



# A deep multi-branch attention model for histopathological breast cancer image classification

Rui Ding<sup>1</sup> · Xiaoping Zhou<sup>1</sup> · Dayu Tan<sup>1</sup>  · Yansen Su<sup>1</sup> · Chao Jiang<sup>2</sup> · Guo Yu<sup>3</sup> · Chunhou Zheng<sup>1</sup>

Received: 3 November 2023 / Accepted: 10 February 2024  
© The Author(s) 2024

## Abstract

Since the impressive superior performance demonstrated by deep learning methods is widely used in histopathological image analysis and diagnosis, existing work cannot fully extract the information in the breast cancer images due to the limited high resolution of histopathological images. In this study, we construct a novel intermediate layer structure that fully extracts feature information and name it DMBANet, which can extract as much feature information as possible from the input image by up-dimensioning the intermediate convolutional layers to improve the performance of the network. Furthermore, we employ the depth-separable convolution method on the Spindle Structure by decoupling the intermediate convolutional layers and convolving them separately, to significantly reduce the number of parameters and computation of the Spindle Structure and improve the overall network operation speed. We also design the Spindle Structure as a multi-branch model and add different attention mechanisms to different branches. Spindle Structure can effectively improve the performance of the network, the branches with added attention can extract richer and more focused feature information, and the branch with residual connections can minimize the degradation phenomenon in our network and speed up network optimization. The comprehensive experiment shows the superior performance of DMBANet compared to the state-of-the-art method, achieving about 98% classification accuracy, which is better than existing methods. The code is available at <https://github.com/Nagi-Dr/DMBANet-main>.

**Keywords** Histopathological images · Breast cancer classification · Convolutional neural networks · Attention mechanism

---

Rui Ding and Xiaoping Zhou have contributed equally to this work.

✉ Dayu Tan  
tandayu19@163.com

Rui Ding  
congroodr@163.com

Xiaoping Zhou  
m15315118203@163.com

Yansen Su  
suyansen@ahu.edu.cn

Chao Jiang  
chaojiang@mail.ecust.edu.cn

Guo Yu  
gysearch@163.com

Chunhou Zheng  
zhengch99@126.com

<sup>1</sup> Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei 230601, China

## Introduction

The task of automatic classification for breast cancer histopathological images is a theme that deserves our deliberation. It is worthwhile for researchers to contrive a model to effectuate the accurate evaluation of breast cancer and automatic classification of breast cancer histopathological images. In recent years, deep learning methods have made significant progress and achieved remarkable performance in the field of computer vision and image processing [36, 44], inspiring many scholars to implement the technique for histopathological image classification [1]. Convolutional neural networks (CNNs) are not only the most widely used type of deep learning networks, but they also have superior performance in

<sup>2</sup> Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing 100124, China

<sup>3</sup> Institute of Intelligent Manufacturing, Nanjing Tech University, Nanjing 211816, China

image classification and image feature extraction [2], which have laid the foundation for the application of convolutional neural networks in histopathological image classification.

Human vision can easily ignore low-value information and quickly find valuable key information. Inspired by the human visual system, many works [8, 10] have apply attention mechanism to the classification of histopathological images. There are two main types of attentional mechanisms [39]: 1) Channel Attention, such as SeNet [7], which focuses on allowing the network to learn what to pay attention; 2) Spatial Attention, such as self-attention [24], which focuses on allowing the network to learn where to pay attention. Furthermore, there are some composite attention mechanisms, which combine channel attention and spatial attention, such as CBAM [31]. Due to the high resolution of histopathological images, using the attention mechanism can help us quickly target to the valuable regions and effectively improve the performance of the network.

In addition, due to limited GPU resources, it is not possible to directly input high-resolution histopathological images into the CNN for classification. The high resolution not only results in substantial computational costs but also significantly prolongs the network's training time, requiring preprocessing of raw data [42, 43]. However, downsampling the image to a lower resolution is not practical either since it results in significant loss of valuable image information, which is especially critical in medical imaging, where the available data is already limited. Therefore, we developed a deep spindle structure in our network that can effectively address this challenge by breaking down the image into multiple patches, and then conducting feature extraction and processing in parallel on each patch to fully capture the salient information.

To address the aforementioned challenges, this study proposes a deep multi-branch attention model that optimizes information extraction in three ways: (a) the multi-branch structure consists of three different branches, enabling the network to focus on a wider range of information scales; (b) the addition of spatial attention and channel attention to different branches allows the network to pay attention to different key information; and (c) through the deep spindle structure, the middle layer's up-dimensioning significantly enhances the network's information extraction ability. To preprocess the original histopathological images, we cut them into small patches and input these patches into the network. Evaluated on the PathoIMG data set, our method outperforms existing state-of-the-art methods and significantly improves the classification accuracy of breast cancer histopathological images.

The main contributions of this study are summarized as follows:

1. This study constructs the Deep Spindle Structure that extracts as much information as possible from pathological images to reduce the waste of feature information. It improves the accuracy without additionally increasing the number of parameters of the network, facilitating the deployment of the network on platforms with limited resources.
2. We design a multi-branch model through add residual connections to the Deep Spindle Structure that allows the network to heed more scales of information and extract richer features, which effectively solves the degradation phenomenon of the network so that the network can stack deeper and get a better result.
3. We add channel and spatial attention mechanism to the multi-branch model to help the network better notice and extract key information. Channel and spatial attention in different branches make the network pay attention to more noteworthy channels and information in individual channels, respectively, which allows the network to extract more critical information hidden in the pathological images and effectively improve the classification accuracy of the network.

The remainder of this study is organized as follows. We start with a review of the related work that contain a number of deep ConvNets applied to the field of histopathological image classification and efficient attention mechanisms in the section "[Related work](#)". We present the proposed a novel deep ConvNets which combined with multi-branch structure and attention mechanism, and explain the specific design ideas in detail in the section "[Proposed method](#)". In the section "[Experiments](#)", We compare our approach with several current state-of-the-art deep neural networks based on the PathoIMG breast cancer data set. The conclusions of this study and future works are given in the section "[Conclusion](#)".

## Related work

In recent years, researchers have been exploring the application of convolutional neural networks to histopathological image classification of breast cancer, leveraging CNN's excellent performance in computer vision and natural language processing. Breakhis [3] is a breast cancer histopathological image data set with two major classes, benign and malignant, each subdivided into four subclasses. The author combines AlexNet networks and different integration strategies and then perform classification on the Breakhis data set with improvement in classification accuracy over traditional machine learning methods. BiCNN [4] treats class labels and subclass labels in the data set as prior knowledge and uses a combination of data augmentation methods and migration learning strategies in the training process. BiCNN

effectively improves the robustness and generalization of the network. Yan et al. [5] sliced the original histopathological images into small patches and utilized GoogleNet V3 and bi-directional LSTM for classification. In addition, they published the PathoIMG data set in their paper. Previous approaches have overlooked the intricate high-dimensional relationships among multiple views. In response, Pan et al. [45] introduce a novel method called Low Rank Tensor Regularized Graph Fuzzy Learning (LRTGFL) for processing multi-view data. Although previous works have made various efforts to improve classification accuracy, unfortunately, the classification accuracy of these works is still unsatisfactory.

The attention mechanism has been successfully applied to various image classification tasks [7] because it enables the network to ignore low-value information easily and learn more valuable features. Attention by Selection [8] uses DeNet, which consists of a hard attention mechanism, to select valuable regions in the original image, and then passes these valuable regions as input to the soft attention dominated SaNet network for further processing and feedback. ARL-CNN [9] (attention residual learning convolutional neural network) is a method for skin lesion classification. It consists of multiple residual blocks incorporating an attention mechanism and using a self-attentive mechanism, that widely used in the natural language processing. DA-MIDL [10] (dual attention multi-instance deep learning network) uses spatial attention to extract useful information and combines it with multi-instance learning (MIL) pooling to enable the network. DA-MIDL achieve better classification performance in terms of accuracy and generalization. CA-Net [11] is an attention-based network model that makes extensive use of multiple attentions. multiple attentions allowing the network to simultaneously attend to the most important spatial location and channel information by combining different attention mechanisms to obtain superior network performance and interpretability.

Since AlexNet [12] popularize deep convolutional neural networks by winning the ImageNet challenge: ILSVRC 2012 [13], excellent convolutional neural networks start to emerge [14–17]. GoogleNet [18] demonstrate the superiority of the multi-branch structure by utilizing the width structure of Inception, leading to its widespread use in subsequent network designs. Moreover, it introduce the depthwise separable convolution to significantly reduce the number of model parameters, which laid the foundation for the development of lightweight network models. After ResNet [19] proposed residual connectivity, it has been widely used in the design of network models, which improves the performance of the network on the one hand, and effectively solves the problem of network degradation on the other hand, allowing the network to be stacked very deeply.

MobileNet [20, 21, 37] and ShuffleNet [22, 38] achieve network lightweighting with guaranteed network performance, the former mainly by deep separable convolution and inverse residual structure, while the latter reduces the number of parameters of the network using Channel Shuffle in group convolution. EfficientNet [23] uses Neural Architecture Search (NAS) technique in the parameter space to searches for a rationalized configuration of three parameters: image input resolution, network depth, and Channel width, and then uses the compound scaling method to obtain a series of well-performing networks.

Moreover, the transformer structure has recently started to show great strength, and ViT [24] has become a milestone work in the application of transformer in CV field because of its simple, scalable and effective structure, which has triggered subsequent related research. ConvNeXt [25], based on ResNet-50 and ResNet-200, borrowed the advanced ideas of swin transformer [26] from five perspectives: Macro Design, Group Convolution & Deep Separable Convolution, Inverted Bottleneck layer, Large Kernel Sizes, and Micro Design, respectively, to effectively improve the performance of CNNs.

## Proposed method

In this section, we describe in detail the methodological strategies applied to the core layer of the DMBANet in four dimensions: (1) Specific design of Deep Spindle Structure; (2) multi-branch attention; (3) optimizer and learning rate decay; and (4) overall structure of DMBANet. We describe the design ideas for the DMBANet in the previous three parts, and show the overall structure of the DMBANet in the last part.

### Deep spindle structure

Bottleneck block is a structure widely used in many neural networks, such as ResNet [19]. It reduces the dimensionality of the input feature map by first  $1 \times 1$  convolutional kernel, and up-dimensions the output feature map by end  $1 \times 1$  convolutional kernel. The bottleneck block has many advantages, such as it can stack the network easily without calculating the size of the feature map when we want to deepen the network structure.

Regrettably, Although bottleneck block gets some benefits by first reducing the dimension and then raising it. But in this process, the ability of bottleneck block to extract feature information is weakened. Thus, the first thing we have to do is to improve the bottleneck block while preserving the crucial residual connection. We have designed Deep Spindle structure that is the opposite of bottleneck block, it up-dimension the input feature map in first  $1 \times 1$  convolutional kernel, and

**Table 1** Specific parameters of Bottleneck Module and Spindle Structure are shown in the table, where DW conv means depthwise convolution and PW conv means pointwise convolution

Module	Input	Operator	Output
Bottleneck Module	$H \times W \times 4C$	$1 \times 1$ conv, ReLU	$H \times W \times C$
	$H \times W \times C$	$3 \times 3$ conv, ReLU	$H \times W \times C$
Spindle Module	$H \times W \times C$	$1 \times 1$ conv, ReLU	$H \times W \times 4C$
	$H \times W \times C$	$1 \times 1$ conv, ReLU6	$H \times W \times 4C$
	$H \times W \times 4C$	$3 \times 3$ DW conv, ReLU6	$H \times W \times 4C$
	$H \times W \times 4C$	Linear $1 \times 1$ PW conv	$H \times W \times C$

downscale them by last  $1 \times 1$  convolutional kernel. Because the number of channels in the middle  $3 \times 3$  convolutional kernel is high enough, Deep Spindle structure allows the input features to be fully extracted and the original information to be more fully utilized. However, since Deep Spindle structure performs the up-dimension first, it will inevitably increase the number of Spindle Structure parameters. The detailed parameters of the two modules are shown in Table 1.

Due to the dramatic increase in the number of parameters in Spindle Structure, we have to combine the depth-separable convolution for feature extraction. Depth-separable convolution is a key component of many neural networks with excellent performance [18, 20, 30]. The basic idea of it is to replace the full convolution operator with a decomposed convolution operator that decomposes the convolution into two separate layers. The first layer, known as the depthwise convolution, employs a group convolution where the convolution kernel is fully decoupled. This lightweight convolution extracts features by applying a single filter to each input channel. The second layer is a  $1 \times 1$  convolution referred to as pointwise convolution, which recouples the output features extracted by the decoupled convolution kernel. The application of this factorization yields a significant reduction in both computational requirements and model size. Figure 1 shows how the standard convolution is decomposed into depthwise convolution and pointwise convolution.

The standard convolutional layer takes a feature mapping  $S$  of size  $D_s \times D_s \times M$  as input, uses a convolutional kernel  $K$  of size  $D_k \times D_k \times M \times N$  to extract feature information from the input  $S$ , and produces a feature mapping  $E$  of size  $D_e \times D_e \times N$  as output. Where  $D_s$  is the spatial length and width of the input feature  $S$ ,  $M$  is the number of channels of the input feature  $S$ ,  $D_k$  is the spatial length and width of the convolutional kernel  $K$ ,  $D_e$  is the spatial length and width of the output feature  $E$ , and  $N$  is the number of channels of the output feature  $E$ .

The output feature mapping for standard convolution is calculated as follows:

$$E_{k,n} = \sum_{i,j,m} (K_{i,j,m,n} \times S_{k+i-1,k+j-1,m}). \quad (1)$$

The cost of the standard convolution is

$$C_{st} = D_{ker} \times D_{ker} \times M \times N \times D_{in} \times D_{in}, \quad (2)$$

where the computational cost depends multiplicatively on the number of input channels  $M$ , the number of output channels  $N$ , the convolution kernel size  $D_{ker} \times D_{ker}$  and the size of the input feature mapping  $D_{in} \times D_{in}$ .

Depth-separable convolution consists of two parts: depthwise convolution and pointwise convolution. We first use depthwise convolution to apply a single filter to each input channel to extract the features in the input channel individually. Pointwise convolution (a simple  $1 \times 1$  convolution) is then used to couple the new output combinations. In this paper, Batch normalization and the Rectified Linear Unit are used for both depthwise convolution and pointwise convolution.

The computation of the feature mapping for depthwise convolution can be written as

$$\hat{E}_{k,n} = \sum_{i,j} (\hat{K}_{i,j,m} \times S_{k+i-1,k+j-1,m}), \quad (3)$$

where  $\hat{K}$  is a depthwise convolution kernel of size  $D_k \times D_k \times m$ , where the  $m$ th filter in the depthwise convolution kernel  $\hat{K}$  is applied to the  $m$ th channel in the input feature  $S$  and produces the  $m$ th channel of the output feature mapping  $E$  after convolution.

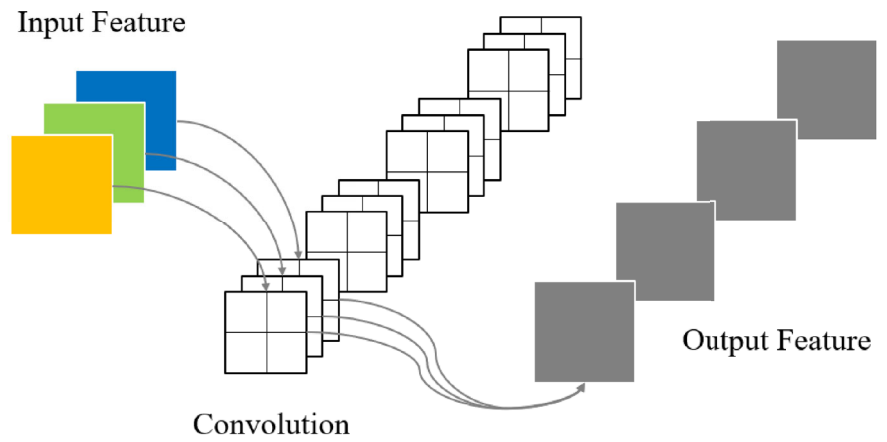
The cost of depthwise convolution and pointwise convolution is

$$C_{dw} = D_{ker} \times D_{ker} \times M \times 1 \times D_{in} \times D_{in}, \quad (4)$$

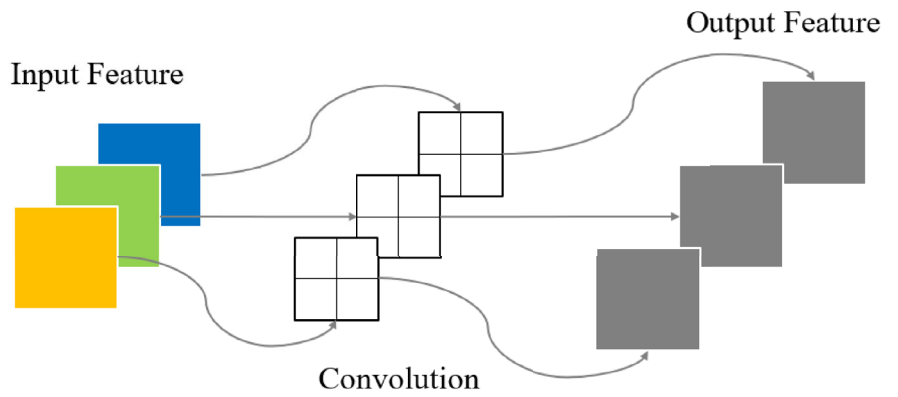
$$C_{pw} = 1 \times 1 \times M \times N \times D_{in} \times D_{in}. \quad (5)$$

In the comparison with standard convolution, depthwise convolution is very effective. However, depthwise convolution only convolves the decoupled input channels and cannot recouple the channel correlations between these feature information. Therefore, to generate these new feature combinations, an additional pointwise convolution layer is needed to couple the linear combinations of the depthwise convolution outputs through a  $1 \times 1$  convolution kernel.

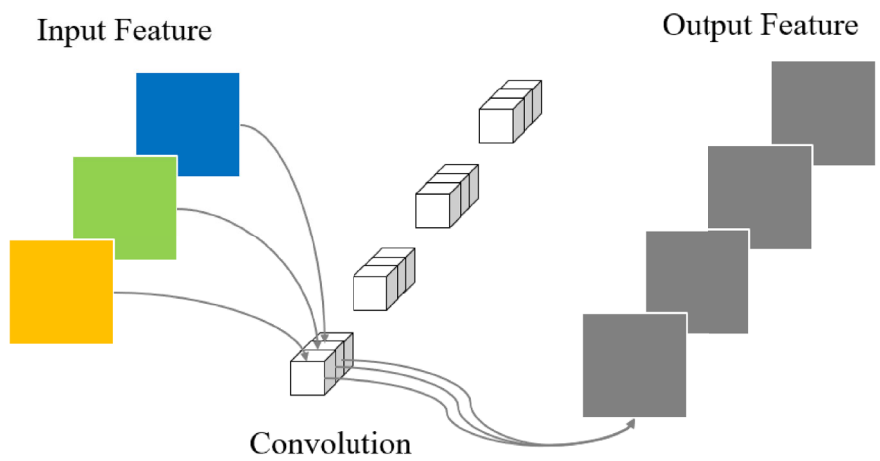
**Fig. 1** Illustration of the three convolution methods



(a) standard convolution



(b) depth-wise convolution



(c) point-wise convolution



The two components, depthwise convolution and pointwise convolution, are collectively referred to as depth-separable convolution, which originally introduced in Xception [18]. Overall cost of depth-separable convolution is

$$C_{dep} = C_{dw} + C_{pw}. \quad (6)$$

By decomposing the standard convolution into a two-step process of depthwise convolution and pointwise convolution, we can reduce the computational effort by

$$Ratio = \frac{C_{dep}}{C_{st}} = \frac{1}{N} + \frac{1}{D_{ker}^2}. \quad (7)$$

In this study, the depth-separable convolution use a convolution kernel of size  $3 \times 3$ , which is 8–9 times less computationally intensive than standard convolution, but only slightly less accurate. It must be noted that factorization in space does not consume much computational resources. Therefore, depth-separable convolution only needs a little computational resources to reduce the number of parameters significantly.

## Multi-branch attention

In recent years, more and more researchers have noticed the excellent performance of attention mechanism and applied it to various application scenarios of deep learning [7, 27, 28]. In different types of tasks, such as image processing, speech recognition or natural language processing, attention mechanism can be well integrated and perform well. In addition, since GoogleNet proposed the Inception [17] structure in 2015 and found that neural networks incorporating multi-scale feature information can perform well, Inception-like multi-branch structure has become a regular guest in various network models [29]. The residual connection proposed in ResNet has been widely used in various network models, and the residual connection is a typical multi-branch structure.

Naturally, we integrated the concept of a multi-branch structure into the design of the Deep Spindle architecture. Specifically, we incorporated both channel attention and spatial attention [31] to account for different types of attention and inserted them into separate branches. Furthermore, we added a residual connection to the network to further enhance its overall performance.

The Channel Attention Module, depicted in Fig. 2a, operates as follows: the input feature map undergoes global max pooling and global average pooling based on its width and height, respectively. These features are then processed through a shared multilayer perceptron. The features processed by the Max pooling and Average Pooling operations within the shared multilayer perceptron undergo an element-

wise summation operation, followed by a sigmoid activation operation to generate the final channel attention feature map.

In simpler terms, channel attention compresses the feature map in the spatial dimension, generating a functional one-dimensional vector. To achieve this spatial compression, we employ techniques like average pooling and maximum pooling, gathering spatial information within the feature map. These compressed features are then directed to the shared multilayer perceptron, further compressing the spatial dimensions of the input feature maps. The resulting vectors undergo elementwise summation and merging, creating a channel attention graph. In the feature map, channel attention enables our model to focus on elements that are more critical and important, this is beneficial for the model to learn information related to categories that require heightened attention.

Channel Attention can be expressed as follows:

$$CA(F) = \sigma(FC_2(FC_1(F_{Max})) + FC_2(FC_1(F_{Avg}))), \quad (8)$$

where  $\sigma$  denotes the Sigmoid activation function,  $F_{Max}$  and  $F_{Avg}$  denote the input that has undergone the MaxPool and AvgPool operations, Shared multilayer perceptron consists of  $FC_1$  and  $FC_2$ ,  $F_{Max}$  and  $F_{Avg}$  share the parameters of  $FC_1$  and  $FC_2$ .

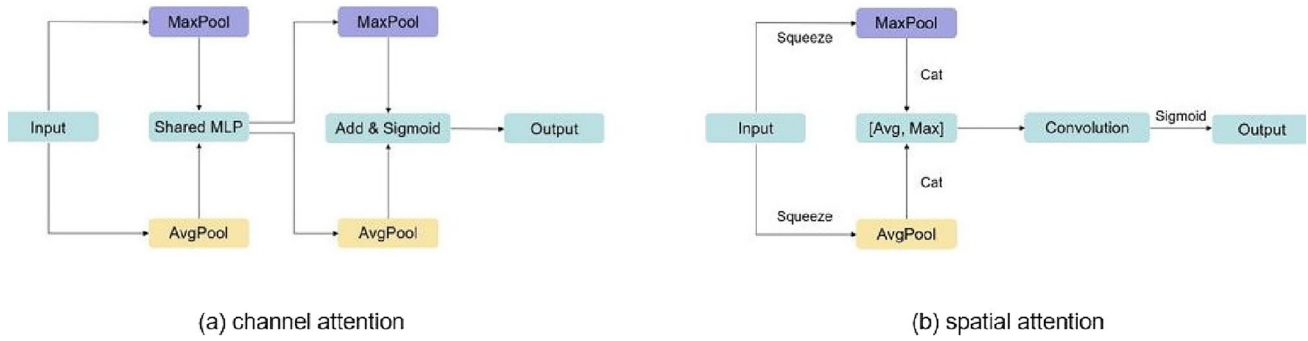
The Spatial Attention Module, as illustrated in Fig. 2b, operates in the following manner: first, global max pooling and global average pooling are applied based on the channel. These results are concatenated based on the channel to generate the spatial attention feature map using a sigmoid activation function. Finally, this feature map is multiplied by the input feature of the module to obtain the final generated feature.

The spatial attention module efficiently condenses channels by employing both average pooling and maximum pooling operations across the channel dimensions. To be more specific, the average pooling operation calculates the mean across all channels a number of times corresponding to the product of the height and width. Subsequently, the resulting feature maps from the preceding channels are merged, or duplicated in the case of a single channel, resulting in a two-channel feature map. Channel attention plays a crucial role in enabling the network to discern valuable location information, ensuring the extraction of critical texture details.

Spatial Attention Module can be expressed as follows:

$$SA(F) = \sigma(Ker_{7 \times 7}(Avg \Leftarrow Max_F)), \quad (9)$$

where  $\sigma$  denotes a sigmoid operation,  $Ker_{7 \times 7}$  denotes a convolution kernel of size  $7 \times 7$ ,  $Avg \Leftarrow Max_F$  indicates that the MaxPool operation is performed on the input  $F$  first, and then perform the AvgPool operation.



**Fig. 2** Overview of channel attention and spatial attention. The purpose of **a** channel attention is make the network pay attention to those channels which contain critical information. **b** Spatial attention can play a complementary role to channel attention, it pay attention to critical information on key channels

**Optimizer and learning rate decay**

The role of an optimizer in deep learning is to guide the backpropagation process by appropriately updating each parameter of the loss (objective) function in the correct direction, with the aim of minimizing the value of the loss function and approaching the global minimum.

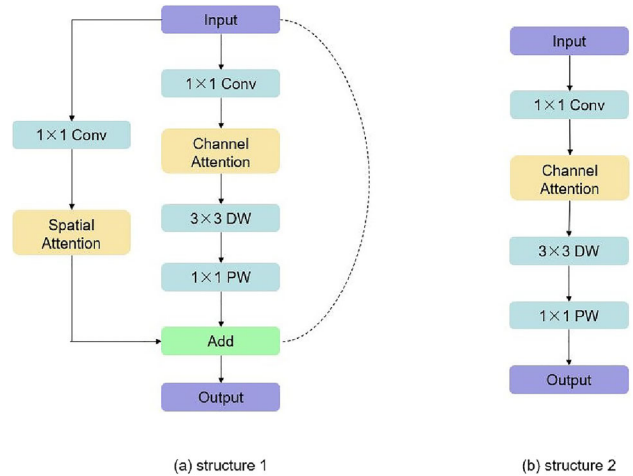
The Adam optimizer is widely adopted in various fields due to its superior performance in deep learning optimization. It offers the advantage of automatically adjusting parameter learning rates, which results in faster convergence and improved overall network stability. However, despite these benefits, Adam has several drawbacks [32–34]. For instance, Adam converges faster than other optimization algorithms and also tends to reduce the learning rate to a negligible value during the later stages of training, which may result in non-convergence and suboptimal performance. Moreover, compared to stochastic gradient descent methods, Adam may produce poorer results in some cases.

Therefore, we propose using the SGD algorithm as the optimization algorithm for our network, and supplement it with momentum to facilitate quicker convergence. In addition, to address the issue of slow convergence, we employ the lambda learning rate decay algorithm, which is characterized by the following decay equation:

$$lr = lambda(x) = \frac{1 + \cos(\frac{x \times \pi}{Epochs})}{2} \times (1 - lrf) + lrf. \tag{10}$$

where *lr* denotes the initial learning rate, and *lrf* represents the learning rate update parameter.

The lambda learning rate decay algorithm is designed to counteract slow convergence by allowing for a larger initial learning rate that speeds up early stage convergence. As the network continues to train, the learning rate gradually decreases, which prevents the network from converging pre-



**Fig. 3** Specific structure of the Deep Spindle module

maturely to a local optimal solution and instead yields better overall training results.

**Architecture of the DMBA Net**

In this section, we provide a detailed description of our architecture. As mentioned earlier, our fundamental building block is a spindle-shaped residual module based on depth-separable convolution. The structure of this module can be observed in Fig. 3. Table 2 provides an overview of the complete architecture of our model. We utilize ReLU6 as the nonlinear activation function due to its ability to maintain stability during low precision computations [22]. For modern networks, we typically employ a kernel size of 3 × 3 and include batch normalization [35] in the training process.

DMBANet leverages a multi-branch structure and attention mechanism to facilitate the efficient extraction of features from medical images. Initially, we optimize the residual block to give rise to the inverted residual structure. This architectural choice entails upscaling the intermediate con-

**Table 2** Detailed structure of the network

Input Size	Operator	Layer Details	Stride	Output Size
$224 \times 224 \times 3$	Conv2d	Kernel Size = $3 \times 3$	2	$112 \times 112 \times 64$
$112 \times 112 \times 64$	Deep Spindle	$\left\{ \begin{array}{l} 1 \times 1, 64 \\ \text{Spatial Attention} \end{array} \right\} + \left[ \begin{array}{l} 1 \times 1, 64 \\ \text{Channel Attention} \\ 3 \times 3, 256 \\ 1 \times 1, 64 \end{array} \right] \times 2$	1	$112 \times 112 \times 64$
$112 \times 112 \times 64$	Deep Spindle	$\left\{ \begin{array}{l} 1 \times 1, 128 \\ \text{Spatial Attention} \end{array} \right\} + \left[ \begin{array}{l} 1 \times 1, 128 \\ \text{Channel Attention} \\ 3 \times 3, 512 \\ 1 \times 1, 128 \end{array} \right] \times 3$	2	$56 \times 56 \times 128$
$56 \times 56 \times 128$	Deep Spindle	$\left\{ \begin{array}{l} 1 \times 1, 256 \\ \text{Spatial Attention} \end{array} \right\} + \left[ \begin{array}{l} 1 \times 1, 256 \\ \text{Channel Attention} \\ 3 \times 3, 1024 \\ 1 \times 1, 256 \end{array} \right] \times 5$	2	$28 \times 28 \times 256$
$28 \times 28 \times 256$	Deep Spindle	$\left\{ \begin{array}{l} 1 \times 1, 512 \\ \text{Spatial Attention} \end{array} \right\} + \left[ \begin{array}{l} 1 \times 1, 512 \\ \text{Channel Attention} \\ 3 \times 3, 2048 \\ 1 \times 1, 512 \end{array} \right] \times 3$	2	$14 \times 14 \times 512$
$14 \times 14 \times 512$	Adaptive AvgPool		–	$1 \times 1 \times 512$
$1 \times 1 \times 512$	Flatten + Fully Connected Layer + Softmax		–	$1 \times \text{NumClass}$

It is worth noting that the Deep Spindle layer consists of plural Deep Spindle blocks, and when the stride of the Deep Spindle layer is 2, it only means that the stride of the first Deep Spindle block in the layer is 2, which is used for downsampling the feature map, while the stride of the remaining Deep Spindle blocks is still 1

volutional layers, enhancing the network's performance by maximizing the extraction of feature information from the input feature map. Furthermore, the introduction of depth-separable convolution within the inverted residual structure achieves the decoupling of intermediate convolutional layers, enabling channel-by-channel convolution. This strategic move markedly diminishes network parameters and computations, leading to an improvement in the overall operation speed of the network. DMBANet adopts a multi-branch design, incorporating diverse attention mechanisms into distinct branches. The key branches include the inverted residual branch combined with the channel attention mechanism, significantly enhancing network performance as the primary branch of DMBANet. The spatial attention branch excels at capturing feature information overlooked by the primary branch, while the residual connections in another branch serve to minimize neural network degradation and expedite training speed.

## Experiments

### Implementation details

To implement our designed model, we leverage the PyTorch framework and conduct training on an RTX 3080Ti. The evaluation of our model's performance is carried out on the PathoIMG data set. Since this data set is not inherently par-

tioned into training and test sets, we manually perform the split, allocating 80% of the histopathological images to the training set and the remaining 20% to the test set. Specific details regarding the training configuration used in our experiments are provided in Table 3.

For comparison purposes, we selected several state-of-the-art models for testing in this study. These models include ResNet [19], ResNeXt [29], MobileNet [21, 37], ShuffleNet [22], EfficientNet [23, 46], Vision Transformer (ViT) [24], and ConvNeXt [25]. As shown in Table 3, all of these models are trained under the same experimental conditions, configuration settings, and hyperparameters. The only variation was that the batch size was dynamically chosen based on GPU memory utilization.

In evaluating the performance of our method, we utilized accuracy (Acc), which is one of the most widely adopted metrics in classification tasks. This metric provides a straightforward indication of a model's efficacy by calculating the ratio between the number of correctly classified samples and the total number of samples. The accuracy metric can be expressed as

$$Acc = \frac{S_{rig}}{S_a}, \quad (11)$$

where  $S_{rig}$  is the number of samples correctly classified in the test data and  $S_a$  is the total number of samples in the test data.



**Table 3** PathoIMG data set training settings

Training config	Our approach	Contrast model
Optimizer	SGDM	SGDM
Weight decay	4E-5	4E-5
Optimizer momentum	0.9	0.9
Learning rate schedule	Lambda Decay	Lambda Decay
Base learning rate	0.01	0.01
Learning rate update coefficient	0.1	0.1
Batch size	16	32-256
Training epochs	200	200
Learning rate warmup	None	None

**Table 4** Summary of the PathoIMG data set

Class	Numbers
Normal	299
Benign	1106
In situ carcinoma	1066
Invasive carcinoma	1300
Color model	RGB
Resolution size	2048 × 1536 pixels
Type of label	Imagewise

## Data set

The PathoIMG data set [5] comprises 3771 meticulously curated high-resolution histopathological images of breast cancer, meticulously stained with hematoxylin and eosin. Hematoxylin serves to accentuate the nucleus, while eosin highlights other vital structures. All images adhere to consistent acquisition conditions, with a magnification of either 10x or 20x. The histopathological sections are prepared using the standard paraffin filming method, a widely adopted practice in hospitals. Each image in the data set is meticulously labeled based on the type of tumor depicted in the histopathological section, encompassing categories such as normal, benign, in situ carcinoma, or invasive carcinoma. PathoIMG stands out as a classical data set in the realm of medical imaging. It encapsulates quintessential features inherent to medical imaging, including high image resolution, variations in tissue and cell morphology, instances of cell overlap, and color distribution heterogeneity. Notably, it boasts an extensive data set, making it an ideal benchmark for evaluating the efficacy of our models.

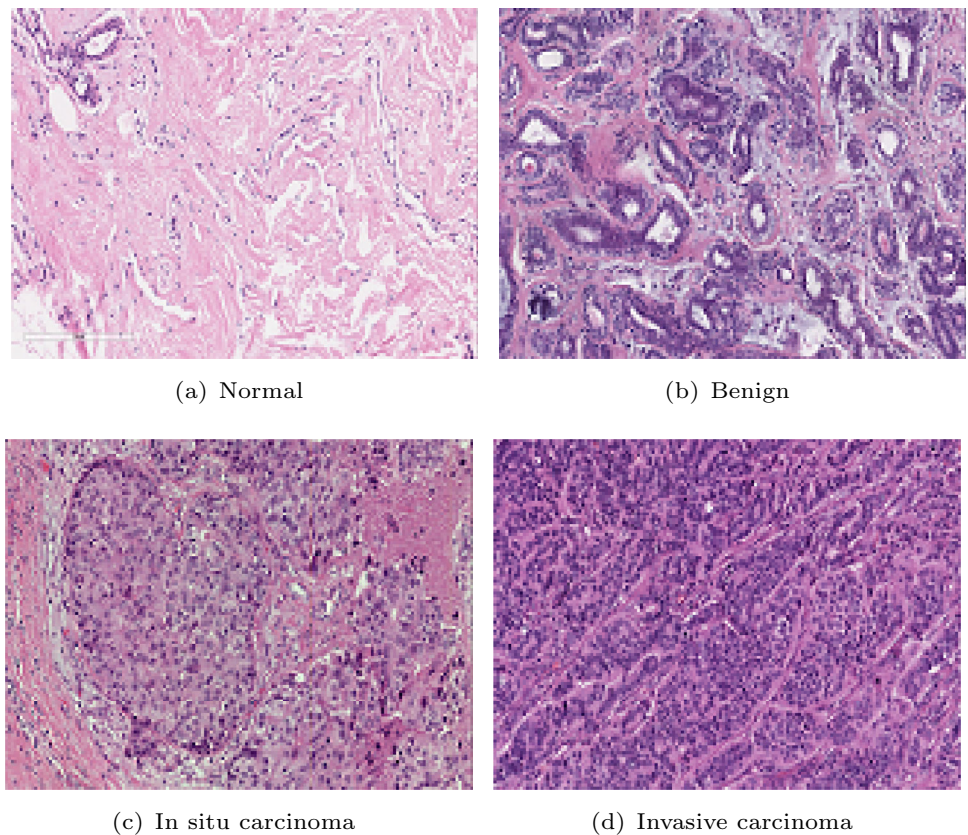
The histopathological images of breast cancer in the PathoIMG image data set were labeled by two experienced pathologists. Any images with objections were reviewed and confirmed by the head of the histopathological department. The quantitative distribution of the 3771 breast cancer images in the data set is presented in Table 4, along with a summary

description. The format of each histopathological image in the data set is also defined. The different classes of breast cancer images included in the data set can be seen in Fig. 4.

The raw histopathology images in the PathoIMG data set boast a resolution of up to  $2048 \times 1536$  pixels, rendering them unsuitable for direct utilization as inputs for a neural network. This limitation stems from two primary reasons. First, employing high-resolution images as inputs results in an escalation of parameters in the convolutional and fully connected layers during convolutional operations. This amplifies the computational resource consumption, reaching potentially impractical levels. Second, the constrained memory of GPUs necessitates a reduction in batch sizes when deploying high-resolution images, significantly impeding the network training speed. To circumvent these challenges, a viable solution involves slicing the original histopathological images into smaller patches with lower resolution, which can then serve as input. This strategy ensures the comprehensive extraction of feature information without overwhelming the network. Given that the raw data in the PathoIMG data set is labeled at the image level, we can conveniently crop the raw data. In this scenario, we opt to partition the original image into 12 patches, each patch with a resolution of  $512 \times 512$  pixels. This tactical approach retains the crucial features of the original image while downsizing it to a manageable scale suitable for neural network input.

It is crucial to note that the structure and function of the breast change significantly throughout a woman's life due to various factors such as puberty, sexual maturity, pregnancy, lactation, and old age. Therefore, to ensure the diversity of data and enable machine learning algorithms to learn a sufficient number of representative features, it is essential to include a wide range of histopathological images from patients of different ages. Doing so adequately reflects the morphology of the breast tissue at different stages of a woman's life. Therefore, PathoIMG data set strive to cover as many histopathological images of different patients and ages as possible to ensure diversity. By including images from different age groups, PathoIMG data set can better represent

**Fig. 4** Example of histopathological images of breast cancer from the PathoIMG data set



the morphological diversity of the breast tissue and improve the accuracy of the machine learning algorithm.

### Results under basic benchmark

In this work, we employ advanced techniques such as residual connections and deep separable convolution [18]. To evaluate the effectiveness of our model, we compare it with several state-of-the-art deep learning networks such as ResNet [19], MobileNet [21], ShuffleNet [22], EfficientNet [23], and ViT [24]. To ensure fairness in our evaluation, we consider the difference in parameters and computation cost for different networks. By running each network using the same memory limits, we can compare their performance under similar conditions. Thus, we perform the comparison by adjusting the batch size of each network (32–256) so that it consumes precisely the entire memory of a RTX 3080Ti. This approach enables us to provide a fair comparison of the various models in terms of their efficiency and accuracy.

The outcomes illustrated in Table 5 unequivocally establish the superior performance of our DMBANet model in comparison to other cutting-edge deep learning models, particularly in terms of accuracy. We attribute the efficacy of our model to two key factors: First, the introduction of the Deep Spindle Structure enables the up-dimensioning of the intermediate layer (using a  $3 \times 3$  depthwise convolutional ker-

nel), which was originally down-dimensioned. This unique structure facilitates the comprehensive extraction of feature information from histopathological images, mitigating the risk of accuracy loss due to information loss. Second, our approach involves employing diverse sizes of convolutional kernels, attention mechanisms, and varying branch depths across different branches of the Deep Spindle Structure. This strategy imparts the network with a rich array of multi-scale feature information, further enhancing accuracy. Furthermore, we present the confusion matrix for different models, specifically at 12 times the cutoff multiplication rate, in Fig. 5. The matrix distinctly indicates that our model not only excels in accuracy but also outperforms in additional metrics such as precision and recall. In summary, these results strongly affirm that our DMBANet model constitutes an effective and efficient solution for histopathological image classification.

After analyzing the confusion matrix results, a comparative assessment between DMBANet and EfficientNet unfolds. EfficientNet, acclaimed as the standout performer among the models under consideration, not only excels in terms of accuracy but also demonstrates remarkable efficiency in managing parameters and computational resources. The detailed performance outcomes, meticulously presented in the ensuing Table 6, unmistakably highlight DMBANet's superiority across all evaluated dimensions. In each facet of evaluation, ranging from accuracy to the Precision, Recall,

**Table 5** Compared with other state-of-the-art models on the basis of slicing the original histopathological images into 12 patches

Methods	Details	Batch size	Accuracy (%)
ResNet [19]	ResNet-50 [19]	96	96.1
	ResNet-101 [19]	64	96.5
ResNeXt [29]	ResNeXt-50, Groups = 32 Width Per Group = 4	80	97.3
	ResNeXt-101, Groups = 32 Width Per Group = 8	32	96.7
MobileNet-V2 [21]	$\alpha = 1.0$	96	96.7
MobileNet-V3 [37]	Small	256	95.0
	Large	128	96.8%
ShuffleNet-V2 [22]	Width Coefficient = 0.5	256	94.1
	Width Coefficient = 1.0	256	95.6
EfficientNet-V1 [23]	B0	96	97.1
EfficientNet-V2 [46]	Small	48	97.4
	Middle	32	97.2
	Base, Embed Dim = 768	48	93.6
ConvNeXt [25]	Tiny, Depths = [3, 3, 9, 3]	64	92.9
	Small, Depths = [3, 3, 27, 3]	48	92.7
	Base, Depths = [3, 3, 27, 3]	32	93.6
DMBANet	–	16	<b>98.1</b>

Specificity, F1-score and so on, DMBANet emerges as the superior choice. This comprehensive comparison solidifies the position of DMBANet as a formidable model, surpassing EfficientNet in all aspects and affirming its prowess in real-world applications.

Figure 6 visually illustrates the effectiveness of our method, which outperforms other methods not just in accuracy, but also features smoother performance without significant fluctuations. To enhance the aesthetics of the graph, only the last 100 epochs of the training process are presented.

The choice of a batch size of 16 for the DMBANet model is not solely based on the number of model parameters. Instead, it is influenced by several other factors related to the architecture and memory requirements of the network. One key consideration is the stride of the first deep spindle block, which is set to 1. This results in a larger overall feature map size for the model. Consequently, all subsequent deep spindle blocks operate on a larger feature map, leading to a significant increase in the memory footprint of the network. This larger memory requirement can impact the batch size selection. Moreover, the construction of the deep spindle structure, along with the increase in remaining connections, leads to longer GPU memory usage per convolution process. This prolonged memory usage can result in additional performance degradation. However, it is important to note that including these remaining connections remains beneficial, as observed in the results of RepVGG [41]. While a smaller batch size might alleviate the memory constraints, it is essential to find a balance that considers both memory lim-

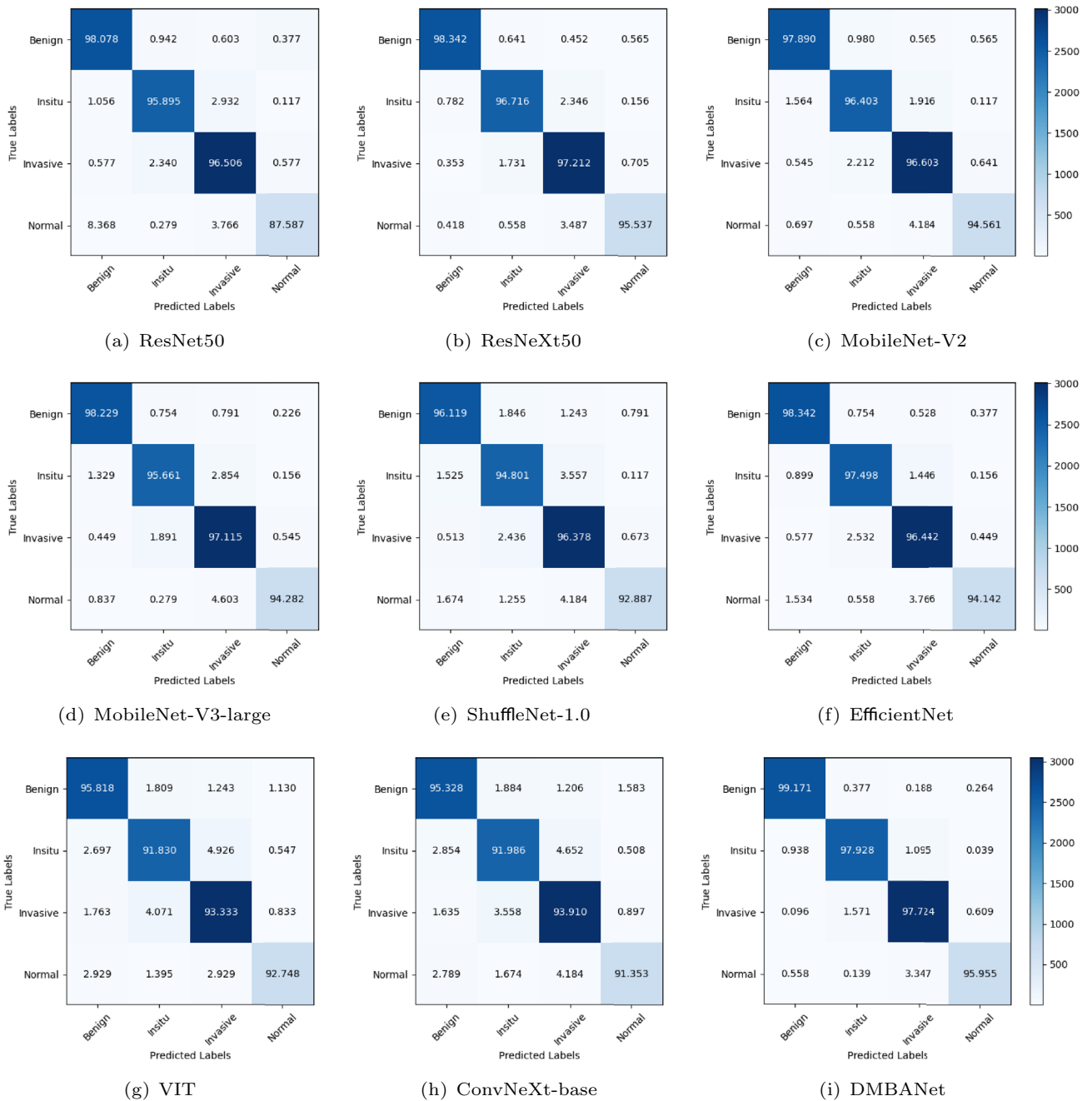
itations and model performance. Through experimentation and empirical analysis, a batch size of 16 was determined to be a suitable compromise that effectively manages memory consumption while still achieving favorable results with the DMBANet model.

### Results at other cutoff multipliers

To extract the complete feature information from the histopathological image, we divided the original image into 12 smaller patches. We conducted three comparison tests in this study to verify the reliability and validity of this process. These tests include: “Original”, which uses the original data, “2 patches”, which uses 2 patches of the original data, and “4 patches”, which uses 4 patches of the original data.

Table 7 shows that our proposed method is consistently more effective than the comparison models, regardless of whether the original image is sliced or how many patches it is divided into. This confirms the efficiency of our approach in extracting feature information. In addition, using the original data as network input resulted in the worst performance for all models. However, when we sliced the image, the performance of all models improved. Interestingly, the more patches the image is divided into, the better the performance. Our proposed method achieved the best results when the original image was sliced into 12 patches, as demonstrated in the previous section.

It is natural to question whether better performance can be achieved by slicing the original image into more patches.

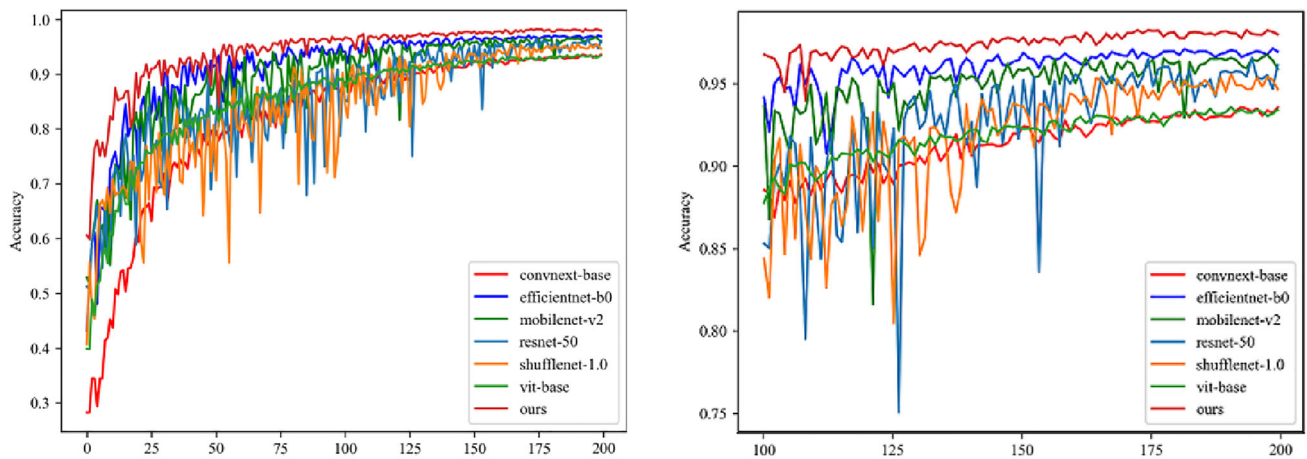


**Fig. 5** Confusion matrix for different models at 12× cutoff multiplier, we have selected the optimal version of each SOTA model based on their performance

**Table 6** Comparison of DMBANet and EfficientNet results on precision, recall, specificity, F1-Score, etc

	DMBANet					EfficientNet			
	Precision	Recall	Specificity	F1-Score		Precision	Recall	Specificity	F1-Score
Benign	0.988	0.992	0.995	0.990	Benign	0.980	0.983	0.992	0.981
Insitu	0.977	0.979	0.991	0.978	Insitu	0.960	0.975	0.984	0.967
Invasive	0.982	0.977	0.990	0.980	Invasive	0.975	0.964	0.987	0.969
Normal	0.962	0.960	0.997	0.961	Normal	0.960	0.941	0.997	0.950





**Fig. 6** For aesthetic reasons, each state-of-the-art model selects only one version for visualization. The left figure represents the entire training process, while the right figure displays the accuracy for the last one hundred epochs of training

**Table 7** Compared with other models on different slice magnification of the histopathological images

Methods	Details	Parameters	Flops	Original (%)	2 Patches (%)	4 Patches (%)
ResNet [19]	Layers = 50	23.516M	4.132G	92.8	94.8	96.0
	Layers = 101	42.508M	7.864G	91.4	93.8	95.3
ResNeXt [29]	ResNeXt-50, Groups = 32 Width Per Group = 4	22.988M	4.286G	94.0	95.0	96.2
	ResNeXt-101, Groups = 32 Width Per Group = 8	86.751M	16.537G	92.0	94.3	95.7
MobileNet-V2 [21]	$\alpha = 1.0$	2.229M	326.276M	93.9	95.9	96.0
MobileNet-V3 [37]	Small	1.522M	61.17M	90.6	92.0	94.0
	Large	4.207M	232.962M	93.5	94.1	96.0
ShuffleNet-V2 [22]	Width Coefficient = 0.5	345.862K	43.552M	89.5	91.6	92.4
	Width Coefficient = 1.0	1.258M	151.689M	90.7	93.8	94.4
EfficientNet-V1 [23]	B0	4.013M	411.555M	94.3	96.1	96.5
EfficientNet-V2 [46]	Small	20.183M	2.897G	95.4	96.1	96.9
	Middle	52.863M	5.439G	92.3	94.8	95.5
ViT [24]	Base, Embed Dim = 768	85.65M	16.863G	85.9	86.3	91.6
ConvNeXt [25]	Tiny, Depths = [3, 3, 9, 3]	27.802M	4.455G	71.9	79.3	85.9
	Small, Depths = [3, 3, 27, 3]	49.416M	8.683G	72.9	82.4	87.6
	Base, Depths = [3, 3, 27, 3]	87.514M	15.354G	74.8	86.86	90.0
DMBANet	–	10.559M	6.065G	<b>96.8</b>	<b>97.0</b>	<b>97.9</b>

Original(%) indicates that the original image is directly used as the network input, and X-patches represents the number of patches of the histopathological image after slicing

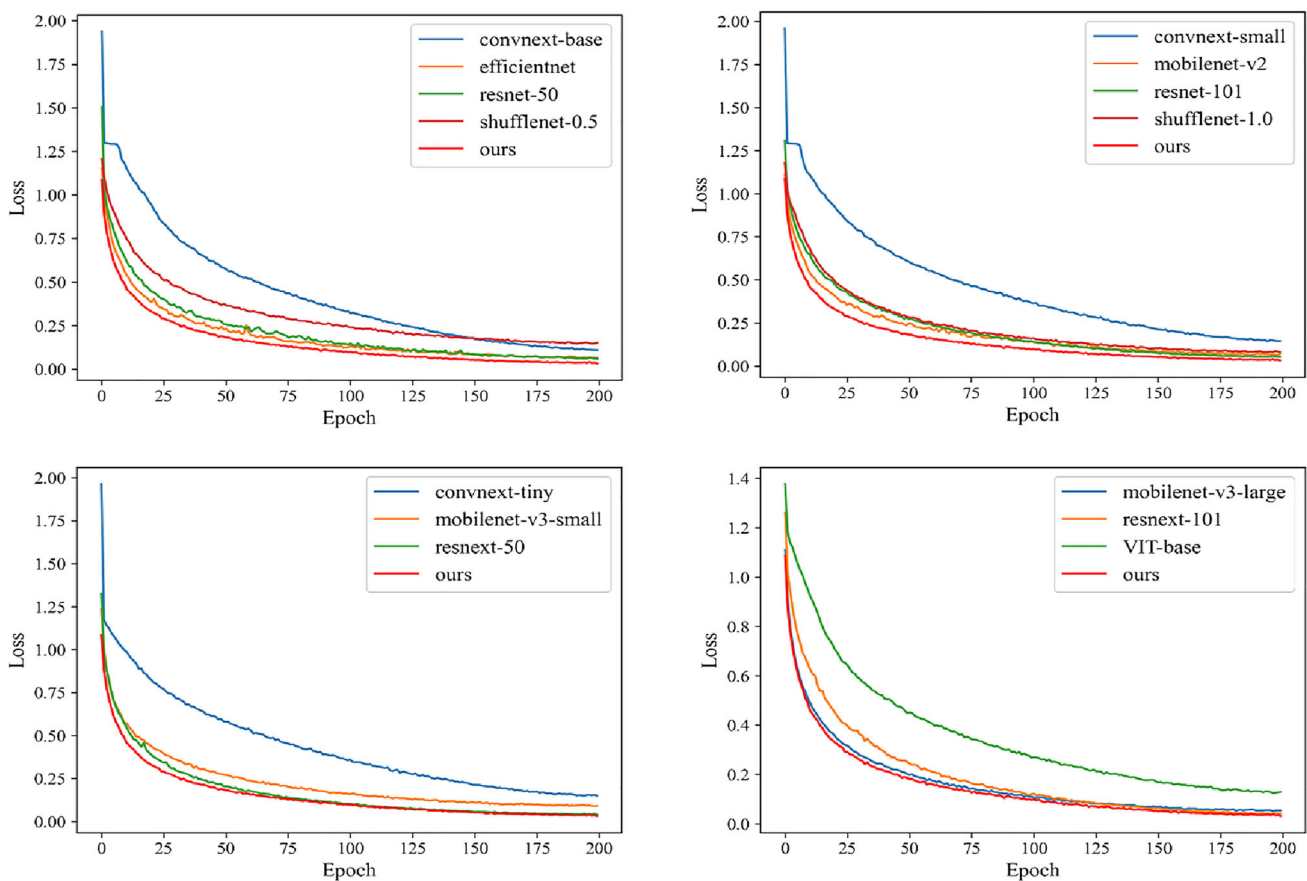
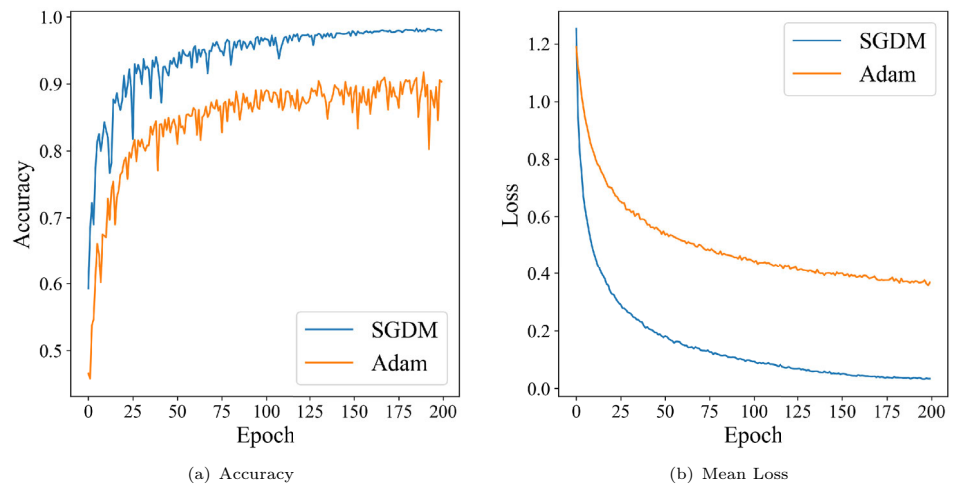
However, our experiments show that this is not the case. Slicing the original image into 48 patches ( $256 \times 256$ ) resulted in worse performance than using 12 patches. The reason for this is that the simple slicing operation leads to blank areas in some patches, which impacts the feature extraction process. We believe that conducting pathological screening on the sliced patches could improve the results, but would increase manual labor costs. Moreover, due to the large amount of data generated using 48 patches, running experiments takes significantly longer (around four to five days), ultimately

limiting computational resources. Therefore, we conducted only a few experiments using our proposed method, which resulted in worse performance than with 12 patches.

### The effect between Adam and SGDM

To overcome the potential drawback of the Adam optimization algorithm not converging, we used the SGDM optimization algorithm in this study. We also employed the lambda learning rate decay strategy to accelerate network



**Fig. 7** Difference between SGDM and Adam's results**Fig. 8** Convergence analysis with other methods

convergence and address the slow convergence of the SGDM optimization algorithm. Figure 7 presents the results of using the Adam optimization algorithm and SGDM optimization algorithm with lambda learning rate decay.

The results in Fig. 7 demonstrate that using the SGDM optimization algorithm with the lambda learning rate decay algorithm produces better performance for the network

model. First, using SGDM optimizer outperforms Adam optimizer in terms of classification accuracy throughout the training process, as shown in Fig. 7a. SGDM optimizer has higher and more stable classification accuracy compared to Adam optimizer which has lower accuracy and keeps fluctuating until late stages of the training where large fluctuations make it difficult to obtain a stable result. Second, as expected,

the results in Fig. 7b show that Adam optimization algorithm is difficult to converge, and mean loss remains very high even after 200 epochs of training. In contrast, the SGDM optimizer with lambda learning rate decay algorithm successfully converges the network after the same training run, and achieves lower mean loss.

## Convergence analysis

In Fig. 8, we performed convergence analysis to illustrate the mean loss of our network compared to others. It is evident that our DMBANet has the best convergence result and achieves the highest accuracy at any point during the entire training process. Our approach maintains high accuracy while also achieving comparable convergence speeds with lightweight networks. We set the number of epochs to 200 to maximize the performance of transformer models such as ViT [24] and ConvNext [25]. However, if the number of epochs is reduced to 100 or even lower, as is commonly done in previous works for convolutional neural networks, our method can achieve an even greater lead compared to other models.

The exceptional convergence speed of the proposed DMBANet suggests that it can effectively extract information from the data to a greater extent. Moreover, Exceedingly fast convergence confirms that the methods we proposed, including the deep spindle structure and multi-branch attention, are effective at improving the network's ability to extract pathological information from perspectives other than accuracy.

## Ablation study

Table 8 demonstrates that experimental results are significantly improved by optimizing the bottleneck structure to the Deep Spindle structure while maintaining the same residual connections. This improvement confirms that the bottleneck structure can lead to a loss of image information due to its initial dimensionality reduction. We incorporate dual channels with lighter channel attention and spatial attention in the network to focus on different types of information. The use of these attention mechanisms significantly improves the classification accuracy of the network.

It needs to be noted that we do not conduct experiments to remove the depthwise separable convolution in DMBANet, because the resulting explosive increase in the number of parameters would make it difficult to train the network on a single 3080Ti GPU. However, as shown in MobileNet, the use of depthwise separable convolution can significantly reduce the number of parameters in the network. Nonetheless, this reduction can also have an impact on network accuracy.

**Table 8** For runtime reasons, the ablation study was done on the basis of slicing the histopathological images to a resolution of  $1024 \times 768$  pixels

Methods	Deep Spindle	CA and SA	Accuracy (%)
ResNet-50	×	×	96.0
DMBANet	✓	×	97.0
DMBANet	✓	✓	97.9

## Discussion

It is essential to note that the current study has some restrictions. First, the images in the PathoIMG data set have been cropped and selected by a pathologist, which can help to extract pathological information better. However, the model has not yet been evaluated on the original entire slide data set. Second, the size of the PathoIMG data set is relatively small, with only 3771 histopathological images in total. Therefore, the size of both the training set and test set is small, which may introduce bias in the results. Third, the model has been evaluated on only one data set in this paper. In future work, we plan to extend our model to more histopathological data sets to test its generalization ability. In conclusion, we aim to evaluate our method on a larger histopathology image data set to further improve the robustness of our model.

## Conclusion

In summary, this paper presents a novel method for breast cancer histopathological image classification using deep spindle structure and multi-branch attention. The deep spindle structure incorporated in the network addresses the issue of losing feature information due to the large Whole Slide Image resolution. This enables the network to efficiently extract feature information from the histopathological images, resulting in more accurate classification results. Incorporating channel attention and spatial attention in different branches further enhances the network's ability to notice richer feature information, improving its overall performance.

We test DMBANet on the PathoIMG data set, and compare it not only with popular Convolutional Neural Networks like ResNet, MobileNet, and EfficientNet, but also with Transformer-based networks such as ViT and ConvNeXt. By benchmarking against these high-performance networks, we demonstrate that our approach outperforms the state-of-the-art classification methods for breast cancer histopathological images, implying the potential for our methodology to achieve higher accuracy and efficiency in clinical practice.

**Acknowledgements** This work was supported in part by the National Key Research and Development Program of China (2021YFE0102100),

in part by National Natural Science Foundation of China (62303014, 62172002, 62302007), in part by the University Synergy Innovation Program of Anhui Province (GXXT-2022-035), in part by Anhui Provincial Natural Science Foundation (2108085QF267, 2308085QF225) and in part by Education Department of Anhui Province (2023AH050061).

**Data availability** The data generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

1. Esteva A, Chou K, Yeung S et al (2021) Deep learning-enabled medical computer vision. *NPJ Digit Med* 4(1):5
2. Aggarwal R, Sounderajah V, Martin G et al (2021) Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med* 4(1):65
3. Spanhol FA, Oliveira LS, Petitjean C, Heutte L (2015) A dataset for breast cancer histopathological image classification. *Proc IEEE Trans Biomed Eng* 63(7):1455–1462
4. Wei B, Han Z, He X, Yin Y (2017) Deep learning model based breast cancer histopathological image classification. In: *Proc IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp 348–353
5. Yan R, Ren F, Wang Z, Wang L, Zhang T, Liu Y, Rao X, Zheng C, Zhang F (2020) Breast cancer histopathological image classification using a hybrid deep neural network. *Methods* 173:52–60
6. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception Architecture for Computer Vision. In: *Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 2818–2826
7. Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks. In: *Proc IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 7132–7141
8. Xu B, Liu J, Hou X, Liu B, Garibaldi J et al (2020) Attention by selection: a deep selective attention approach to breast cancer classification. *Proc IEEE Trans Med Imaging* 39(6):1930–1941
9. Zhang J, Xie Y, Xia Y, Shen C (2019) Attention residual learning for skin lesion classification. *Proc IEEE Trans Med Imaging* 38(9):2092–2103
10. Zhu W, Sun L, Huang J, Han L, Zhang D (2021) Dual attention multi-instance deep learning for Alzheimer's disease diagnosis with structural MRI. *Proc IEEE Trans Med Imaging* 40(9):2354–2366
11. Gu R et al (2021) CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation. *Proc IEEE Trans Med Imaging* 40(2):699–711
12. Krizhevsky A, Sutskever I, Hinton G (2017) Imagenet classification with deep convolutional neural networks. *Commun ACM* 60(6):84–90
13. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: *Proc IEEE Conference on Computer Vision and Pattern Recognition*, pp 248–255
14. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely Connected Convolutional Networks. In: *Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 2261–2269
15. Lin T-Y, Dollár P et al (2017) Feature Pyramid Networks for Object Detection. In: *Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 936–944
16. Tan D et al (2023) Large-scale data-driven optimization in deep modeling with an intelligent decision-making mechanism. In: *Proc IEEE Transactions on Cybernetics*
17. Tan D et al (2023) Deep adaptive fuzzy clustering for evolutionary unsupervised representation learning. In: *Proc IEEE Transactions on Neural Networks and Learning Systems*
18. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: *Proc IEEE conference on computer vision and pattern recognition*
19. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proc IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 770–778
20. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint arXiv:1704.04861*
21. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C (2018) MobileNetV2: Inverted residuals and linear bottlenecks. In: *Proc IEEE Conference on Computer Vision and Pattern Recognition*, pp 4510–4520
22. Ma N, Zhang X, Zheng H, Sun J (2018) Shufflenet v2: Practical guidelines for efficient CNN architecture design. In: *Proc European conference on computer vision (ECCV)*, pp 116–131
23. Tan M, Le Q (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In: *Proc International Conference on Machine Learning*, pp 6105–6114
24. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X et al (2021) An image is worth 16 x 16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations*
25. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S (2022) A ConvNet for the 2020s. In: *Proc IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 11966–11976
26. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin Transformer: hierarchical vision transformer using shifted windows. In: *Proc IEEE Conference on Computer Vision and Pattern Recognition*, pp 10012–10022
27. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: *Proc IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June
28. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: *Proc. 30th Int. Adv. Neural Inf. Process. Syst*
29. Xie S et al (2017) Aggregated Residual Transformations for Deep Neural Networks. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 5987–5995
30. Han D, Kim J, Kim J (2017) Deep pyramidal residual networks. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 6307–6315
31. Woo S, Park J, Lee J-Y, Kweon IS (2018) CBAM: Convolutional block attention module. In: *Proc. European Conference on Computer Vision (ECCV)*, pp 3–19

32. Reddi SJ, Kale S, Kumar S (2019) On the Convergence of Adam and Beyond. arXiv preprint [arXiv:1904.09237](https://arxiv.org/abs/1904.09237)
33. Wilson AC, Roelofs R, Stern M, Srebro N, Recht B (2017) The Marginal Value of Adaptive Gradient Methods in Machine Learning. arXiv preprint [arXiv:1705.08292](https://arxiv.org/abs/1705.08292)
34. Keskar NS, Socher R (2017) Improving Generalization Performance by Switching from Adam to SGD. arXiv preprint [arXiv:1712.07628](https://arxiv.org/abs/1712.07628)
35. Loffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR
36. Liu Q, Li D, Ge SS, Ouyang Z (2021) Adaptive feedforward neural network control with an optimized hidden node distribution. *Proc IEEE Trans Artif Intell* 2(1):71–82
37. Howard A et al (2019) Searching for MobileNetV3. In: *Proc IEEE/CVF International Conference on Computer Vision (ICCV)*, pp 1314–1324
38. Zhang X, Zhou X, Lin M, Sun J (2018) ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: *Proc IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp 6848–6856
39. Guo M, Xu T, Liu J et al (2022) Attention mechanisms in computer vision: a survey. *Comput Vis Media* 8(3):331–68
40. Yang J, Zheng W-S, Yang Q, Chen Y-C, Tian Q (2020) Spatial-temporal graph convolutional network for video-based person re-identification. In: *Proc IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp 3289–3299
41. Ding X et al (2021) RepVGG: Making VGG-style ConvNets Great Again. In: *Proc IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 13728–13737
42. Hou L, Samaras D et al (2016) Patch-based convolutional neural network for whole slide tissue image classification. In: *Proc IEEE conference on computer vision and pattern recognition*
43. Wei JW, Tafe LJ et al (2019) Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Sci Rep* 9(1):3358
44. Vente CD et al (2022) Automated COVID-19 grading with convolutional neural networks in computed tomography scans: a systematic comparison. In: *Proc IEEE Transactions on Artificial Intelligence*, vol. 3, no. 2, pp 129–138
45. Pan B, Li C, Che H, Leung M-F, Yu K (2023) Low-rank tensor regularized graph fuzzy learning for multi-view data processing. *Proc IEEE Trans Consum Electron*. <https://doi.org/10.1109/TCE.2023.3301067>
46. Tan MX, Le Q (2021) Efficientnetv2: smaller models and faster training. In: *International conference on machine learning*. PMLR

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.