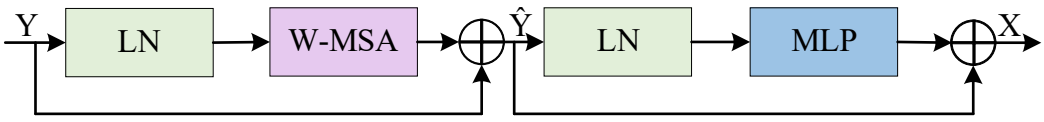


Fig. 6 Detailed architecture of the Swin Transformer sub-module

Here, W-MSA denotes the Window-based Multi-Head Self-Attention operation performed within local windows, designed to model the correlations among features inside each window. Specifically, the input features are first linearly projected to generate Query (Q), Key (K), and Value (V) vectors, after which self-attention is computed within each window. Its basic form is expressed as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V \tag{9}$$

where ddenotes the feature dimension, and B is the relative position bias term, which encodes the relative spatial relationships between different positions within a window. By employing a multi-head parallel computation strategy, the attention results from different subspaces are fused, thereby enhancing the diversity and discriminability of the feature representations. After the self-attention computation, the features are further fed into a feed-forward network (MLP) to enhance the non-linear expressive capacity, which can be formulated as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V \tag{10}$$

where$\sigma(\cdot)$enotes a nonlinear activation function. Through the above W-MSA-based feature modeling process, the network is able to effectively capture the structural correlations embedded in the multi-scale fused features while preserving local detail sensitivity. This further enhances the representation capability of critical lesion regions and provides more discriminative feature support for the subsequent classification decisions.

2.4 Experimental setup

2.4.1 Training configuration

Based on histopathological images, both the BreakHis dataset and the BACH dataset are selected to evaluate the model's performance in classification tasks; the diversity of these datasets is expected to help demonstrate the robustness of the model. A total of 80% of the data is used for training, while the remaining 20% is further split into two equal halves, resulting in 10% for validation and 10% for testing. All experiments are conducted on a device equipped with an NVIDIA GPU (RTX 4060, 8 GB) running the Windows operating system, with Python 3.7, PyTorch 1.13.1, and CUDA 11.6 installed. The initial learning rate of RFPN-Net is set to 0.0001, the batch size is set to 16, and the Adam optimizer is adopted to adaptively update the model parameters. All training images are resized to 224 × 224 pixels, and the model is trained for 100 epochs.

2.5 Evaluation metrics

To comprehensively evaluate the model's performance in multi-class classification tasks, the following evaluation metrics are adopted: Accuracy, Precision, Recall, and F1-Score. In addition, the Receiver Operating Characteristic (ROC) curve is plotted and the Area Under the Curve (AUC) is computed. All these metrics are derived from the confusion matrix. The relevant notations are defined as follows: True Positive (TPk) represents the number of samples that truly belong to class k and are correctly predicted as class k; False Positive (FPk) represents the number of samples that do not belong to class k but are incorrectly predicted as class k; False Negative (FNk) represents the number of samples that belong to class k but are predicted as not k; True Negative (TNk) represents the number of samples that neither belong to class k nor are predicted as class k.

Accuracy is calculated using Equation (11) to obtain the percentage of correctly identified samples.

$$Accuracy = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} (TP_k + FP_k + FN_k + TN_k)} \tag{11}$$

Precision is calculated using Equation (12), which represents the proportion of samples that truly belong to a class among all samples predicted as that class, reflecting the accuracy of the model's positive predictions.

$$Precision = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} (TP_k + FP_k)} \tag{12}$$

Recall is calculated using Equation (13), which represents the proportion of actual positive samples that are correctly identified, reflecting the comprehensiveness of the model's predictions.

$$Recall = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} (TP_k + FN_k)} \tag{13}$$

The F1-score for each class is defined in Equation (14). The F1-score addresses the trade-off between precision and recall, with a higher value indicating better performance.

$$\text{F1-Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

For class k, it is treated as the positive class, and all other classes are treated as negative. The False Positive Rate (FPRk) and the True Positive Rate (TPRk) for this class are then computed. The calculation methods are defined in Equations (15) and (16), respectively.

$$FPR_k = \frac{FP_k}{FP_k + TN_k} \tag{15}$$

$$TPR_k = \frac{TP_k}{TP_k + FN_k} \tag{16}$$

2.6 Datasets

2.6.1 BreakHis dataset

Spanhol et al. [6] publicly released the BreakHis dataset, a widely used and publicly available benchmark for breast cancer histopathological image classification. It comprises microscopic biopsy images of benign and malignant breast tumors, covering eight fine-grained breast tumor subtypes. Specifically, the four benign categories include adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA); the four malignant categories include ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). The dataset was collected at the P&D Laboratory (Brazil), to which all patients were referred for clinical diagnosis of breast cancer, and all data were anonymized. Representative images are shown in Fig. 7.

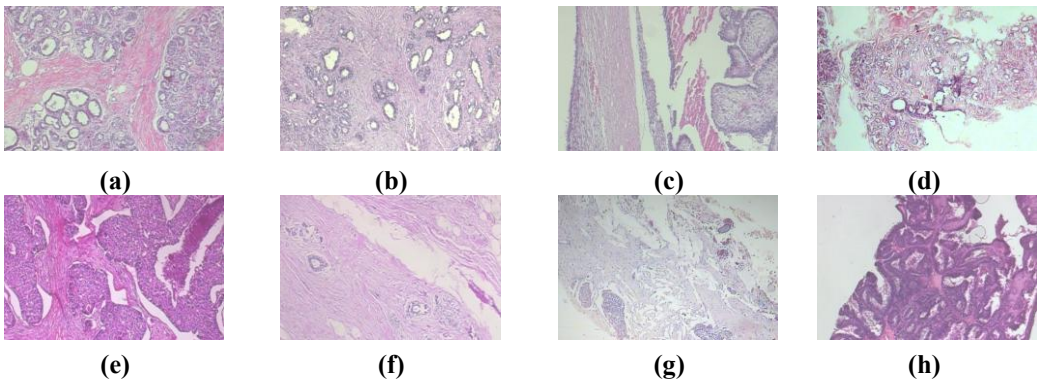


Fig. 7 40× histopathology sample images from the BreakHis dataset; (a) adenosis, (b) fibroadenoma, (c) phyllodes tumor, (d) tubular adenoma, (e) ductal carcinoma, (f) lobular carcinoma, (g) mucinous carcinoma, and (h) papillary carcinoma

The breast cancer samples were obtained through surgical biopsy and annotated by expert pathologists. All images have a resolution of 700 × 460 pixels and are divided into four magnification levels: 40×, 100×, 200×, and 400×. The dataset contains a total of 7,909 images; however, the original distribution is severely imbalanced across subtypes and magnifications. For example, at 40× magnification, the ductal carcinoma category contains 864 images, whereas adenosis has only 114 images.

To alleviate the problem of class imbalance, data augmentation was applied to the original images. The specific augmentation operations included random rotation, random horizontal flipping, and random vertical flipping. These augmentation techniques effectively increased the sample size, improved model robustness, and resulted in a more balanced class distribution. Supplementary Fig. S1 illustrates representative original images and their corresponding augmented counterparts. Table 1 presents the data distribution before and after augmentation, with the numbers in parentheses indicating the post-augmentation counts. As can be seen from Table 1, the augmented dataset exhibits a more uniform distribution, thereby providing

a more stable data foundation with stronger generalization capability for subsequent model training.

Table 1 Distribution of image counts in the BreakHis dataset before and after augmentation

magnification	benign				malignant			
	A	F	PT	TA	DC	LC	MC	PC
40×	114 (798)	253 (759)	109 (763)	149 (894)	864 (864)	156 (780)	205 (820)	145 (870)
100×	113 (791)	260 (780)	70 (874)	150 (750)	903 (903)	170 (850)	222 (888)	142 (852)
200×	111 (777)	264 (792)	108 (756)	140 (840)	896 (896)	163 (815)	196 (784)	135 (810)
400×	106 (742)	237 (711)	115 (815)	130 (910)	788 (788)	137 (822)	169 (845)	138 (828)
overall	444 (3108)	1014 (3042)	402 (3198)	569 (3544)	3451 (3451)	626 (3267)	792 (3337)	560 (3360)

2.6.2 BACH dataset

To promote the advancement of digital pathology, two major Portuguese research institutions, Ipatimup and INEB, jointly constructed a histological image dataset and organized the Breast Cancer Histology Challenge (BACH) based on this dataset, with the aim of supporting research on the automatic classification of breast cancer malignancy grade. The dataset comprises hematoxylin and eosin (H&E)-stained histopathological images of breast cancer, containing a total of 400 sample images. All images were digitized under uniform acquisition conditions at a magnification of 200×. Representative examples are shown in Supplementary Fig. S2.

Based on histological characteristics, the BACH dataset classifies samples into four types: normal tissue (Normal, N), benign lesion (Benign, B), in situ carcinoma (In Situ Carcinoma, ISC), and invasive carcinoma (Invasive Carcinoma, IVC). Among them, normal tissue and benign lesions are typically categorized as non-cancerous, while in situ carcinoma and invasive carcinoma are classified as cancerous. Each category contains 100 images, yielding a total of 400 images. The detailed distribution is shown in Table 2.

Table 2 Image distribution of the BACH dataset

Category	Subtype	Number of Images	Total
Non-Carcinoma	Normal	100	200
	Benign	100	
Carcinoma	In Situ Carcinoma	100	200
	In vasive Carcinoma	100	

3 Results

3.1 Five-fold cross-validation

To evaluate the effectiveness of the proposed model, a 5-fold cross-validation study was conducted. This approach not only helps assess the generalization capability of the model, but also allows us to identify potential overfitting, particularly in the case of small datasets. In this experiment, all dataset samples were utilized for training and validation at different

stages, thereby enabling a thorough evaluation of the model's performance. We performed five runs of cross-validation to assess the model's ability to handle class imbalance. The dataset was divided into 80% for training, 10% for validation, and 10% for testing. At the end of the five training runs, the mean and standard deviation of the performance metrics were computed, providing a comprehensive performance evaluation. Table 3 summarizes the results, demonstrating that the performance achieved through 5-fold cross-validation is consistent with that of the optimal model, highlighting the resilience of the proposed model in addressing class imbalance while maintaining satisfactory performance. By averaging the TP, FP, TN, and FN across all folds, the confusion matrices for both datasets under the 5-fold evaluation were generated as shown in the figures.

Table 3 Performance of five-fold cross-validation

40×						
Fold	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Specificity**	**AUC**
Fold1	98.78	98.77	98.79	98.78	99.65	98.90
Fold2	98.94	98.95	98.94	98.94	99.70	98.78
Fold3	98.48	98.52	98.54	98.53	99.61	98.63
Fold4	98.63	98.69	98.63	98.65	99.69	98.92
Fold5	98.63	98.67	98.69	98.68	99.61	98.50
Mean±STD	98.68±0.16	98.72±0.14	98.72±0.14	98.72±0.14	99.65±0.04	98.75±0.18
100×						
Fold1	99.11	99.16	99.17	99.16	99.57	97.88
Fold2	98.50	98.54	98.60	98.57	99.66	98.43
Fold3	97.75	97.87	97.84	97.85	99.52	98.03
Fold4	98.50	98.56	98.57	98.57	99.78	98.46
Fold5	98.35	98.45	98.46	98.45	99.61	98.11
Mean±STD	98.44±0.43	98.52±0.41	98.53±0.42	98.52±0.42	99.63±0.10	98.18±0.25
200×						
Fold1	98.46	98.51	98.50	98.51	99.69	98.67
Fold2	97.85	97.90	97.94	97.92	99.56	97.59
Fold3	98.92	98.95	99.00	98.96	99.74	98.87
Fold4	98.31	98.39	98.37	98.37	99.74	98.19
Fold5	97.69	97.72	97.80	97.73	99.73	97.67
Mean±STD	98.24±0.44	98.29±0.44	98.32±0.43	98.31±0.43	98.30±0.07	98.20±0.57
400×						
Fold1	98.92	98.95	98.87	98.91	99.82	98.92
Fold2	98.31	98.32	98.28	98.30	99.69	98.72
Fold3	98.30	98.36	98.23	98.29	99.87	98.16
Fold4	98.61	98.69	98.56	98.62	99.78	98.45
Fold5	98.76	98.83	98.74	98.75	99.82	98.56
Mean±STD	98.58±0.28	98.64±0.27	98.53±0.28	98.59±0.28	99.80±0.07	98.58±0.29

The five-fold cross-validation results on the BreakHis dataset at four magnification levels (40×, 100×, 200×, and 400×) demonstrate that our method achieves highly stable and consistent performance across all metrics. Specifically, at 40× magnification, the model attains the highest average accuracy (98.68% ± 0.16), while the mean Precision, Recall, and F1-

Score all remain at 98.72%, with extremely low standard deviations, reflecting strong consistency. At 100× and 200× magnifications, the average accuracies are 98.44% ± 0.43 and 98.24% ± 0.44, respectively. Although the fluctuations of the metrics slightly increase, the overall performance remains at a high level above 98%. At 400× magnification, the overall performance is equally stable, with an average accuracy of 98.58% ± 0.28. Overall, the model exhibits excellent generalization capability and robustness across different magnification levels, indicating that the proposed method can effectively adapt to the texture and structural variations in histopathological images caused by changes in magnification.

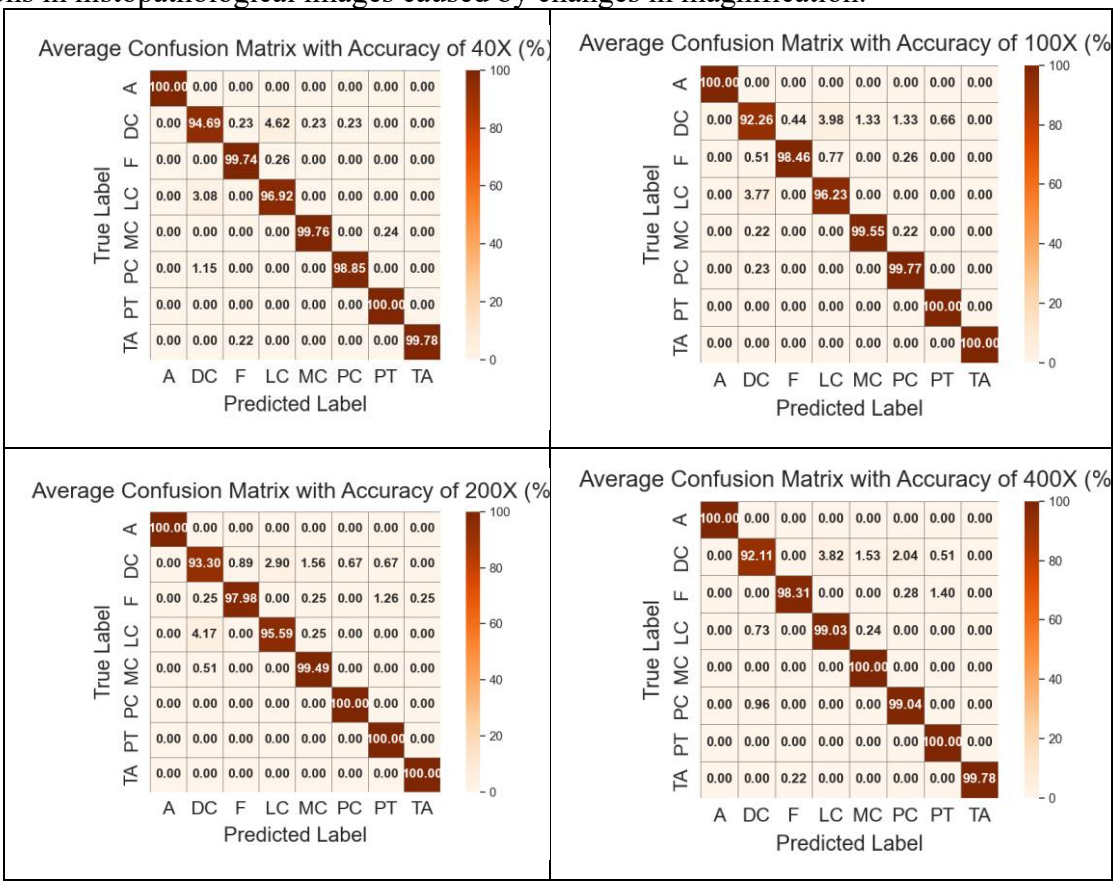


Fig. 8 Confusion matrices of RFPN-Net for breast cancer subtype classification on the BreakHis dataset at different magnification levels

Fig. 8 presents the average confusion matrices at four magnification levels (40×, 100×, 200×, and 400×). At 40× magnification, the classification accuracy for almost all categories reached or exceeded 99%, with only a few categories exhibiting minor misclassifications, indicating that the model effectively extracts global tissue structural features from low-magnification images and thereby achieves high accuracy. At 100× magnification, the confusion matrix still displays strong diagonal dominance, with very few categories showing slight confusion, suggesting robust classification performance at moderate magnification. At 200× magnification, the increased complexity of cellular structures led to some misclassification between similar categories, such as F and LC; however, the recognition rate for most categories remained high, demonstrating the model's strong capability in handling complex features. At 400× magnification, the confusion matrix again exhibits a well-defined diagonal distribution, with the recognition accuracy for each category further improved. Notably, categories such as MC, PC, and PT were classified almost perfectly, indicating that the model possesses enhanced recognition capability for high-resolution fine-grained features.

It can be observed from all confusion matrices that no large-scale systematic confusion occurs among categories, with only a few categories exhibiting relatively higher misclassification rates. Consequently, the proposed model achieves not only high overall classification accuracy but also well-balanced performance across categories. The confusion matrix analysis thus further confirms that RFPN-Net is stable and reliable at different magnification levels. The model can correctly identify the majority of classes while effectively distinguishing tissues with similar morphologies, demonstrating strong fine-grained classification capability and practical value.

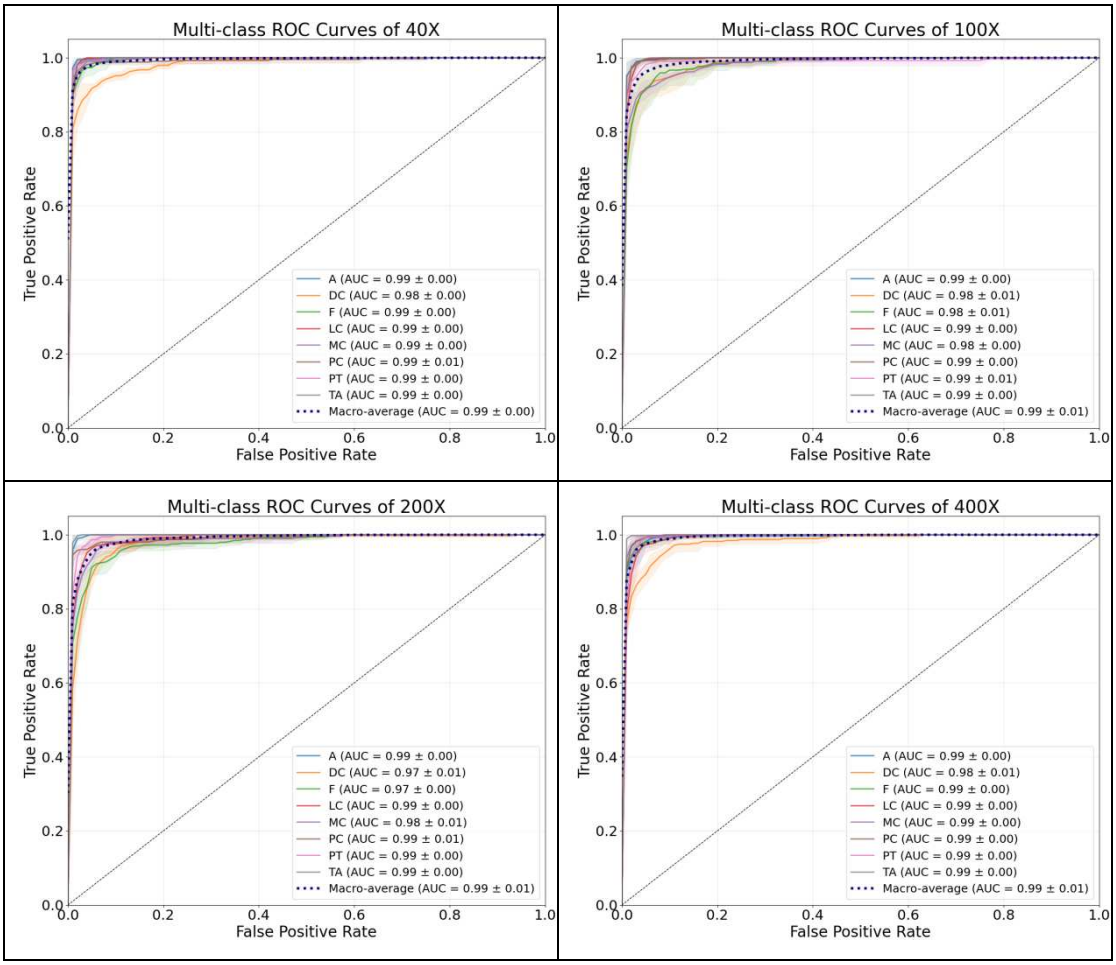


Fig. 9 ROC curves of RFPN-Net for breast cancer subtype classification on the BreakHis dataset at different magnification levels

Fig. 9 shows the ROC curves at different magnification levels (40×, 100×, 200×, and 400×), which can be used to analyze the model's discriminative ability for each class and to observe the overall stability of the model. The experimental results demonstrate that the proposed model achieves favorable classification performance across all magnifications. Specifically, the AUC values for all classes (A, DC, F, LC, MC, PC, PT, TA) exceed 0.98, with the majority surpassing 0.99, and the average AUC remains stable at around 0.99, indicating a strong capability to distinguish among various classes. A comparison across magnifications reveals no significant performance degradation from 40× to 400×, suggesting that the model adapts well to variations in image details. Whether extracting global structural information at low magnification or fine-grained cellular features at high magnification, the model can effectively capture discriminative features, thereby yielding stable and consistent classification results. The ROC curves are generally concentrated in the upper-left corner, meaning that the model can achieve a near-perfect true positive rate, demonstrating an excellent balance between sensitivity and specificity.

Table 4 reports the classification performance results of the proposed method on the BACH dataset obtained through five-fold cross-validation. As can be seen from the results, the proposed RFPN-Net model continues to maintain excellent classification performance on the BACH dataset, with all evaluation metrics achieving consistently high levels.

Table 4 Performance metrics of RFPN-Net on the BACH test dataset

BACH						
Fold	Accuracy	Precision	Recall	F1-Score	Specificity	AUC
Fold1	98.75	98.81	98.75	98.78	99.58	99.17
Fold2	99.38	99.39	99.38	99.38	99.79	99.58
Fold3	99.38	99.39	99.38	99.38	99.79	99.58
Fold4	99.38	99.39	99.38	99.38	99.79	99.58
Fold5	1.00	1.00	1.00	1.00	1.00	1.00
Mean±STD	99.38±0.44	99.40±0.42	99.38±0.44	99.39±0.43	99.79±0.15	99.58±0.29

As shown in the table, the proposed method consistently achieves excellent performance across all folds, with an average accuracy of 99.38% ± 0.44%, precision of 99.40% ± 0.42%, recall of 99.38% ± 0.44%, and F1-score of 99.39% ± 0.43%. Furthermore, the model also demonstrates outstanding class discrimination capability, attaining a specificity of 99.79% ± 0.15% and an AUC of 99.58% ± 0.29%. The low standard deviations across all evaluation metrics indicate that the proposed model possesses strong robustness and stability on the BACH dataset.

Supplementary Fig. S3 presents the average confusion matrix and the multi-class ROC curves of the proposed method on the BACH dataset. It can be observed from the confusion matrix that the model achieves high classification accuracy for the four categories—Benign, InSitu, Invasive, and Normal—with only a small number of misclassifications occurring between classes.

The ROC curves also demonstrate that the curves for all categories are positioned close to the upper-left corner, achieving high true positive rates at low false positive rates. The AUC values for all classes are close to 1, with a macro-average AUC of approximately 0.99, indicating that the model possesses excellent and stable discrimination capability for the subtype classification task. The confusion matrix analysis further reveals that the proposed model performs favorably across all performance metrics and maintains good balance and reliability across categories. These results confirm that the experimental outcomes of RFPN-Net on the BACH dataset once again validate its strong generalization ability and stability under various data distributions, suggesting promising application prospects in breast cancer histopathological image subtype classification.

3.2 Comparative experiments

3.2.1 Comparison with classical algorithms

To validate the classification performance of RFPN-Net, the classification accuracy was evaluated on the test data at each of the four magnification levels (40×, 100×, 200×, and 400×) of the BreakHis dataset. Several mainstream deep learning models, including MobileNetV2, ShuffleNet, EfficientNet, Swin Transformer, and DCDNet, were selected as comparison algorithms. To ensure the reliability of the experimental results, each model was fully trained for 100 epochs. A comprehensive evaluation was conducted on all models at the four different magnification levels of the BreakHis dataset, and the results are presented in Table 5. By comparing the classification accuracy of these models at different magnifications, the performance advantages of the proposed model can be more intuitively assessed.

Table 5 Performance metrics of different CNN methods on the BreakHis dataset

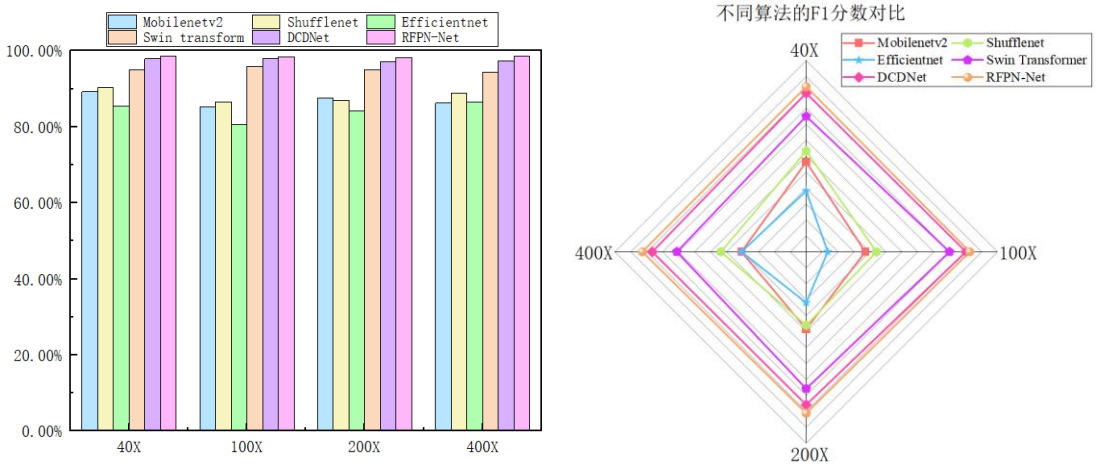
Model	Metric	Results on different scales of the BreakHis dataset			
		40×	100×	200×	400×
MobileNetV2 (2016)	Accuracy	89.18%	85.25%	87.50%	86.24%
	Precision	89.36%	85.29%	87.88%	86.23%
	Recall	89.19%	85.53%	87.48%	86.00%
	F1-Score	89.27%	85.41%	87.68%	86.12%
ShuffleNet (2018)	Accuracy	90.40%	86.59%	86.88%	88.87%
	Precision	91.02%	86.71%	87.37%	88.75%
	Recall	90.28%	86.94%	87.07%	88.64%
	F1-Score	90.65%	86.82%	87.22%	88.70%
EfficientNet (2019)	Accuracy	85.52%	80.63%	84.10%	86.40%
	Precision	85.67%	80.53%	84.48%	86.16%
	Recall	85.60%	80.86%	84.28%	86.06%
	F1-Score	85.64%	80.70%	84.38%	86.11%
Swin Transformer (2021)	Accuracy	94.97%	95.83%	95.06%	94.28%
	Precision	95.02%	95.99%	95.14%	94.35%
	Recall	94.97%	95.98%	95.22%	94.11%
	F1-Score	95.00%	95.98%	95.18%	94.23%
DCDNet (2021)	Accuracy	97.87%	97.91%	97.07%	97.37%
	Precision	97.86%	98.02%	97.17%	97.31%
	Recall	97.94%	98.04%	97.21%	97.32%
	F1-Score	97.90%	98.03%	97.19%	97.31%
Our	Accuracy	98.68%	98.44%	98.21%	98.58%
	Precision	98.72%	98.52%	98.26%	98.64%
	Recall	98.72%	98.53%	98.30%	98.53%
	F1-Score	98.72%	98.52%	98.26%	98.55%

It can be observed that lightweight networks such as MobileNetV2, ShuffleNet, and EfficientNet generally achieve accuracies in the range of 80%–90% across different magnifications, with ShuffleNet slightly outperforming the other two lightweight models. However, these networks still exhibit noticeable deficiencies in distinguishing among multiple breast tissue classes. In contrast, Swin Transformer and DCDNet, as more recent architectures, achieve significantly improved overall performance, with accuracies generally exceeding 94% for Swin Transformer and 97% for DCDNet, and demonstrate greater robustness in recognizing more complex tissue structures. Notably, our proposed method achieves the highest performance at all magnification levels, attaining accuracies of 98.68%, 98.44%, 98.21%, and 98.58% for the 40×, 100×, 200×, and 400× magnifications, respectively; Precision, Recall, and F1-Score also remain consistently above 98%, significantly outperforming all compared models. These results indicate that the proposed model possesses superior generalization ability and robustness across pathological images with varying resolutions and tissue morphologies, and can more accurately differentiate among fine-grained breast cancer subtypes.

To gain a comprehensive understanding of the classification performance of the proposed method on multi-scale pathological images, this paper graphically presents the accuracy and F1-score at four magnifications—40×, 100×, 200×, and 400×—as shown in Fig. 10.

As can be seen from the accuracy bar chart on the left, RFPN-Net achieves the highest accuracy at all magnifications, remaining above 98% and substantially outperforming the other compared methods. Swin Transformer and DCDNet rank second, with their accuracy curves closely following and exhibiting minimal variation. In contrast, MobileNetV2, ShuffleNet,

and EfficientNet yield lower accuracy and show slight fluctuations across different magnifications, indicating their relatively weak capability in modeling multi-scale pathological features. The F1-score radar chart on the right further corroborates these findings: at magnifications of 40×, 100×, 200×, and 400×, RFPN-Net's F1-score lies on the outermost ring of the radar chart, approaching 0.98, demonstrating an excellent balance between precision and recall. The F1-scores of DCDNet and Swin Transformer are slightly lower than that of the proposed method, while EfficientNet, ShuffleNet, and MobileNetV2 exhibit overall lower F1-scores, with the discrepancy being more pronounced particularly at intermediate magnifications (100× and 200×).



(a)Accuracy Comparison (b)F1-Score Comparison

Fig. 10 Performance comparison of different CNN methods on the BreakHis dataset

Overall, the proposed method achieves consistently high accuracy and F1-score on histopathological images at different magnifications, and it significantly outperforms classical CNN models as well as the Swin Transformer method built upon the Transformer architecture, demonstrating strong performance and favorable robustness in the histopathological image subtype classification task.

Supplementary Fig. S4 presents the average ROC curves and the corresponding AUC values of each model on the BreakHis dataset at the four magnification levels (40×, 100×, 200×, and 400×). Overall, all models demonstrate reasonably good discriminative ability across different magnifications, but their performance varies considerably. Lightweight models such as EfficientNetB0, MobileNetV2, and ShuffleNetV2 yield relatively lower overall AUC values, with a noticeable decline in discrimination at higher magnifications (200× and 400×), where AUCs mostly fall within the range of 0.93–0.96. In contrast, Swin Transformer and DCDNet exhibit higher curve trends and larger AUCs exceeding 0.98 across all four magnifications, suggesting stronger feature representation capability for complex tissue structures. Notably, the proposed RFPN-Net achieves the highest AUCs across all magnifications: 0.996, 0.994, 0.990, and 0.996 at 40×, 100×, 200×, and 400×, respectively, with ROC curves approaching the upper-left corner, indicating excellent sensitivity and specificity at different resolutions. These results demonstrate that the proposed method exhibits strong robustness in multi-magnification breast cancer subtype classification, providing a reliable and efficient solution for this task.

3.2.2 Comparison with state-of-the-art methods

To systematically validate the advancement and clinical applicability of the proposed RFPN-Net model and to ensure the scientific rigor and fairness of the comparison experiments, six representative state-of-the-art models from the current literature, encompassing different

technical paradigms, were selected as baseline methods. DBLCNN [13], ES-CNN [30], and Pretrained-Xception+FCNN [31] are classical improved models based on convolutional neural networks. EfficientNetV2+DSE [20] is a lightweight and efficient network integrated with attention mechanisms. TokenMixer [32] is a vision architecture based on Transformer, and MFF-HistoNet [33] is a dedicated network designed for multi-scale feature fusion in pathological images. Table 6 reports the classification accuracy of all compared models at the four magnification levels.

Table 6 Performance metrics of different state-of-the-art models on the BreakHis dataset

Model	40×	100×	200×	400×
DBLCNN(2022)	96.58%	95.65%	96.43%	95.27%
EfficientNetV2+DSE(2022)	95.57%	94.34%	87.25%	93.99%
P-X+FCNN(2023)	94.69%	93.75%	91.71%	91.03%
TokenMixer(2024)	90.93%	91.67%	90.73%	93.29%
MFF-HistoNet(2025)	98.10%	98.80%	97.90%	97.30%
ES-CNN(2025)	97.41%	96.72%	95.97%	95.97%
RFPN-Net	98.68%	98.44%	98.21%	98.58%

As shown in the table, RFPN-Net exhibits superior overall performance across all four magnification factors. It achieves the highest accuracy at 40×, 200×, and 400×, being only slightly inferior to MFF-HistoNet at 100×. Specifically, its accuracies at 40×, 200×, and 400× reach 98.68%, 98.21%, and 98.58%, respectively, representing state-of-the-art performance. At 100× magnification, its accuracy of 98.44% is marginally lower than that of MFF-HistoNet (98.80%), yet it remains significantly higher than those of all other baseline models. Notably, the accuracies across all four magnifications are consistently maintained above 98%, demonstrating excellent cross-scale robustness.Compared with contemporary mainstream methods, RFPN-Net significantly outperforms earlier approaches such as DBLCNN and EfficientNetV2+DSE, as well as recently proposed advanced models like TokenMixer and ES-CNN. Furthermore, it demonstrates a clear advantage over the highly competitive MFF-HistoNet at 200× and 400× magnifications, with only a marginal gap at 100×. This thoroughly validates the superiority and robustness of the proposed method in multi-magnification pathological image subtype classification tasks. The consistent performance of RFPN-Net across varying magnifications underscores the effectiveness of its multi-scale feature extraction and fusion mechanisms. Ultimately, these results comprehensively demonstrate that the proposed RFPN-Net yields highly accurate and stable classification outcomes across diverse magnification levels.

Furthermore, a comparative analysis of other baseline models across varying magnifications reveals considerable instability. For instance, the accuracy of EfficientNetV2+DSE at 200× is merely 87.25%, which is substantially lower than the 91.03% achieved by P-X+FCNN at 400×. This discrepancy indicates that these models lack robustness when applied to multi-scale pathological image classification. In contrast, RFPN-Net maintains consistent performance across different magnifications, underscoring its efficacy in the extraction and enhancement of multi-scale pathological features. This further demonstrates that the novel feature extraction and enhancement mechanisms proposed in this study yield more accurate and reliable classification results across diverse magnification levels. In conclusion, these experimental results validate the advanced nature and practical applicability of RFPN-Net in multi-magnification pathological image classification tasks, indicating its potential to provide robust support for the intelligent, computer-aided diagnosis of breast cancer.

As illustrated in Supplementary Fig. S5, the tabular data is transformed into a line chart to clearly visualize the performance trends of the various models across different magnification factors. The trajectories of the curves indicate that the proposed model maintains exceptional

accuracy across all magnifications and exhibits the highest stability. Although a slight decline is observed at 100×, the accuracy consistently remains above 0.98, demonstrating strong adaptability to pathological images of varying resolutions. The overall performance of MFF-HistoNet is highly comparable to that of the proposed method. While ES-CNN and DBLCNN maintain relatively high accuracy at certain magnifications, their overall performance remains inferior to that of RFPN-Net. In contrast, P-X+FCNN, TokenMixer, and EfficientNetV2+DSE are significantly affected by the magnification factor, exhibiting substantial fluctuations in accuracy. Notably, TokenMixer experiences a performance trough at 200×, and the accuracy of EfficientNetV2+DSE drops to approximately 0.87 at the same magnification. This indicates that their feature extraction capabilities for medium-resolution images are limited, resulting in comparatively weaker robustness.

3.3 t-SNE visualization experiment

Supplementary Fig. S6 illustrates the performance of the BreakHis dataset at different magnification factors (40×, 100×, 200×, and 400×), with the t-SNE visualization results clearly demonstrating the feature representation capabilities of the various models. Compared with baseline models such as MobileNetV2, ShuffleNet, EfficientNet, Swin Transformer, and DCDNet, the proposed model consistently exhibits more pronounced inter-class separation and intra-class clustering across all magnification levels. As a widely utilized non-linear dimensionality reduction technique, t-SNE effectively projects high-dimensional features into a two- or three-dimensional space while preserving the local neighborhood structure of the samples. Consequently, samples of the same category with high feature similarity form tight clusters, while samples from different categories exhibit clearly separated distributions if they are highly discriminative. The feature points of each category in our proposed model are relatively concentrated with distinct boundaries between classes. This robust feature separation is maintained across all scales, providing a more reliable foundation for subsequent classification tasks.

The t-SNE visualization results shown in the figure reveal significant differences in feature distributions among different models on the BreakHis dataset at four magnifications (40×, 100×, 200×, and 400×). Overall, traditional lightweight networks such as MobileNetV2, ShuffleNet, and EfficientNet exhibit relatively scattered feature distributions, with pronounced overlaps between classes and unclear inter-class boundaries, indicating their limited discriminative capability. Swin Transformer achieves improved feature clustering compared with CNN models, yet partial class overlap persists. DCDNet further enlarges the boundary distances between certain classes through a dual-channel feature interaction mechanism; however, its intra-class feature aggregation remains insufficiently compact. In contrast, RFPN-Net produces more compact and well-separated feature distributions, forming clear cluster structures for each class in the feature space, with markedly enlarged inter-class margins and stronger intra-class consistency. This demonstrates that the proposed method can effectively extract more discriminative deep features and exhibits better robustness across different magnifications. Notably, even at high magnifications (200× and 400×), RFPN-Net maintains a favorable clustering structure, further validating its superior performance in breast cancer histopathological image subtype classification.

3.4 Ablation experiments

To quantitatively analyze the specific contributions of the proposed MFEM and CSFM modules to model performance, a series of ablation experiments were conducted on the BreakHis dataset at four magnification levels (40×, 100×, 200×, and 400×), using ResNet50 as the baseline. Five-fold cross-validation was employed to enhance the robustness and reliability of the results. In the experimental design, the target modules were progressively introduced to construct four comparison models: the baseline model using only ResNet50 (ResNet50 (Base)), ResNet50 with only the CSFM module (ResNet50+CSFM), ResNet50 with only the

MFEM module (ResNet50+MFEM), and the final model RFPN-Net integrating both MFEM and CSFM. By comparing the classification accuracy of different models at each magnification level, the individual contributions and synergistic gains of each module can be clearly decoupled, thereby validating the effectiveness of the designed modules in improving the subtype classification performance of breast cancer histopathological images.

Table 7 Ablation study results of RFPN-Net in terms of accuracy

Model	Fold1	Fold2	Fold3	Fold4	Fold5	Average
40×						
ResNet50(BASE)	91.77%	92.40%	94.52%	95.58%	94.20%	93.69%
ResNet50+CSFM	95.42%	96.96%	97.87%	97.71%	96.64%	96.95%
ResNet50+MFEM	98.17%	97.87%	98.33%	97.56%	98.32%	98.05%
RFPN-Net	98.78%	98.94%	98.48%	98.63%	98.63%	98.69%
100×						
ResNet50(BASE)	89.87%	87.87%	89.34%	91.29%	90.69%	89.81%
ResNet50+CSFM	93.88%	96.86%	96.84%	96.09%	95.34%	95.80%
ResNet50+MFEM	97.02%	95.36%	97.75%	96.40%	96.85%	96.67%
RFPN-Net	99.11%	98.50%	97.75%	98.50%	98.35%	98.44%
200×						
ResNet50(BASE)	92.90%	88.92%	87.98%	91.22%	89.81%	90.17%
ResNet50+CSFM	96.14%	95.08%	97.53%	97.84%	96.14%	96.55%
ResNet50+MFEM	96.60%	96.76%	97.07%	96.92%	96.45%	96.76%
RFPN-Net	98.46%	97.69%	98.92%	98.31%	97.69%	98.21%
400×						
ResNet50(BASE)	90.57%	91.37%	88.85%	90.56%	92.11%	90.72%
ResNet50+CSFM	97.22%	94.76%	95.82%	94.42%	95.20%	95.48%
ResNet50+MFEM	97.37%	95.69%	95.20%	95.82%	94.40%	95.70%
RFPN-Net	98.92%	98.31%	98.30%	98.61%	98.76%	98.58%

As shown in Table 7, when the CSFM module is introduced alone, the classification accuracy improves consistently across all magnifications, with average accuracies of 96.95%, 95.80%, 96.55%, and 95.48% at 40×, 100×, 200×, and 400×, respectively. This indicates that the CSFM module effectively enhances the semantic information of feature representations and strengthens the model's ability to recognize critical pathological features. Employing the MFEM module alone also improves model performance, achieving accuracies of 98.05%, 96.67%, 96.76%, and 95.70% at the four magnifications, respectively, demonstrating that multi-level feature aggregation can fuse semantic and local information from different levels. The advantage of the MFEM module lies in its ability to fully exploit multi-scale features from low-, medium-, and high-magnification images, thereby maintaining high classification accuracy across most magnifications. Furthermore, when both the MFEM and CSFM modules are integrated into the baseline network, the model achieves the best performance at all four magnifications, with average classification accuracies of 98.69%, 98.44%, 98.21%, and 98.58%, respectively, significantly outperforming models with only a single module. These results demonstrate that the MFEM and CSFM modules exhibit a complementary relationship, and their synergistic effect not only enhances the discriminative power of the model but also strengthens its stability and robustness across different magnifications.

In summary, the ablation experiments convincingly demonstrate the effectiveness and necessity of the proposed MFEM and CSFM modules for breast cancer histopathological image subtype classification. MFEM enhances the capability of multi-level feature extraction, while CSFM strengthens the effective fusion of critical features. The combination of both modules significantly improves the subtype classification accuracy and cross-scale adaptability of the model, thereby providing a reliable technical foundation for breast cancer histopathological image analysis.

4 Discussion

This paper proposes RFPN-Net, a breast cancer histopathological image classification model based on multi-layer attention and cross-scale fusion, designed to address the multi-class

classification of breast cancer pathology images. The model adopts a dual-branch architecture consisting of an improved ResNet50 and a Swin Transformer, simultaneously capturing local fine-grained features at the cellular level and global semantic information of tissue macrostructures. This enables the assessment of multi-scale pathological features, enhances the ability to recognize subtle pathological differences, and ultimately achieves a hierarchical and multi-scale representation of breast tissue images. Such a design allows the model to capture subtle pathological patterns through the multi-layer attention mechanism (CBAM) while preserving the overall tissue architecture via cross-scale fusion, which is crucial for the multi-subtype classification of breast cancer where the information distribution exhibits pronounced multi-scale characteristics. Ablation studies (MFEM+CSFM) and visualization results (t-SNE) demonstrate that the combination of the proposed Multi-level Feature Extraction Module (MFEM) and Cross-Scale Feature Fusion Module (CSFM) effectively improves classification accuracy and cross-scale adaptability. Specifically, the t-SNE plots show that the designed modules extract highly discriminative pathological features and maintain favorable feature separability across all magnifications. By effectively integrating local details with global tissue semantic information, the model fully exploits the rich multi-scale and multi-type information in histopathological images, significantly enhancing the accuracy, stability, and diagnostic reliability of subtype classification.

To highlight the potential clinical utility of RFPN-Net, the model could facilitate the diagnostic workflow in the following ways: assisting pathologists in accurately identifying various breast cancer subtypes, reducing misdiagnosis, and providing critical references for targeted therapy selection and prognosis evaluation; and, through its synergistic CNN-Transformer architecture, simultaneously capturing local cellular features and global tissue information, thereby supporting multi-scale lesion localization and enabling comprehensive pathological subtyping of breast cancer. In terms of interpretability and validation, ablation studies and feature visualization confirm that the model effectively focuses on critical pathological regions, providing spatial insights that enhance clinical trust. It should be noted that although attention mechanisms and visualization methods offer a certain degree of interpretability, other feature attribution methods such as SHAP were not employed in this study. This is mainly because, under the large-scale global modeling of the Transformer architecture, SHAP incurs extremely high computational cost and implementation complexity, making large-scale application infeasible under current experimental conditions. Nevertheless, SHAP and similar methods have been identified as future research directions to further improve the transparency and clinical utility of multi-subtype model predictions.

Although the model incorporates Focal Loss during training to address class imbalance and has demonstrated strong generalization across multiple datasets, certain limitations remain. In particular, when applied to breast cancer multi-class datasets of limited size, model performance may still be affected by the relatively small number of minority class samples, variations in staining protocols across different laboratories, and differences in digital scanner imaging schemes. Furthermore, the model currently relies primarily on visual features from images alone and does not integrate clinically relevant metadata closely associated with subtypes (such as family history, gene mutation types, or biomarkers). Incorporating such multi-dimensional information may further push the accuracy ceiling beyond what can be achieved by image-only judgment.

To address the aforementioned limitations in data diversity, modality singularity, and computational complexity, future work will focus on the following directions: multimodal fusion, i.e., integrating complementary information from different imaging modalities (e.g., immunohistochemistry IHC images) and other relevant molecular data to enhance diagnostic robustness; cross-dataset transfer learning, i.e., leveraging heterogeneous datasets from multiple

medical centers to improve model generalization across different populations, imaging devices, and staining conditions; integration of clinical metadata, i.e., combining patient demographic information, medical history, and molecular biomarker data to enable more comprehensive and targeted decision support for breast cancer diagnosis; and enhancing interpretability, i.e., further improving interpretability mechanisms, including the planned exploration of SHAP or similar advanced attribution methods, to provide transparent and clinically actionable insights from the model's subtype classification predictions.

5 Conclusion

Based on the PyTorch deep learning framework, this paper proposes a breast cancer histopathological image subtype classification network, RFPN-Net, which integrates a multi-layer attention mechanism with cross-scale feature extraction. The model is trained using an end-to-end optimization strategy, aiming to overcome the limitations of single-scale feature extraction in handling highly heterogeneous pathological images through a hybrid CNN-Transformer architecture, thereby achieving high-precision assessment of breast cancer subtypes. This paper elaborates on the network architecture of RFPN-Net and conducts experiments on both the BreakHis and BACH datasets, evaluating the performance on the BreakHis dataset at different magnification levels as well as its generalization capability on the BACH dataset. The core innovation of RFPN-Net lies in its dual-branch parallel collaborative design, in which the Multi-level Feature Extraction Module (MFEM) employs an improved ResNet50 as the backbone and embeds the Convolutional Block Attention Module (CBAM) into the residual mapping, achieving adaptively enhanced modeling of microscopic fine-grained features such as nuclear morphology and nucleocytoplasmic ratio. Experimental results show that RFPN-Net achieves classification accuracies of 98.68%, 98.44%, 98.21%, and 98.58% at the four magnification levels (40×, 100×, 200×, and 400×) of the BreakHis dataset, demonstrating its stability and effectiveness across different magnifications. Simultaneously, it attains a classification accuracy of 99.3% on the BACH dataset, exhibiting remarkably high accuracy. Compared with state-of-the-art deep learning models and multiple standard models, RFPN-Net demonstrates significant advantages. Ablation experiments and t-SNE-based high-dimensional feature visualization analysis further indicate that RFPN-Net not only validates the contribution of each module to feature separability but also accurately focuses on highly discriminative pathological topological regions, providing intuitive and interpretable insights for clinicians. Although RFPN-Net has achieved a significant breakthrough in classification performance, it still faces technical bottlenecks, including a relatively large number of parameters and high computational overhead that hinder deployment in resource-constrained environments, as well as the under-exploited auxiliary value of clinical metadata. Therefore, future research will focus on model lightweighting, constructing multimodal fusion, and actively exploring domain adaptation learning to mitigate the impact of staining variations arising from different preparation protocols and scanning devices, thereby further enhancing the cross-domain generalization capability and clinical utility of the proposed method.

References

1. World Health Organization (2020) WHO report on cancer: setting priorities, investing wisely and providing care for all. World Health Organization, Geneva
2. Zhang XK, Ma ZQ, Zhao WH, Cui DM (2022) A review of breast cancer histopathological image classification research based on convolutional neural networks. Computer Science 49(11A):210800232-210800239 (in Chinese)
3. Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60-88

4. Gurcan MN, Boucheron LE, Can A et al (2009) Histopathological image analysis: a review. IEEE Rev Biomed Eng 2:147-171
5. Macenko M, Niethammer M, Marron JS et al (2009) A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, Boston, pp 1107-1110
6. Spanhol FA, Oliveira LS, Petitjean C et al (2016) A dataset for breast cancer histopathological image classification. IEEE Trans Biomed Eng 63(7):1455-1462
7. Benhammou Y, Achchab B, Hermosilla G et al (2020) BreakHis based breast cancer automatic diagnosis using deep learning: taxonomy, survey and insights. Neurocomputing 375:9-24
8. Boumaraf S, Liu X, Zheng Z et al (2021) A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images. Biomed Signal Process Control 63:102192
9. Ciga O, Xu T, Nofech-Mozes S et al (2021) Overcoming the limitations of patch-based learning to detect cancer in whole slide images. Sci Rep 11:8894
10. Sharma S, Mehra R (2020) Conventional machine learning and deep learning approach for multi-classification of breast cancer histopathology images - a comparative insight. J Digit Imaging 33(3):632-654
11. Bardou D, Zhang K, Ahmad SM (2018) Classification of breast cancer based on histology images using convolutional neural networks. IEEE Access 6:24680-24693
12. Boumaraf S, Liu XB, Wan YC et al (2021) Conventional machine learning versus deep learning for magnification dependent histopathological breast cancer image classification: a comparative study with visual explanation. Diagnostics 11(3):528
13. Wang CQ, Gong WJ, Cheng JL et al (2022) DBLCNN: dependency-based lightweight convolutional neural network for multi-classification of breast histopathology images. Biomed Signal Process Control 73:103451
14. Deniz E, Sengur A, Kadiroglu Z et al (2018) Transfer learning based histopathologic image classification for breast cancer detection. Health Inf Sci Syst 6:18
15. Wang P, Wang JX, Li YM et al (2021) Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing. Biomed Signal Process Control 63:102341
16. Nahid AA, Kong Y (2021) Optimised CNN in conjunction with efficient pooling strategy for the multi-classification of breast cancer. IET Image Process 15(3):728-739
17. Jiao L, Huang Q, Zhang H et al (2025) Multiscale deep learning for detection and recognition: a comprehensive survey. IEEE Trans Neural Netw Learn Syst 36(4):5900-5920
18. Eliza RE, Zulkifley MA, Muharar R et al (2022) A review on multiscale-deep-learning applications. Sensors 22(19):7384
19. Yang ZB, Ran LY, Zhang SZ et al (2019) EMS-Net: ensemble of multiscale convolutional neural networks for classification of breast cancer histology images. Neurocomputing 366:46-53
20. Sarker MMK, Akram F, Alsharid M et al (2022) Efficient breast cancer classification network with dual squeeze and excitation in histopathological images. Diagnostics 13(1):103
21. Xie JY, Liu RR, Luttrell J et al (2019) Deep learning based analysis of histopathological images of breast cancer. Front Genet 10:80
22. Deng J, Dong W, Socher R et al (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami, pp 248-255
23. Das A, Nair MS, Peter SD (2020) Computer-aided histopathological image analysis techniques for automated nuclear atypia scoring of breast cancer: a review. J Digit Imaging 33(5):1091-1121
24. Gandomkar Z, Brennan PC, Mello-Thoms C (2018) MuDeRN: multi-category classification of breast histopathological image using deep residual networks. Artif Intell Med 88:14-24
25. Hao Y, Qiao SC, Zhang LY et al (2022) Breast cancer histopathological images classification based on deep semantic features and gray level co-occurrence matrix. PLoS ONE 17(5):e0268211
26. Sanchez Morillo D, Gonzalez J, Garcia-Rojo M et al (2018) Classification of breast cancer histopathological images using KAZE features. In: Proceedings of the 6th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2018). Springer, Granada, pp 276-286
27. Kahya MA (2017) Classification of breast cancer histopathology images based on adaptive sparse support vector machine. J Appl Math Bioinform 7(1):47-63
28. Guo Y, Dong HH, Song FZ et al (2018) Breast cancer histology image classification based on deep neural networks. In: Lecture Notes in Computer Science, vol 10882. Springer, Cham, pp 827-836

29. Wei B, Han ZQ, He XY et al (2017) Deep learning model based breast cancer histopathological image classification. In: 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE, Chengdu, pp 348-353
30. Yusuf M, Kana AFD, Bagiwa MA et al (2025) Multi-classification of breast cancer histopathological image using enhanced shallow convolutional neural network. J Eng Appl Sci 72:24
31. Pandey A, Kumar A (2023) An integrated approach for breast cancer classification. Multimed Tools Appl 82:33357-33377
32. Abimouloud ML, Bensid K, Elleuch M et al (2025) Advancing breast cancer diagnosis: token vision transformers for faster and accurate classification of histopathology images. Vis Comput Ind Biomed Art 8:1
33. Mahmood T, Saba T, Rehman A (2025) Breast cancer diagnosis with MFF-HistoNet: a multi-modal feature fusion network integrating CNNs and quantum tensor networks. J Big Data 12:60-82
34. Aresta G, Araújo T, Kwok S et al (2019) BACH: Grand challenge on breast cancer histology images. Med Image Anal 56:122-139. https://doi.org/10.1016/j.media.2019.05.010

Statements and Declarations

Funding

This study was supported by the Horizontal Project of the Jilin Provincial Department of Science and Technology (Project No. 2021JBH02L38) and the Science and Technology Development Plan Project of Jilin Province (Project No. YDZJ202301ZYTS180).

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Not applicable. This study used publicly available, de-identified histopathological image datasets and did not involve human participants or animals.

Consent to participate

Not applicable. The study used only publicly available datasets.

Consent to publish

Not applicable. The manuscript contains no identifiable participant data.

Data availability

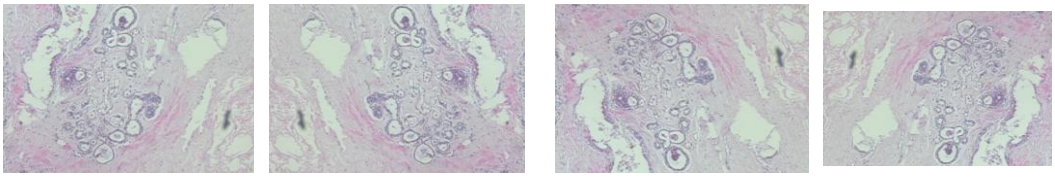
The BreakHis and BACH datasets analyzed in this study are publicly available. BreakHis can be accessed at https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/. BACH can be accessed through the ICIAR 2018 Grand Challenge at https://iciar2018-challenge.grand-challenge.org/Dataset/.

Author contributions

All authors contributed to the study conception and design. Shuai Li, Guowei Wang, Luhan Li, Junpeng Cui, and Qingxing Su contributed to data preparation, implementation, and analysis. Xiaoli Wang supervised the project. Shuai Li and Xiaoli Wang drafted the manuscript. All authors reviewed and approved the final manuscript.

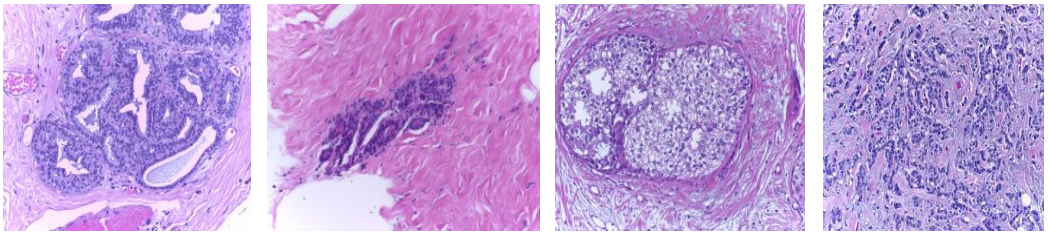
Supplementary Information

The following supplementary figures provide additional dataset examples and complementary performance visualizations.



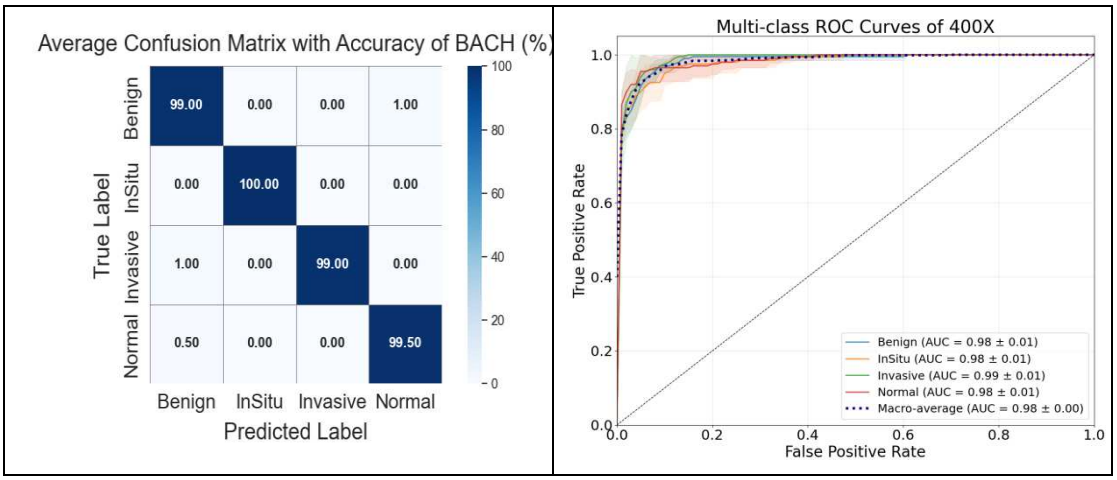
(a) **(b)** **(c)** **(d)**

Supplementary Fig. S1 Representative original images and their augmented counterparts; (a) original image, (b) horizontal flip, (c) vertical flip, and (d) 180° rotation

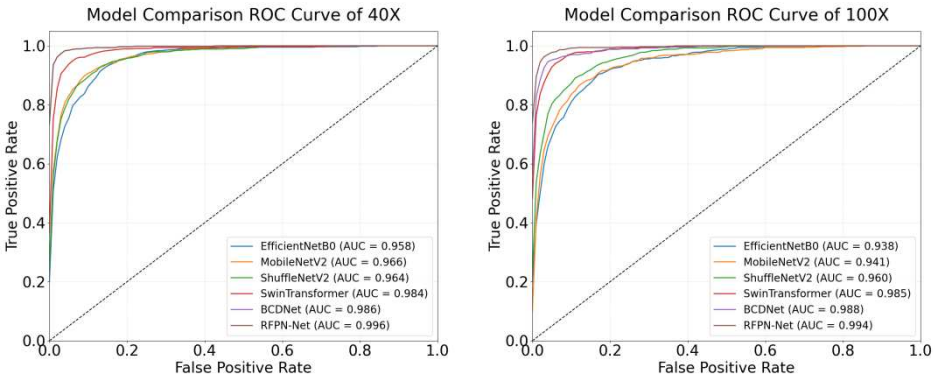


(a) **(b)** **(c)** **(d)**

Supplementary Fig. S2 Representative images from the BACH dataset; (a) normal, (b) benign, (c) carcinoma in situ, and (d) invasive carcinoma

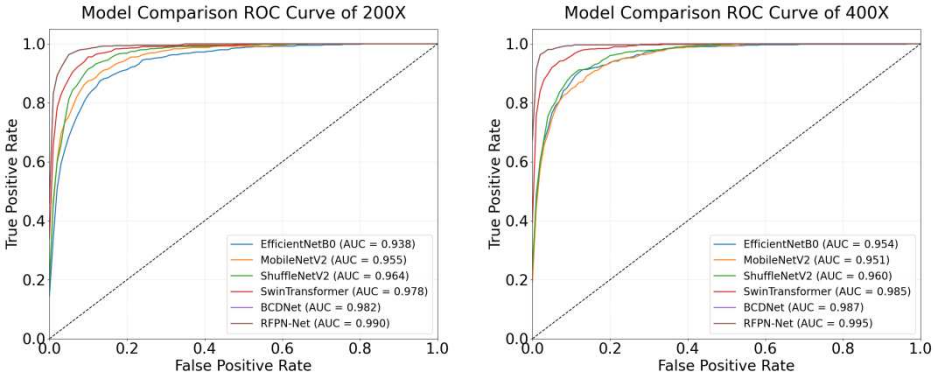


Supplementary Fig. S3 ROC curves of RFPN-Net for multi-class classification on the BACH dataset



Comparison of ROC curves for 40 different methods

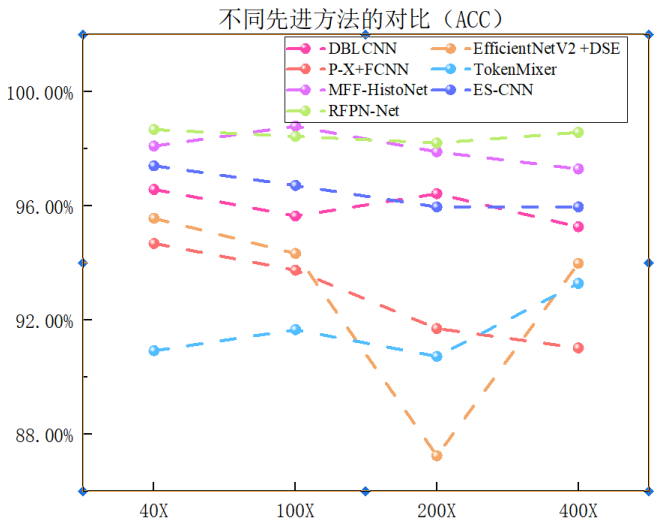
Comparison of ROC curves for 100 different methods



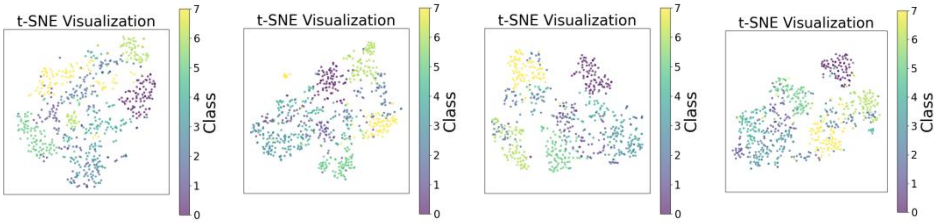
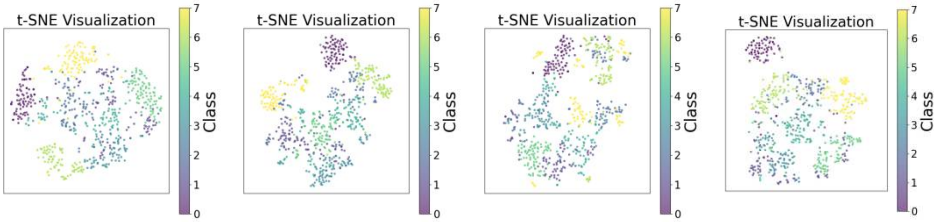
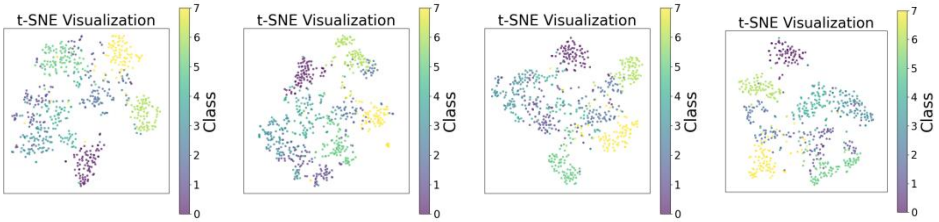
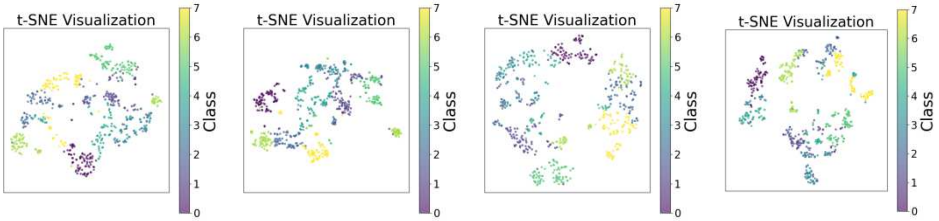
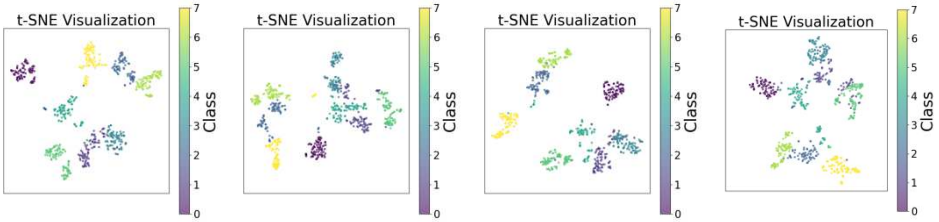
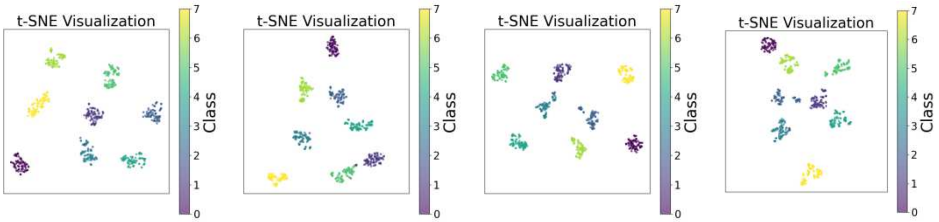
Comparison of ROC curves for 200 different methods

Comparison of ROC curves for 400 different methods

Supplementary Fig. S4 ROC curves of different methods on the BreakHis dataset



Supplementary Fig. S5 Comparison of classification accuracies among different state-of-the-art methods under various magnifications on the BreakHis dataset

	BreakHisDataset			
model	40×	100×	200×	400×
MobileNetV2(2016)				
Shufflenet (2018)				
Efficientnet(2019)				
Swin Transform (2021)				
DCDNet (2021)				
Our				

Supplementary Fig. S6 t-SNE visualization results of RFPN-Net on the BreakHis test set