Received: 6 January 2025 | Revised: 23 February 2025 | Accepted: 31 March 2025 | Published online: 14 April 2025

RESEARCH ARTICLE

Medinformatics
2025, Vol. 00(00) 1–10
DOI: 10.47852/bonviewMEDIN52025163

BON VIEW PUBLISHING

A Universal Intelligent Classification Algorithm for Pathological Images Based on Sliding Window Attention Mechanism

Huiqin Jiang[1,2,3,†], Zhiheng Tong[1,†], Xiaonan Yang[1,†], Fangjie Zhao[1], Jinhong Tan[1], Xing Dong[4], Zhenying Ban[4], Xianxu Zeng[2,4,*], Xin Zhao[2,4,*] and Ling Ma[3]

[1]*School of Electrical and Information Engineering, Zhengzhou University, China*

[2]*Tianjian Laboratory of Advanced Biomedical Sciences, Institute of Advanced Biomedical Sciences, Zhengzhou University, China*

[3]*Zhengzhou Zhililkang Co. Ltd., China*

[4]*The Third Affiliated Hospital of Zhengzhou University, China*

Abstract: The pathological diagnosis is the gold standard for qualitative analysis of tumors. Due to the difficulty in extracting complete features from high-resolution whole-slice pathological images, the generality and accuracy of traditional deep learning classification algorithms are limited. This paper proposes an intelligent classification algorithm by combining convolutional neural networks (CNN) and Transformer for pathological images. Firstly, the local and global features of pathological images are extracted using the designed CNN and Transformer hybrid network architecture. Furthermore, the Mish activation function is introduced to improve the nonlinear expression ability of the feature extraction network. Finally, by stacking multiple convolutional blocks and residual attention blocks to increase model depth, the classification accuracy is improved. The main contribution lies in the design of a residual module that introduces a sliding window multi-head attention mechanism, which enhances the algorithm's ability to extract contextual information. While effectively reducing computational complexity, it also improves classification accuracy. Experimental results show that the proposed algorithm attains accuracies of 0.987 and 0.947 for classifying benign and malignant lung and breast pathological images, respectively. The algorithm also achieves an accuracy of 0.932 in classifying benign, adenocarcinoma, and squamous cell carcinoma images and 0.841 in distinguishing benign and four subtypes of breast cancer. Moreover, it achieves an accuracy of 0.976 on a private dataset for breast cancer tissue pathological grading, which shows that the algorithm is universal and feasible in multidisease multiclassification tasks and clinical applications.

Keywords: pathological image classification, attention mechanism, histological grading, residual network, intelligent diagnosis

1. Introduction

Pathological diagnosis serves as the definitive method for tumor classification [1]. In tumor treatment, pathological diagnosis directly impacts the selection of subsequent treatment plans for patients and is crucial for predicting prognosis [2]. However, with the rapid growth in demand for pathological diagnosis in China, there are challenges such as a shortage of pathologists and uneven distribution of pathological resources [3]. There is an urgent need for AI-assisted clinical pathological diagnosis solutions.

AI-assisted pathological diagnosis has become a recent research focus, aiming to utilize image analysis techniques for classifying histopathological images, assisting in early diagnosis and treatment of tumors. The primary methods include traditional machine learning and deep learning-based classification techniques.

1.1. The traditional machine learning methods

Traditional machine learning methods involve manually selecting features from pathological images and using classifiers like support vector machines (SVM) for prediction. For example, George et al. [4] combined Otsu thresholding with fuzzy c-means clustering to extract image features. They utilized various neural network models, including multilayer perceptron to classify the malignancy of breast tissue cells. This approach effectively reduced false positives. Gupta and Bhavsar [5] utilized color and texture features and combined classifiers such as SVM, k-nearest neighbors, and decision trees with a majority voting strategy, achieving 88% accuracy in classifying breast pathology images at various magnifications. Peikari et al. [6] proposed a clustering-first labeling method to identify high-density regions in the semi-supervised learning space, enhancing SVM decision boundaries

***Corresponding authors:** Xianxu Zeng, Tianjian Laboratory of Advanced Biomedical Sciences, Institute of Advanced Biomedical Sciences, Zhengzhou University and The Third Affiliated Hospital of Zhengzhou University, China. Email: xianxuzeng@zzu.edu.cn and Xin Zhao, Tianjian Laboratory of Advanced Biomedical Sciences, Institute of Advanced Biomedical Sciences, Zhengzhou University and The Third Affiliated Hospital of Zhengzhou University, China. Email: zdsfyzx@zzu.edu.cn

[†]**Co-first author**

and achieving 92% accuracy in breast tissue classification. Trivizakis et al. [7] utilized artificial neural networks with local binary patterns, wavelet transforms, and Gabor filters to extract features from multi-scale pathological images and achieved 87.4% accuracy in classifying colorectal cancer into eight categories by using extracted features. Alqudah and Alqudah [8] extracted local binary pattern texture features within each window by introducing a sliding window feature extraction technique, constructed an SVM classifier, and achieved 91.12% accuracy in benign vs. malignant classification for breast pathology image.

Overall, the traditional machine learning classification methods, due to their reliance on extracting only shallow features from images and depending on human understanding and selection of these features, result in limited accuracy in pathological image classification.

1.2. Deep learning-based classification methods

Deep learning methods can improve classification accuracy by extracting hidden features from pathological images, enabling end-to-end disease diagnosis and prediction [9]. Spanhol et al. [10] effectively improved classification accuracy by employing the convolutional neural networks (CNNs) to classify breast tissue pathological images based on malignancy. Koné and Boulmane [11] achieved 81% accuracy in classifying normal breast tissue, benign tumors, ResNet50 is used to classify ductal carcinoma in situ and invasive carcinoma [12]. Teramoto et al. [13] attained 85.3% accuracy in lung malignancy classification by augmenting training datasets with generative adversarial networks. Phankokkruad [14] proposed an ensemble model combining VGG16, ResNet50V2, and DenseNet201, achieving 91% accuracy in a five-class classification task for pulmonary pathological images. Adu et al. [15] introduced a dual-level compressed capsule network for lung pathology classification, achieving 99.23% accuracy in distinguishing benign, squamous carcinoma, and adenocarcinoma. Srikantamurthy et al. [16] presented a hybrid model combining CNNs and long short-term memory networks for malignancy classification in histopathological images, achieving 99% accuracy. Liu et al. [9] proposed a classification algorithm for breast tumor pathological images by optimizing the cross-entropy loss function of AlexNet. Zou et al. [17] integrated attention mechanisms and higher-order statistical features into residual convolutional networks, achieving 99.29% and 85% classification accuracies on the BreakHis and BACH datasets, respectively.

Most CNN-based methods for classifying pathological images extract features through multiple convolutional and pooling layers. These networks can improve classification performance by capturing semantic information from low to high levels. However, since the lesion areas in pathological images are closely related to surrounding tissues, CNNs tend to focus only on local information and overlook contextual information, which needs further addressing.

Transformer models have gained popularity in recent years for their ability to capture global features from images using attention mechanisms [18]. Alotaibi et al. [19] designed a classification model that integrates ViT (Vision Transformer) [20] and DeiT (Data-efficient image Transformer) [21] to improve the performance of breast histopathology image classification. Thomas et al. [22] achieved an accuracy of 96% in classifying benign and malignant breast tissue pathology images using the ViT model. Tummala et al. [23] studied BreaST-Net and conducted experiments with the Swin Transformer [24] model for binary and eight-class classification tasks. The results showed that integrating four different Swin Transformer models significantly improves recognition accuracy compared to using a single model.

However, faced with information-rich pathological images, local features can provide detailed information, while global features can offer contextual information across the entire image. The challenge lies in how to effectively extract and organically integrate local and global features to improve classification performance. Therefore, this study combines CNNs with attention mechanisms to extract both local and global features from pathological images, enhancing classification accuracy.

Our contributions to this work are threefold:

1) We design a universal intelligent classification algorithm for pathological images by combining the respective strengths of CNN and Transformer.
2) We design a residual module that introduces a sliding window multi-head attention mechanism, which enhances the algorithm's ability to extract contextual information and reduce computational complexity.
3) Experimental results on several public datasets demonstrate that the proposed algorithm outperforms most existing method in both binary and multi-class classification tasks. In addition, the validation experiment on the private dataset of breast cancer histopathological grading also showed good performance, indicating the algorithm's generalizability, robustness, and feasibility for practical applications.

2. Related Work

2.1. Sliding window attention mechanism

In the pathology image classification, images can be viewed as two-dimensional sequences where each position corresponds to a patch. Self-attention mechanisms weigh different positions in the image, capturing comprehensive contextual information. Combining attention mechanisms with CNNs is crucial for improving the global feature extraction capability, as demonstrated by Chattopadyay et al. [25], who introduced channel attention mechanisms based on the bottleneck modules in ShuffleNet, which improves accuracy in breast tissue pathology image classification tasks.

Window Multi-head Self-Attention (W-MSA) is an attention mechanism used for processing local region in images. It independently captures relationships and importance within each window by dividing the input data into multiple windows.

Self-attention is computed independently within each individual window, restricting interaction between windows and potentially impacting comprehensive feature extraction. To overcome this limitation, the sliding window self-attention mechanism is introduced.

Sliding Window Multi-head Self-Attention (SW-MSA) introduces the concept of sliding windows using overlapping local windows on the input sequence. SW-MSA is particularly effective for handling long sequences, performing multi-head self-attention calculations within local windows to improve computational efficiency and resource utilization.

The window partition diagram of sliding window attention is illustrated in Figure 1, where four pre-segmented windows are depicted in different colors. Figure 1(A) shows the standard window self-attention segmentation after W-MSA at layer l, Figure 1(B) displays the sliding window self-attention segmentation after SW-MSA at layer l+1. The windows start from the top-left corner of Figure 1(A) and are offset by M/2 pixels along the right and bottom sides. Here, M is set to 4.

As shown in Figure 1, the sliding window self-attention mechanism leads to an increase in the number of windows. In

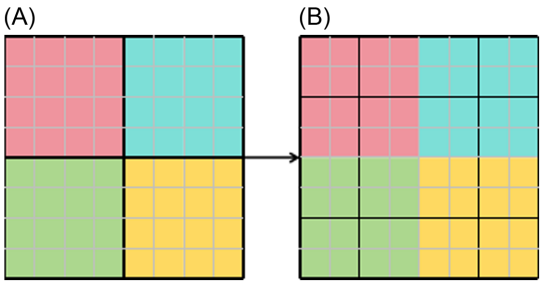

Figure 1. Schematic illustration of window partitioning for the sliding window attention mechanism

order to reduce computational complexity, an efficient calculation method of cyclic shifting is adopted.

Figure 2 illustrates the specific details of how cyclic shifting achieves interaction between windows, using a 2-pixel shift as an example. In Figure 1(B), the segmented nine windows are considered as the initial state in Figure 2(A). By cyclically shifting, four 2×2 windows, four 2×4 windows, and one 4×4 windows are rearranged into four 4×4 windows, as shown in Figure 2(C). Figure 2(B) depicts an intermediate step. Different colors are used to clearly illustrate the entire cyclic shifting process. Through this cyclic operation, the nine windows are eventually reduced back to four windows, ensuring consistent computational complexity.

The self-attention score calculation method is shown in Equation (1), which introduces relative position encoding to solve the permutation invariance problem in self-attention calculation.

$$\text{Attention}\,(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d}}+B\right)V \tag{1}$$

where Q, K and V represent the query matrix, the key matrix, and the value matrix, respectively. $Q,K,V\in R^{M^2\times d}$. The M^2 and d denote the number of image patches within a window and the dimension of either Q or K, respectively. B denotes the bias matrix, and $B\in R^{(2M-1)\times(2M-1)}$.

The sliding window attention mechanism divides the input pathological image into multiple local windows and independently calculates the self-attention score of each window. The self-attention score of each window reflects the contribution of its corresponding feature to the subsequent classification task. This method not only ensures the accuracy of self-attention computation but also enhances the efficiency of the model in handling image features.

2.2. Residual attention

Residual structure is a popular network used to solve problems such as gradient vanishing and exploding during the training process of deep networks. It adds skip connections to pass along residual information, helping the network learn the differences better. By mixing the residual structure with a sliding window, it can make good use of both local and overall information, improving classification precision.

3. Methodology

To improve the accuracy of intelligent diagnosis in pathological image analysis, this paper proposes a pathology image classification algorithm. Figure 3 shows the structure diagram of the proposed algorithm. It mainly includes three parts: the first convolutional block, the residual attention block, and the classifier Block.

As shown in Figure 3, the input is a 3-channel RGB pathological image, denoted as X_{in}. Firstly, X_{in} applies the first convolution module twice to obtain the output X_{in+2}. Then, the X_{in+2} is sent the first residual attention module and obtain the output X_{in+8}. X_{in+8} again applies the residual attention module twice to obtain the output X_{in+13}. This method allows for capturing both local and global details of pathological images at the same time. After that, all the gathered features are sent to a classifier to predict the categories.

3.1. First convolutional block

The purpose of designing the first convolutional block is to extract local features from the original input image.

The computational process of the first convolutional block is described in Equation (2)

$$X_{in+1}=\text{Leaky ReLU}\,(\text{Batch Norm}\,(\text{Conv}\,(X_{in}))) \tag{2}$$

where X_{in} and X_{in+1} represent the input and the output data, respectively. $X_{in}\in R^{C\times W\times H}$, C, W and H represent the number of channels, the width, and the height of the input data, respectively. Conv (X_{in}) performs the two-dimensional convolutional operations by using a 3×3 convolutional kernel. BatchNorm normalizes the output of the convolutional layer to accelerate model training and enhance model stability. LeakyReLU is an activation function with a negative slope parameter.

3.2. Residual attention block

The residual attention block embeds the sliding window attention mechanism into a residual structure. The purpose is to extract contextual features and fuse the input local feature.

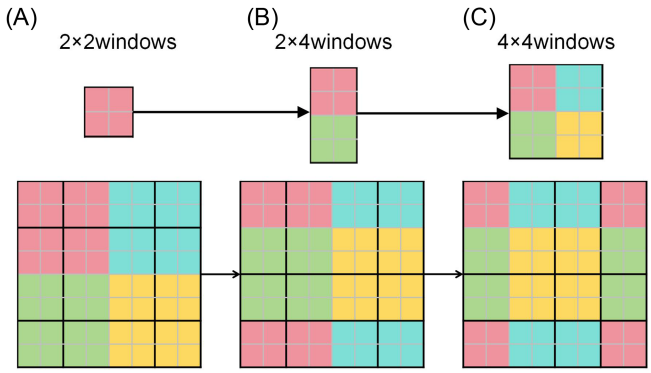

Figure 2. Schematic illustration of cyclic shifting operation

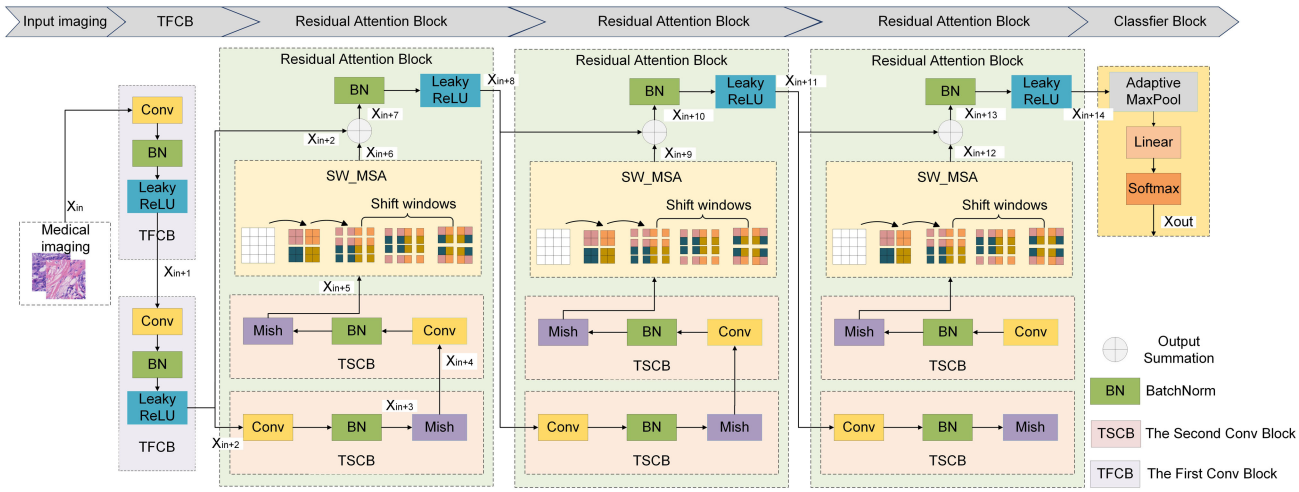

Figure 3. The proposed algorithm architecture diagram. X_{in} represents the input, and X_{out} represents the output.

In this module, the first step introduces the nonlinear transformation to enhance the model's feature representation capabilities through two secondary convolution blocks. The secondary convolution block is defined by Equations (3) and (4), comprising a convolutional layer, batch normalization layer, and Mish activation function layer. The convolutional layers allow residual attention modules to gradually stack and combine feature representations, thereby increasing the depth and breadth of the model.

In this module, the selected Mish activation function is shown in Equation (4) and illustrated in Figure 4. Mish activation function possesses unique characteristics compared to other activation functions: firstly, it has no upper limit, which ensures that the function value can increase infinitely, thereby maintaining a faster model training speed. The lower limit can reduce the risk of overfitting by limiting the function values to a finite range. Secondly, Mish is a nonlinear monotonic function that helps to confine values within a small range of negative values to stabilize gradient propagation. Lastly, the continuous smoothing property of Mish activation function near zero makes it easier to solve gradient vanishing and exploding problems in algorithm optimization.

$$X_{in+4} = \text{Mish}\left(\text{Batch Norm}\left(\text{Conv}\left(X_{in+2}\right)\right)\right) \tag{3}$$

$$X_{\text{in}+4} = X_{\text{in}+3} * \tanh\left(\ln(1 + e^{X_{in+3}})\right) \tag{4}$$

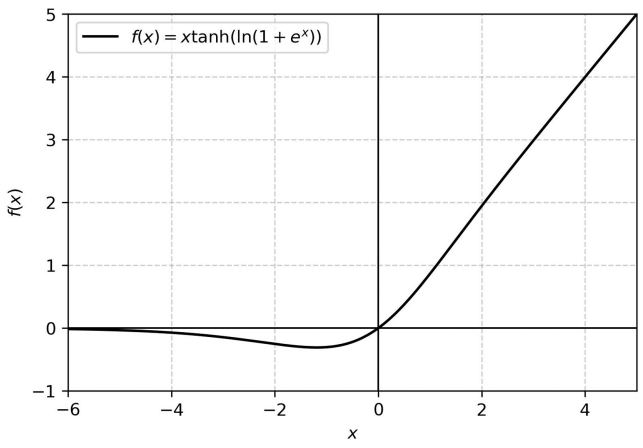

Figure 4. The Mish activation function

where the residual attention block inputs are the extracted local features X_{in+2}, from the first convolutional block, the X_{in+2}, is sent to the second convolution block for further feature extraction and feature enhancement using the Mish Activation Function. X_{in+3} is the output of a convolutional layer and batch normalization layer in the first second convolution block.

where X_{in+4} is calculated using the Mish activation function as defined by Equation (4). The second convolution block again is applied to the data X_{in+4} for obtaining the enhanced features X_{in+5} as shown in Equation (5). This nonlinear transformation enables better capturing of fine-grained features and high-level representations within the input data, as depicted in Figure 4.

$$X_{in+5} = \text{Mish}(\text{BatchNorm}(\text{Conv}(X_{in+4}))) \tag{5}$$

The residual attention module also introduces a sliding window attention layer, further enhancing the model's ability to perceive global contextual information. The sliding window attention layer employs the SW-MSA to handle relationships between input features of different positions and calculate the attention weights using Equation (1). In the windowed version of multi-head self-attention, the query, key, and value vectors are first used to weigh the value matrix and obtain a weighted result. These weighted results are then combined to produce the output of the windowed multi-head self-attention.

$$X_{in+6} = \text{SW} - \text{MSA}(X_{in+5}) \tag{6}$$

As shown in Equation (7), X_{in+7} is obtained by adding the output of the sliding window attention layer X_{in+6} to the output of the first convolution module X_{in+2}, and then, the multi-scale fusion feature X_{in+8} across layers is calculated using Equation (8). The specific calculation process is shown in Equations (7) and (8).

$$X_{in+7} = X_{in+6} + X_{in+2} \tag{7}$$

$$X_{in+8} = \text{LeakyReLU}(\text{BatchNorm}(X_{in+7})) \tag{8}$$

where X_{in+2}, X_{in+7}, and X_{in+8} denote the input to the residual block, the residual connections, and the output of the residual block, respectively. Additionally, the input and output at each stage are subject to the condition X_{in+2}, $X_{in+4}, X_{in+5}, X_{in+6}$, $X_{in+7}, X_{in+8} \in R^{C\times W\times H}$.

3.3. Classifier

The classifier aims to capture high-level semantic information from the input image and map it to a predefined category space, where it uses adaptive max pooling, flattening, and fully connected layers to process feature maps extracted with convolutional and residual blocks to predict the category of the input image.

The classifier has an adaptive max pooling layer, a flattening step, fully connected layers, and Softmax. The pooling layer changes the size of input feature maps to a fixed one, letting the model take in different-sized inputs and turn them into same-sized feature maps. The flattening step turns these feature maps into long vectors, taking away their spatial layout for the fully connected layers. These layers then convert the vectors into outputs for each category, matching the number of classes the model recognizes. The whole process is shown in Equation (9).

$$X_{out} = \mathrm{Linear}(\mathrm{AdaptiveMaxPool}(X_{in+14})) \tag{9}$$

where X_{in+14} represents the input to the Classifier, $X_{in+14} \in R^{C\times W\times H}$, C, W, and H represent the number of channels in the input, the width and the height of the input, respectively. While X_{out} denotes the one-dimensional array utilized for classification, $X_{out} \in R^{class\times 1\times 1}$, Class denotes the number of categories in the dataset used for model training.

By integrating the sliding window self-attention mechanism of Swin Transformer and the residual structure of ResNet, the proposed algorithm's feature extraction ability can be enhanced, and superior performance can be achieved in intelligent pathological image classification.

4. Experiments and Analysis

4.1. Datasets and experimental setup

To validate the proposed algorithm, we used two public datasets and one private dataset:

1) The lung pathology image dataset LC25000

In the LC25000 [26], there are 5000 images of benign tissues, 5000 images of lung squamous cell carcinoma tissues, and 5000 images of lung adenocarcinoma tissues respectively. These images are 768*768 pixels in size and stored in JPEG format. To support a binary classification task for lung pathology images, we randomly selected 2500 images from squamous cell carcinoma and adenocarcinoma tissues respectively, totaling 5000 malignant tissue images. These were paired with 5000 benign tissue images to form the binary classification task dataset. Figure 5(A) and (B) are examples of pathological images of lung squamous cell carcinoma and lung adenocarcinoma, respectively.

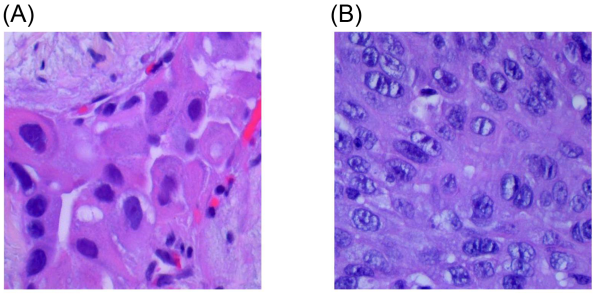

Figure 5. Example of pathological image dataset of lung. (A) The image of lung squamous cell carcinoma and (B) the image of lung adenocarcinoma.

2) The breast tissue pathology image dataset BreakHis

For the BreakHis [27], it comprises 2480 benign samples and 5429 malignant samples with a dimension of 700*460 pixels and stored in PNG format. To meet the experimental requirements of the algorithm, the BreakHis dataset is preprocessed for both binary and five-class classification tasks, resulting in breast tissue binary and five-class datasets. Benign breast tumors, including adenosis, fibroadenoma, phyllodes tumor, and tubular adenoma, are merged into a benign sample dataset, while malignant breast tumors, including ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma, are merged into a malignant sample dataset for the binary classification of breast tumors. Furthermore, the merged benign samples are combined with the four malignant samples to form a five-class classification task dataset for breast tissue pathology images. The pathological images of these eight categories of breast tissues are shown in Figure 6. Figure 6(A), (B), (C), and (D) are examples of pathological images of benign breast lesions including adenosis, fibroadenoma, lobulated tumor, and tubular adenoma, respectively. Figure 6(E), (F), (G), and (H) are examples of pathological images of ductal adenoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma, respectively, of malignant tumors.

3) A private dataset for breast cancer pathology tissue grading BreastCancerZZU3th

BreastCancerZZU3th is a histological grading dataset of breast tissue pathology images collected from the Third Affiliated Hospital of Zhengzhou University. The pathological tissue sections of 10 patients with grade I and II breast cancer were scanned with Motic digital slide scanner, and the corresponding full scan pathological images were obtained. The exported original slice images are in tif format, with a pixel size of 65500*65500. These images have been annotated by clinical pathologists, as shown in Figure 7. In Figure 7, (A) represents the scanned slice image, (B) represents the lesion range image annotated by the pathologist extracted from the original image, and (C) shows the preprocessed pathology image.

To validate the algorithm designed in this study, 1500 pathological images of Grade I and Grade II breast cancer were obtained through preprocessing methods such as cropping and filtering. Each image is 224*224 in size and in JPG format, as shown in Figure 8. Figure 8(A) and (B) represent pathological image examples of Grade I and Grade II breast cancer, respectively.

4.2. Evaluation metrics

This study utilizes evaluation metrics like accuracy, precision, recall, and $F1$ Score, with their respective calculation formulas provided as follows in Equations (10)–(13).

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{12}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{13}$$

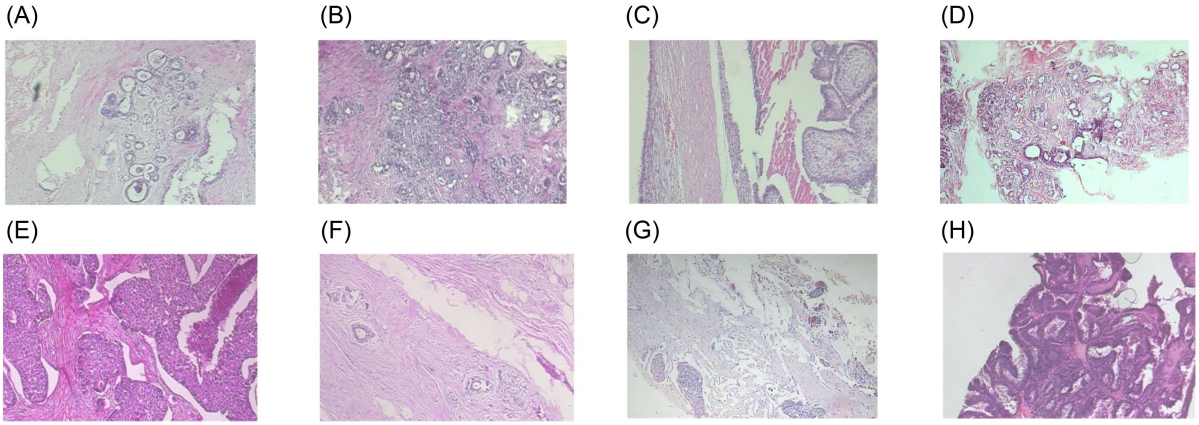

Figure 6. Examples of breast tissue pathological image. (A) Adenosis, (B) fibroadenoma, (C) lobulated tumor, (D) tubular adenoma, (E) ductal adenoma, (F) lobular carcinoma, (G) mucinous carcinoma, and (H) papillary carcinoma.

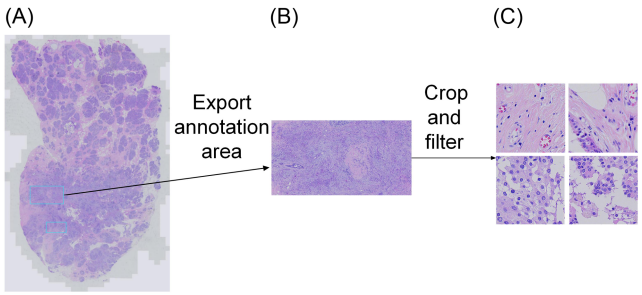

Figure 7. Example of extracting the lesion range from the original slice image of a private breast cancer dataset. (A) The scanned slice image, (B) the lesion area image annotated by the pathologists, and (C) the preprocessed pathology image.

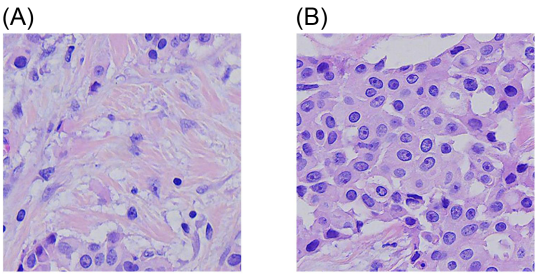

Figure 8. The pathological image examples of breast cancer. (A) Grade I and (B) Grade II.

where *TP*, *TN*, *FN*, and *FP* indicate true positive, true negative, false negative, and false positive prediction, respectively.

4.3. Comparative experiments

This paper compares the designed model with the following seven different models.

ResNet34: the classification model introducing residual structure in CNN.

ViT: the classic Transformer model.

Swin-tiny: the original model with sliding window self-attention mechanism.

The following comparison model replaces the sliding window self-attention module of the design model with different attention modules.

Residual + CBAM [28]: With convolutional block attention module (CBAM), a lightweight attention module capable of operating attention on spatial and channel dimensions.

Residual + NAM [29]: With the normalization-based attention module (NAM), introducing sparse weight penalties to enhance computational efficiency and utilizing regularization to suppress insignificant features.

Residual + ECA [30]: With regularization efficient channel attention (ECA), a channel attention mechanism into convolution operations to capture relationships between different channels and adaptively adjust channel feature weights.

Residual + MSA: With multi-head self-attention (MSA), the classic multi-head self-attention mechanism in Transformers.

This study conducted the following five comparative experiments:

Experiment 1 and Experiment 2 implemented a binary classification (benign vs. malignant)) and a three-class classification task for lung pathology images using the LC25000 dataset, respectively. The results of the quantitative comparison experiments on the lung pathological image datasets are shown in Tables 1 and 2, respectively.

It is clearly observed from the results in Tables 1 and 2, compared to the literature model Residual + MSA with the best comprehensive performance, our algorithm respectively achieves an improvement of 0.01, 0.011 in accuracy 0.012, 0.014 in precision, 0.008, 0.012 in recall, and 0.01, 0.015 in *F*1 score for the binary classification and the three-class classification task. These results indicate that our algorithm not only achieves the best classification performance in binary classification tasks for lung pathology images but also outperforms the comparative models in the three-class classification task, demonstrating the potential of this algorithm in multi-class classification tasks. Furthermore, compared with models combining other attention mechanisms and residual structures, our approach of integrating sliding window self-attention mechanism with residual modules demonstrates the superior effectiveness in improving the model's performance.

Experiment 3 did a binary classification (benign or malignant) using the BreakHis dataset for breast pathology images. Experiment 4, on the other hand, performed a five-class classification task with the same dataset. The results of the quantitative comparison experiments on the breast pathology image are shown in Tables 3 and 4, respectively.

As shown in Tables 3 and 4, our algorithm achieves improvements over the literature model Residual + MSA with the best comprehensive performance, respectively, in the binary and five-class classification of breast pathology images: 0.019, 0.026 in accuracy, 0.025, 0.035 in precision, 0.014, 0.027 in recall, and 0.019, 0.031 in *F*1 score.

Table 1. Experimental results for the binary classification task on lung pathological image dataset

Model	Accuracy	Precision	Recall	*F*1
ResNet34	0.925	0.918	0.926	0.922
ViT	0.933	0.942	0.929	0.936
Swin-tiny	0.966	0.959	0.965	0.962
Residual + CBAM	0.956	0.964	0.971	0.967
Residual + NAM	0.945	0.941	0.949	0.945
Residual + ECA	0.964	0.959	0.968	0.964
Residual + MSA	0.968	0.963	0.972	0.967
Ours	**0.978**	**0.975**	**0.980**	**0.977**

Table 2. Experimental results for the three-class classification task on lung pathological image dataset

Model	Accuracy	Precision	Recall	*F*1
Resnet34	0.894	0.898	0.874	0.886
ViT	0.879	0.881	0.868	0.875
Swin-tiny	0.897	0.897	0.885	0.892
Residual + CBAM	0.918	0.920	0.904	0.912
Residual + NAM	0.896	0.897	0.885	0.891
Residual + ECA	0.914	0.902	0.897	0.900
Resisual +MSA	0.921	0.914	0.929	0.922
Ours	**0.932**	**0.928**	**0.941**	**0.937**

Table 3. Experimental results table for the binary classification task on breast pathology image dataset

Model	Accuracy	Precision	Recall	*F*1
Resnet34	0.818	0.852	0.798	0.824
ViT	0.853	0.841	0.864	0.853
Swin-tiny	0.885	0.886	0.884	0.885
Residual + CBAM	0.924	0.920	0.927	0.924
Residual + NAM	0.882	0.883	0.881	0.882
Residual + ECA	0.891	0.872	0.906	0.889
Resisual + MSA	0.928	0.914	0.940	0.927
Ours	**0.947**	**0.939**	**0.954**	**0.946**

The experimental results indicate that our algorithm not only performs better in binary classification tasks for breast pathology images but also outperforms the comparative models in the five-class classification task.

4.4. Clinical data validation experiment

The histological grading of breast tissue is closely related to the prognosis of patients. At present, pathologists have divided breast cancer into three grades by observing pathological tissue sections under the microscope and scoring gland tube formation, mitosis count and nuclear atypia. The higher the grade, the worse the biological behavior and prognosis of breast cancer.

In order to verify the effectiveness of the proposed algorithm in the classification diagnosis of pathological images of early and mid-term breast cancer patients, Experiment 5 is conducted using the private datasets BreastCancerZZU3. Table 5 shows the comparative experimental results of the proposed algorithm and the seven literature method models in the previous section for the histological grading of breast tissue.

Table 5 shows our algorithm performs best in the two-level breast cancer histopathology classification. It works well on both public and private datasets, proving it is versatile and strong. Our algorithm also does great in tasks with multiple diseases and classes, meaning it is very robust.

Table 4. Experimental results for the five-class classification task on the breast histopathology image dataset

Model	Accuracy	Precision	Recall	*F*1
Resnet34	0.758	0.744	0.751	0.748
ViT	0.794	0.815	0.788	0.802
Swin-tiny	0.821	0.803	0.818	0.811
Residual + CBAM	0.802	0.791	0.806	0.799
Residual + NAM	0.811	0.821	0.811	0.816
Residual + ECA	0.784	0.798	0.778	0.788
Resisual + MSA	0.815	0.822	0.819	0.821
Ours	**0.841**	**0.857**	**0.846**	**0.852**

Table 5. Experimental results of histological grading of breast cancer on a private breast cancer dataset

Model	Accuracy	Precision	Recall	*F*1
Resnet34	0.936	0.942	0.933	0.938
ViT	0.952	0.950	0.946	0.948
Swin-tiny	0.960	0.964	0.958	0.962
Residual + CBAM	0.953	0.947	0.950	0.949
Residual + NAM	0.939	0.945	0.933	0.939
Residual + ECA	0.944	0.925	0.931	0.928
Resisual + MSA	0.959	0.945	0.952	0.949
Ours	**0.976**	**0.973**	**0.967**	**0.970**

Table 6. Ablation study results

Model	Accuracy	Precision	Recall	*F*1
Baseline	0.908	0.896	0.901	0.899
+Residual	0.944	0.951	0.969	0.960
+Residual(Mish)	0.958	0.963	0.972	0.967
+SW-MSA	0.971	0.966	0.972	0.969
Ours	**0.978**	**0.975**	**0.980**	**0.977**

4.5. Ablation study

To investigate the impact of the residual structures, Mish activation function, and the sliding window attention module on the performance enhancement of the proposed algorithm, we conduct an ablation study by creating variants of the proposed algorithm. In each variant, we selectively remove or replace one of the components from the proposed algorithm, resulting in the following models:

1) Baseline: The initial model of the proposed approach without residual structures and with the sliding window attention module replaced by a regular convolution module.
2) +Residual: Adding the residual structures to the Baseline.
3) +Residual(Mish): Replacing ReLU with Mish activation functions in +Residual.
4) +SW-MSA: Adding the sliding window attention module to the Baseline. The ablation experiment used the same dataset as Experiment 1, The results of experiments on the binary dataset of lung pathology image are shown in the Table 6.

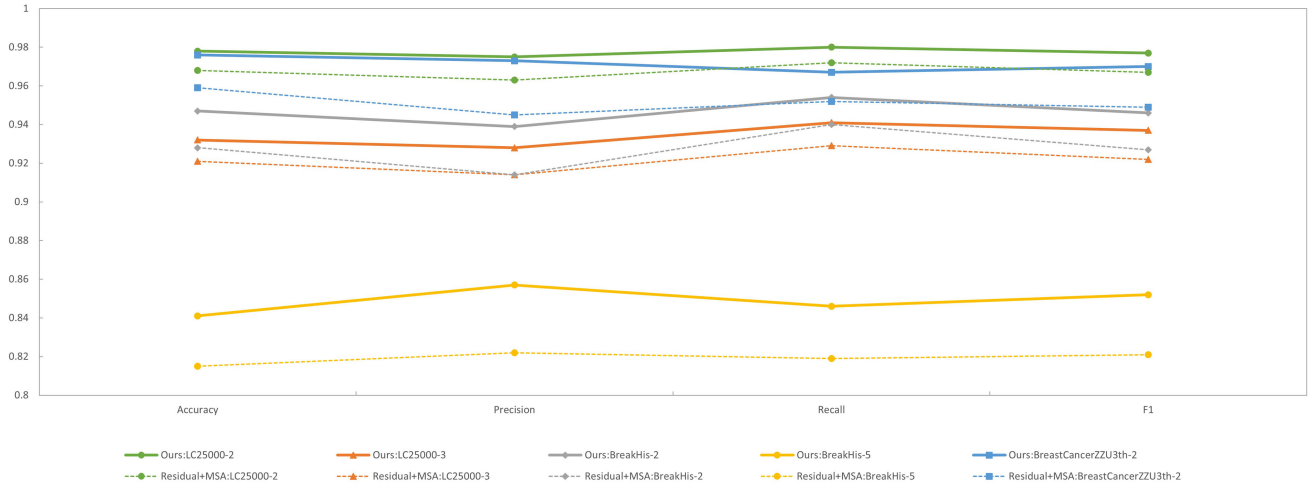

Figure 9. A model comparison chart

As shown in Table 6, the proposed algorithmic demonstrates positive impacts on the model's classification performance. The residual structure improves accuracy by 0.036, Mish activation function integrated with residual structure improves accuracy by 0.05, and the sliding window self-attention mechanism improves accuracy by 0.063.

Overall, the comprehensive improvement approach raises the baseline model's accuracy by 0.07. Furthermore, compared to the baseline, the various enhancements and the final model also show improvements across other metrics. These results indicate that the residual structure, Mish activation function, and sliding window self-attention module in the proposed algorithm are beneficial for pathological image classification tasks, contributing to higher accuracy and performance metrics in classification tasks.

4.6. Discussion and analysis

Based on the experimental results from Tables 1–5, our method is compared with the Resisual + MSA method with the best classification performance, as shown in Figure 9.

From Figure 9, we can clearly observe that our method outperforms the literature method in all five classification tasks. Our algorithm has shown good performance in both benign and malignant classification of breast pathological images, subtype multi classification and breast cancer patient grading. The model achieved accuracies of 0.987 and 0.932 for binary and three-class lung pathology image tasks, respectively, and accuracies of 0.947 and 0.841 for binary and five-class breast pathology image tasks, respectively, indicating good generalization and robustness of the algorithm.

Among the three types of classification tasks (two, three, and five), whether it is a public dataset or a private dataset, our algorithm achieved the best results in the two classification tasks. In particular, the accuracy, precision, recall, and $F1$ of grade 1 and grade 2 clinical verification of breast cancer diagnosis, which are difficult for pathologists to identify, reach 97.6%, 97.3%, 96.7% and 97% respectively, validating its effectiveness in practical applications.

Its theoretical basis lies in: CNNs consist of multiple convolutional and pooling layers, focusing primarily on extracting local features, but they have limited ability to perceive global information. Moreover, CNNs cannot directly model the positional relationships between pixels, thus overlooking spatial dependencies among pixels. In contrast, Transformer is a model based on self-attention mechanism, which can better capture global information and positional relationships. However, Transformer lacks in extracting local features, and the traditional Transformer model's self-attention mechanism requires computing fully connected attention weight matrices, resulting in higher computational complexity. This limitation restricts the application of Transformer in large-scale pathological image classification tasks.

Our algorithm combines sliding window self-attention and residual structure, using Mish to enhance the nonlinear expression ability of features, which helps extract local and global features from pathological images to improve classification accuracy.

On the other hand, as the number of classification tasks increases, the classification performance of the algorithm decreases. How, it also shows promising results in three-class and five-class classification tasks for lung and breast pathological images, indicating its potential in multi-class tasks. This indicates that the proposed algorithm has outstanding potential in the early diagnosis and fine classification of various common tumors.

5. Conclusion

In this paper, we have proposed an intelligent classification algorithm that combines CNN and Transformer for pathological images. In order to address the issues of algorithm accuracy and generality, we conducted comparative experiments on lung pathology datasets and breast pathology datasets for binary and multi-class classification tasks. Extensive experiments showed that the proposed algorithm work outperforms most previous SOTA methods based on ResNet with residual modules, Swin Transformer with sliding window attention mechanism, and other models combining residual attention mechanisms. It shows improvements to outperform the baseline by 7.0%, 7.9%, 7.9%, and 7.8% in accuracy, precision, recall, and $F1$ score for a binary task. Unlike previous Swin Transformers that used sliding window attention mechanism, our algorithm has designed a residual module by introducing a sliding window multi head attention mechanism, which enhances the ability to extract contextual information and reduce computational complexity. This algorithm is particularly suitable for high-resolution whole-slice pathological images.

Funding Support

This work was supported by Tianjian Laboratory of Advanced Biomedical Sciences and Zhengzhou collaborative innovation major special project under Grant 20XTZX11020.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

Author Contribution Statement

Huiqin Jiang: Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Zhiheng Tong:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Xiaonan Yang:** Conceptualization. **Fangjie Zhao:** Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jinhong Tan:** Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Xing Dong:** Investigation, Resources, Data curation. **Zhenying Ban:** Investigation, Resources, Data curation. **Xianxu Zeng:** Investigation, Resources, Supervision. **Xin Zhao:** Investigation, Resources, Supervision, Project administration, Funding acquisition. **Ling Ma:** Software, Validation, Formal analysis, Visualization.

References

[1] Marini, N., Marchesin, S., Otálora, S., Wodzinski, M., Caputo, A., van Rijthoven, M., ..., & Atzori, M. (2022). Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. *npj Digital Medicine*, *5*(1), 102. https://doi.org/10.1038/s41746-022-00635-4

[2] Komura, D., & Ishikawa, S. (2019). Machine learning approaches for pathologic diagnosis. *Virchows Archiv*, *475*, 131–138. https://doi.org/10.1007/s00428-019-02594-w

[3] Song, Z., Zou, S., Zhou, W., Huang, Y., Shao, L., Yuan, J., ..., & Shi, H. (2020). Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nature Communications*, *11*(1), 4294. https://doi.org/10.1038/s41467-020-18147-8

[4] George, Y. M., Zayed, H. H., Roushdy, M. I., & Elbagoury, B. M. (2014). Remote computer-aided breast cancer detection and diagnosis system based on cytological images. *IEEE Systems Journal*, *8*(3), 949–964. https://doi.org/10.1109/JSYST.2013.2279415

[5] Gupta, V., & Bhavsar, A. (2017). Breast cancer histopathological image classification: Is magnification important? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 17–24. https://doi.org/10.1109/CVPRW.2017.107

[6] Peikari, M., Salama, S., Nofech-Mozes, S., & Martel, A. L. (2018). A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific Reports*, *8*(1), 7193. https://doi.org/10.1038/s41598-018-24876-0

[7] Trivizakis, E., Ioannidis, G. S., Souglakos, I., Karantanas, A. H., Tzardi, M., & Marias, K. (2021). A neural pathomics framework for classifying colorectal cancer histopathology images based on wavelet multi-scale texture analysis. *Scientific Reports*, *11*(1), 15546. https://doi.org/10.1038/s41598-021-94781-6

[8] Alqudah, A., & Alqudah, A. M. (2022). Sliding window based support vector machine system for classification of breast cancer using histopathological microscopic images. *IETE Journal of Research*, *68*(1), 59–67. https://doi.org/10.1080/03772063.2019.1583610

[9] Liu, M., Hu, L., Tang, Y., Wang, C., He, Y., Zeng, C., ..., & Huo, W. (2022). A deep learning method for breast cancer classification in the pathology images. *IEEE Journal of Biomedical and Health Informatics*, *26*(10), 5025–5032. https://doi.org/10.1109/JBHI.2022.3187765

[10] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). Breast cancer histopathological image classification using convolutional neural networks. In *International Joint Conference on Neural Networks*, 2560–2567. https://doi.org/10.1109/IJCNN.2016.7727519

[11] Koné, I., & Boulmane, L. (2018). Hierarchical ResNeXt models for breast cancer histology image classification. In *Image Analysis and Recognition: 15th International Conference*, 796–803. https://doi.org/10.1007/978-3-319-93000-8_90

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[13] Teramoto, A., Tsukamoto, T., Yamada, A., Kiriyama, Y., Imaizumi, K., Saito, K., & Fujita, H. (2020). Deep learning approach to classification of lung cytological images: Two-step training using actual and synthesized images by progressive growing of generative adversarial networks. *PLoS ONE*, *15*(3), e0229951. https://doi.org/10.1371/journal.pone.0229951

[14] Phankokkruad, M. (2021). Ensemble transfer learning for lung cancer detection. In *DSIT 2021: 2021 4th International Conference on Data Science and Information Technology*, 438–442. https://doi.org/10.1145/3478905.3478995

[15] Adu, K., Yu, Y., Cai, J., Owusu-Agyemang, K., Twumasi, B. A., & Wang, X. (2021). DHS-CapsNet: Dual horizontal squash capsule networks for lung and colon cancer classification from whole slide histopathological images. *International Journal of Imaging Systems and Technology*, *31*(4), 2075–2092. https://doi.org/10.1002/ima.22569

[16] Srikantamurthy, M. M., Rallabandi, V. S., Dudekula, D. B., Natarajan, S., & Park, J. (2023). Classification of benign and malignant subtypes of breast cancer histopathology imaging using hybrid CNN-LSTM based transfer learning. *BMC Medical Imaging*, *23*(1), 19. https://doi.org/10.1186/s12880-023-00964-0

[17] Zou, Y., Zhang, J., Huang, S., & Liu, B. (2022). Breast cancer histopathological image classification using attention high-order deep network. *International Journal of Imaging Systems and Technology*, *32*(1), 266–279. https://doi.org/10.1002/ima.22628

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ..., & Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems*.

[19] Alotaibi, A., Alafif, T., Alkhilaiwi, F., Alatawi, Y., Althobaiti, H., Alrefaei, A., ..., & Nguyen, T. (2023). ViT-deiT: An ensemble model for breast cancer histopathological images classification. In *1st International Conference on Advanced Innovations in Smart Cities*, 1–6. https://doi.org/10.1109/ICAISC56366.2023.10085467

[20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., . . . , & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint:2010.11929*.

[21] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, 10347–10357.

[22] Thomas, A. M., Adithya, G., Arunselvan, A. S., & Karthik, R. (2022). Detection of breast cancer from histopathological images using image processing and deep-learning. In *Third International Conference on Intelligent Computing Instrumentation and Control Technologies*, 1008–1015. https://doi.org/10.1109/ICICICT54557.2022.9917784

[23] Tummala, S., Kim, J., & Kadry, S. (2022). BreaST-Net: Multi-class classification of breast cancer from histopathological images using ensemble of swin transformers. *Mathematics*, *10*(21), 4109. https://doi.org/10.3390/math10214109

[24] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., . . . , & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

[25] Chattopadhyay, S., Dey, A., Singh, P. K., & Sarkar, R. (2022). DRDA-Net: Dense residual dual-shuffle attention network for breast cancer classification using histopathological images. *Computers in Biology and Medicine*, *145*, 105437. https://doi.org/10.1016/j.compbiomed.2022.105437

[26] Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A., & Mastorides, S. M. (2019). Lung and colon cancer histopathological image dataset (LC25000). *arXiv Preprint:1912.12142*.

[27] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, *63*(7), 1455–1462. https://doi.org/10.1109/TBME.2015.2496264

[28] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Computer Vision – ECCV 2018: 15th European Conference*, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

[29] Liu, Y., Shao, Z., Teng, Y., & Hoffmann, N. (2021). NAM: Normalization-based attention module. *arXiv Preprint: 2111.12419*.

[30] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155

How to Cite: Jiang, H., Tong, Z., Yang, X., Zhao, F., Tan, J., Dong, X., . . . , & Ma, L. (2025). A Universal Intelligent Classification Algorithm for Pathological Images Based on Sliding Window Attention Mechanism. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN52025163