Breast cancer histopathological image classification using a hybrid deep neural network

Rui Yan\textsuperscript{a,b}, Fei Ren\textsuperscript{b}, Zihao Wang\textsuperscript{b,c}, Lihua Wang\textsuperscript{d}, Tong Zhang\textsuperscript{d}, Yudong Liu\textsuperscript{b}, Xiaosong Rao\textsuperscript{d,e}, Chunhou Zheng\textsuperscript{a,*}, Fa Zhang\textsuperscript{b,*}

\textsuperscript{a} College of Computer Science and Technology, Anhui University, Hefei, China
\textsuperscript{b} High Performance Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
\textsuperscript{c} University of Chinese Academy of Sciences, Beijing, China
\textsuperscript{d} Department of Pathology, Peking University International Hospital, Beijing, China

\textbf{ARTICLE INFO}

Keywords: Breast cancer Histopathological images Image classification Deep neural network Dataset

\textbf{ABSTRACT}

Even with the rapid advances in medical sciences, histopathological diagnosis is still considered the gold standard in diagnosing cancer. However, the complexity of histopathological images and the dramatic increase in workload make this task time consuming, and the results may be subject to pathologist subjectivity. Therefore, the development of automatic and precise histopathological image analysis methods is essential for the field. In this paper, we propose a new hybrid convolutional and recurrent deep neural network for breast cancer histopathological image classification. Based on the richer multilevel feature representation of the histopathological image patches, our method integrates the advantages of convolutional and recurrent neural networks, and the short-term and long-term spatial correlations between patches are preserved. The experimental results show that our method outperforms the state-of-the-art method with an obtained average accuracy of 91.3\% for the 4-class classification task. We also release a dataset with 3771 breast cancer histopathological images to the scientific community that is now publicly available at http://ear.ict.ac.cn/?page_id=1616. Our dataset is not only the largest publicly released dataset for breast cancer histopathological image classification, but it covers as many different subclasses spanning different age groups as possible, thus providing enough data diversity to alleviate the problem of relatively low classification accuracy of benign images.

1. Introduction

Cancer is a critical public health problem worldwide. Among the cancer types, breast cancer incidence rates are second highest for women, excluding lung cancer. In addition, the mortality of breast cancer is very high when compared to other types of cancer [1]. Even with the rapid advances in medicine, the analysis of histopathological images remains the most widely used method for breast cancer diagnosis [2]. In all the histopathological image analysis tasks, the most important is the classification task. Because the automatic and precise classification of high-resolution histopathological images is the cornerstone and bottleneck of other in-depth studies, such as nuclei localization, mitosis detection and gland segmentation.

Currently, histopathological imaging in clinical practice is mainly based on the manual qualitative analysis of pathologists. However, at least three problems arise from this analysis method. First, there is a shortage of pathologists in the world, especially in less developed areas and small hospitals. This resource shortage and unbalanced distribution is an urgent problem to be solved. Second, whether the histopathological diagnosis is correct or not correct completely depends on the pathologist’s profound professional knowledge and long-term accumulated diagnostic experience. This pathologist subjectivity has led to a proliferation of diagnostic inconsistencies. Third, the complexity of the histopathological images makes pathologists prone to fatigue and inattention. Facing these problems, it is urgent to develop automatic and precise histopathological image analytical methods, especially classification methods, to alleviate these problems.

Recently, deep learning methods have made considerable progress and achieved remarkable performance in the field of computer vision and image processing, which has inspired many scholars to apply this technique to histopathological image classification [3]. Convolutional neural networks (CNNs) are the most widely used type of deep learning network, and they perform equally well on image classification and image feature extraction [4]. These results have laid the foundation for...
the application of the CNN in histopathological image classification. However, in contrast to natural images, histopathological images are characterized by high resolution. Limited by the memory of the graphics processing unit (GPU), it is impossible to feed high-resolution images as a whole into the CNN for classification. Additionally, it is impractical to simply resize a high-resolution image into a low-resolution image because considerable useful image information will be lost. This is especially true in medical imaging, where data volumes are already small.

The mainstream method of histopathological image classification divides a whole image into smaller patches, then uses a CNN to classify each patch, and finally integrates the classification results of these patches, such as majority voting, to determine the classification results. A CNN is also used to extract the feature representation vector of each patch, and then the traditional machine learning classification algorithm, such as support vector machine (SVM), is used to make the classification result of the whole histopathological image [5].

However, the current mainstream method faces three challenges. First, the high-resolution characteristics of histopathological images have not been fully utilized to improve the classification accuracy but have caused great negative effects. The main reason is that the current best patch-based method does not adequately integrate these patches to make the classification result of the whole histopathological image. Specifically, these methods integrate only the short-distance dependency between patches but ignore the long-distance spatial dependency, which is very helpful for the context understanding of the whole image. Second, the feature representation of the pathological image patch is not sufficiently richer. Thus, a large amount of information is lost before image-wise fusion, making the fusion insufficient. Finally, there are considerable challenges in terms of datasets. Many very important advances in computer vision fields have benefited from an open research environment enabled by publicly available datasets for benchmarking, such as ImageNet, for object recognition in natural images. Medical imaging researchers have eventually started to follow this lead with the release of well-annotated datasets such as the BreaKHis dataset [6] and Bioimaging2015 dataset [7]. However, these datasets are still relatively small. In particular, the diversity of the dataset is not guaranteed.

In this paper, we propose a method that extracts richer multilevel features and integrates the advantages of the CNN and recurrent neural network (RNN), thus, the short-term and long-term spatial correlations between patches are preserved. We first split the high-resolution pathology images into small patches. Then, the CNN is used to extract the richer multilevel image features of each patch. Finally, the RNN is used to fuse the patch features to make the final image classification. For the 4-class classification task, we obtained an average accuracy of 91.3%, which outperforms the state-of-the-art method.

Additionally, cooperating with Peking University International Hospital, we released a dataset with 3771 breast cancer histopathological images, which led to an order of magnitude increase in the current dataset volume. Experimental results show that the average sensitivity for normal, benign, in situ carcinoma and invasive carcinoma improved by 2.9%, 16.4%, 7.8% and 2.3%, respectively, compared with the results on the Bioimaging2015 dataset using the same method. It is especially worth emphasizing that, due to our dataset covering as many different subsets spanning different age groups as possible to ensure sufficient data diversity, the classification sensitivity of benign images improved significantly from 68.7% to 85.1%. This increase indicates that both a high-performance deep learning algorithm and a sufficiently large and diverse dataset are essential to improve the ability of histopathological image classification.

2. Related work

Although many studies have been conducted and a series of important advances have been made in the automatic classification of breast cancer histoplastic images, the characteristics of histopathological images, such as the inconsistency of tissue and cell morphology, the phenomenon of cell overlapping, the appearance variability of stained histological sections and the uneven color distribution, have created considerable difficulties in image classification [8]. These problems result in considerable challenges for automatic and precise classification of breast cancer pathological images. It should also be noted that the resolution of pathological images is very high, which makes it impossible to directly transplant some methods that are successful in the field of natural images to the field of pathological images.

Early research methods for breast cancer pathological image classification mainly focused on the 2-class classification of cancer and noncancer [9–14] or a more complex 3-class classification of normal, in situ carcinoma and invasive carcinoma [15,16]. Most of the works were carried out on the entire image or extracted nuclei using textural, morphological and architectural features based on the traditional machine learning method. It is worth noting that most of the above classification approaches were carried out on low-resolution images at different magnifications. In addition, these approaches used artificial-based feature extraction methods, which require not only considerable effort and professional domain knowledge but also have certain difficulties in extracting distinguishing high-quality features that seriously restrict the application of traditional machine learning methods in the classification of breast cancer histopathological images.

Later, deep learning methods [17] achieved remarkable results in a wide array of computer vision tasks. The most important deep learning methods are the CNN and the RNN. CNNs have been widely used in the classification of pathological images. Spanhol et al. [6] released a breast cancer pathological image dataset named BreaKHis. Based on the dataset, they used the AlexNet network and used different integration strategies for classification, with a classification accuracy of 6% higher than traditional machine learning methods. Bayramoglu et al. [18] also used the magnification-independent deep learning method on the BreaKHis dataset, with a classification accuracy of approximately 83%. Araújo et al. [19] first considered 4-class classifications for breast cancer pathological images. They first extracted features based on a CNN similar to AlexNet and then used SVM to classify the extracted features. In contrast, RNNs are rarely used in pathological image classification tasks. Unlike the CNN, the RNN can use its internal state to process input data, and this characteristic ensures that the RNN has long-distance memory.

Recently, several excellent CNN-based methods for automatic and precise classification of breast cancer pathological images were developed for the ICIAR2018 challenge [20]. These methods have significantly advanced the state-of-the-art. The core ideas of these methods are much the same. The high-resolution histopathological images are first preprocessed and data-enhanced and then divided into equal-sized patches, and each patch is classified or the features extracted by a CNN. An image-wise classification is then made based on the vote of patch-wise classification results or fusion of extracted features. Vescal et al. [21] proposed a transfer learning method. Based on the pretraining model of GoogLeNet and ResNet, they first classified each patch of one image and then used the majority voting method to obtain image-wise classification results. Vang et al. [22] first used Google Inception-V3 to perform patch-wise classification. The patch-wise predictions were then passed through an ensemble fusion framework involving majority voting, a gradient boosting machine and logistic regression to obtain an image-wise prediction. Rakhlın et al. [23] proposed a different method named deep convolutional feature representation. In this method, pathological images are first encoded with the general convolutional neural network to obtain sparse descriptors of low dimensionality (1408 or 2048). Finally, they use gradient boosted trees for the final classification result. Awan et al. [24] utilized ResNet to obtain twelve 8192-dimensional feature vectors that represented twelve nonoverlapping patches of 512 × 512 pixels from one input image. To train a classifier with a larger context, they then trained an
SVM classifier with the bound features of $2 \times 2$ overlapping blocks of patches, which is equivalent to training the classifier with the features of $1024 \times 1024$ pixels in size. The majority voting result on the classification of $1024 \times 1024$ pixel overlapping blocks of patches was used as the final image-wise classification result.

To summarize, the development of the pathological image classification method based on deep learning in chronological order is as follows: 1, CNN + majority voting; 2, CNN + SVM; 3, CNN + transfer learning + majority voting or SVM; 4, CNN + transfer learning + patch-wise binding + majority voting or SVM. SVM can also be replaced with other traditional methods of machine learning. Although recently proposed methods are starting to focus on the result of patch-wise fusion to obtain the final image-wise classification results, these methods are either directly using majority voting and SVM or simply integrating short-distance patch dependencies. They all ignore the important role of long-distance spatial dependence of the histopathological image. Moreover, most of these recently proposed methods only average pool the last convolutional layer of the CNN into a one-dimensional feature vector and use it as the feature representation of image patches. This feature representation is not sufficiently rich for patch-wise fusion.

3. Dataset

One main characteristic of the deep learning method is that it can learn from large amounts of training data. Breakthrough results in the computer vision field were obtained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [25] based on the ImageNet dataset. In contrast, there are few publicly available large-scale image datasets in the medical image domain. Additionally, most of these datasets are not labeled. The cost for scarce medical experts to label data is very expensive. Moreover, traditional methods of annotating natural images, such as crowdsourcing [26], cannot be transplanted to the medical image domain because these tasks are very complex and often require long-term professional training and extensive domain knowledge. Thus, most of the early research on breast cancer pathological image analysis is performed on a small dataset, and other large datasets are usually not publicly available. Veta et al. [2] noted that the main obstacle in developing new analysis methods for pathological images is the lack of large, labeled and open datasets.

The open challenge of the medical image field has greatly contributed to the development of medical image analysis. Since 2007, medical imaging conferences and workshops, such as the International Conference on Image Analysis and Recognition (ICIAR), the International Symposium on Biomedical Imaging (ISBI), the International Conference on Pattern Recognition (ICPR) and Medical Image Computing and Computer-Assisted Intervention (MICCAI), have published a large number of medical image datasets for benchmark research, available at http://www.grand-challenge.org. The main advantage of these public benchmark datasets is that they provide a precise definition of tasks and assessment metrics to facilitate a fair comparison of the performance of various methods. Related to breast cancer pathological image classification, one of the largest open datasets containing 249 images was released by “Bioimaging2015: 4th International Symposium in Applied Bioimaging”. The goal of this challenge was to provide an automatic and precise classification for each input breast cancer pathological image. The Grand Challenge on Breast Cancer Histology images (BACH) [20] was organized as part of the ICIAR 2018 conference (15th International Conference on Image Analysis and Recognition). The organizer of the BACH challenge provided 400 pathological images that were consistent with the format of the Bioimaging2015 dataset. The pathological images were divided into 4 categories, each with 100 images. To the best of our knowledge, this is by far the largest dataset of breast cancer pathological images but is available only during the challenge for the competition. Although these open datasets have played a very significant role in improving the classification accuracy of breast cancer pathological images, the dataset sizes of 249 and 400 images are still too small compared with the open datasets of natural images.

To further promote and complement the research on the breast cancer pathological image classification field, we cooperated with Peking University International Hospital to release a new pathological image dataset of breast cancer. The format of our increased breast cancer pathological image dataset is completely consistent with the dataset published by Bioimaging2015. The institutional review board approved the study, and all the released data are anonymous. All images were acquired from March 2015 to March 2018 using a Leica Aperio AT2 slide scanner, and all patients were from China.

Our image dataset consists of 3771 high-resolution ($2048 \times 1536$ pixels) and annotated hematoxylin and eosin (H&E) stained breast pathological images. Hematoxylin highlights nuclei by staining DNA and eosin highlights other structures by staining proteins. All images have the same acquisition conditions: 100x or 200x magnification. The preparation procedure for pathological sections used in this work was the standard paraffin process, which is widely used in the pathological routine. According to the cancer type in each image, each image is labeled as normal, benign, in situ carcinoma or invasive carcinoma. The annotation was performed by two medical experts, and the images where there was disagreement were taken to the chief of pathology for final confirmation. Table 1 summarizes the image distribution. The initial image is the Bioimaging2015 dataset, and the extended image is our enhanced dataset, which can be regarded as an extension of the dataset in the article. Table 2 describes the image format in our dataset. Overall, based on the initial 249 images, we increased the number of images in the dataset to 4020.

In particular, the structure and function of breasts vary with age, such as puberty, sexual maturity, pregnancy, lactation and old age. To ensure the diversity of data to improve the learning ability of the machine learning algorithm, our dataset covers as many different subsets spanning different age groups as possible, which can fully reflect the morphology of breast tissue.

4. Methods

When a pathological image with high resolution ($2048 \times 1536$ pixels) is input, our goal is to accurately classify the image into one of four categories: normal, benign, in situ carcinoma and invasive carcinoma, as shown in Fig. 1. To this end, we propose a new hybrid convolutional and recurrent deep learning method, and the general workflow of our method is as follows (Fig. 2).

In the training stage, the pathological images are preprocessed and enhanced to improve quantitative analysis. After preprocessing, we first fine-tune the pretrained Inception-V3 [27] model. For each image, the trained patch-wise model is used to extract the feature representation

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Summary of our dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Normal</td>
</tr>
<tr>
<td>Initial</td>
<td>55</td>
</tr>
<tr>
<td>Extended</td>
<td>299</td>
</tr>
<tr>
<td>Overall</td>
<td>354</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Description of pathological images in our dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color model</td>
<td>R(ed)G(reen)B(lue)</td>
</tr>
<tr>
<td>Size</td>
<td>2048 × 1536 pixels</td>
</tr>
<tr>
<td>Memory space</td>
<td>3–20 MB (approx.)</td>
</tr>
<tr>
<td>Type of label</td>
<td>image-wise</td>
</tr>
</tbody>
</table>
vectors of 12 patches. Then, these 12 vectors are used as input to train image-wise long short-term memory (LSTM) [28]. In the testing stage, one pathological image is divided into an average of 12 small patches. Then, a fine-tuned Inception-V3 is used to extract the patch-wise image features. Each patch is extracted to a feature vector of $1 \times 5376$ dimensions. That is, 12 feature vectors can be extracted from one pathological image. Finally, the 12 feature vectors ($12 \times 1 \times 5376$) are input into a bidirectional LSTM to fuse the features of the 12 small patches to make the final complete image-wise classification. Since our method integrates the advantages of CNN and RNN, the short-term and long-term spatial correlations between patches can be preserved. We cover the method in more detail in the following sections.

4.1. Image preprocessing and augmentation

To alleviate many of the known inconsistencies in the staining process, thereby bringing sections that were processed under different conditions into a normalized space to enable better analysis, we used the method described in this paper [29] to normalize the pathological images of H&E staining. In our study, we perform 50 random color augmentations for each image.

4.2. Patch-wise method

Transfer learning [30] is an approach that applies CNNs that are pretrained on large annotated image databases, such as ImageNet, from different domains for various tasks. With such an approach, the original network architecture is retained, and the pretrained weights are used to initialize the network. The initialized weights are constantly updated during the following fine-tuning stage, enabling the network to learn features specific to the target task. Currently, a large number of studies have demonstrated that fine-tuning is efficient for a variety of classification tasks in the medical domain. For example, a recent study showed that utilizing pretrained Google’s Inception-V3 network on ImageNet and fine-tuned using images of skin lesions achieved very high accuracy for classification of skin cancer, comparable to that of numerous dermatologists [31].

In this paper, we use a well-known CNN architecture, named Google’s Inception-V3, which is pretrained to 93.33% top-five accuracy on the 1000 object classes (1.28 million images) of the 2014 ImageNet Challenge. Then, we fine-tune Google’s Inception-V3 to learn domain- and modality-specific features for classifying breast pathological images. Such a fine-tuning approach is easier to optimize and enables the training of deeper networks, which correspondingly leads to an overall improvement in network capacity and performance.

We choose Inception-V3 for two main reasons. First, Inception-V3 network employs factorized inception modules, allowing the network to choose suitable kernel sizes for the convolution layers, which enables the network to learn both low-level features with small convolutions and high-level features with larger convolutions. Second, the computational efficiency and low parameter count advantages of Inception have made it feasible to utilize Inception networks in high-resolution scenarios.

It is well known that deep learning methods are heavily dependent on the size of the training dataset, with a network structure of higher complexity requiring more data to avoid overfitting and generalizing. Meanwhile, the breast pathological images provided are very large spanning 2048 × 1536 pixels. To address the problems of large image sizes and insufficient data, we extract patches from each image and augment the image by applying varying degrees of rotation and flipping the extracted patches. This mode of data augmentation is consistent with a real-world scenario, as there is no fixed orientation adopted by pathologists when observing and analyzing pathological images under a microscope. The label for each patch is inherited from the class assigned to the original image.

4.2. Patch-wise method

Transfer learning [30] is an approach that applies CNNs that are pretrained on large annotated image databases, such as ImageNet, from different domains for various tasks. With such an approach, the original network architecture is retained, and the pretrained weights are used to initialize the network. The initialized weights are constantly updated during the following fine-tuning stage, enabling the network to learn features specific to the target task. Currently, a large number of studies have demonstrated that fine-tuning is efficient for a variety of classification tasks in the medical domain. For example, a recent study showed that utilizing pretrained Google’s Inception-V3 network on ImageNet and fine-tuned using images of skin lesions achieved very high accuracy for classification of skin cancer, comparable to that of numerous dermatologists [31].

In this paper, we use a well-known CNN architecture, named Google’s Inception-V3, which is pretrained to 93.33% top-five accuracy on the 1000 object classes (1.28 million images) of the 2014 ImageNet Challenge. Then, we fine-tune Google’s Inception-V3 to learn domain- and modality-specific features for classifying breast pathological images. Such a fine-tuning approach is easier to optimize and enables the training of deeper networks, which correspondingly leads to an overall improvement in network capacity and performance.

We choose Inception-V3 for two main reasons. First, Inception-V3 network employs factorized inception modules, allowing the network to choose suitable kernel sizes for the convolution layers, which enables the network to learn both low-level features with small convolutions and high-level features with larger convolutions. Second, the computational efficiency and low parameter count advantages of Inception have made it feasible to utilize Inception networks in high-resolution scenarios.

4.2. Patch-wise method

Transfer learning [30] is an approach that applies CNNs that are pretrained on large annotated image databases, such as ImageNet, from different domains for various tasks. With such an approach, the original network architecture is retained, and the pretrained weights are used to initialize the network. The initialized weights are constantly updated during the following fine-tuning stage, enabling the network to learn features specific to the target task. Currently, a large number of studies have demonstrated that fine-tuning is efficient for a variety of classification tasks in the medical domain. For example, a recent study showed that utilizing pretrained Google’s Inception-V3 network on ImageNet and fine-tuned using images of skin lesions achieved very high accuracy for classification of skin cancer, comparable to that of numerous dermatologists [31].

In this paper, we use a well-known CNN architecture, named Google’s Inception-V3, which is pretrained to 93.33% top-five accuracy on the 1000 object classes (1.28 million images) of the 2014 ImageNet Challenge. Then, we fine-tune Google’s Inception-V3 to learn domain- and modality-specific features for classifying breast pathological images. Such a fine-tuning approach is easier to optimize and enables the training of deeper networks, which correspondingly leads to an overall improvement in network capacity and performance.

We choose Inception-V3 for two main reasons. First, Inception-V3 network employs factorized inception modules, allowing the network to choose suitable kernel sizes for the convolution layers, which enables the network to learn both low-level features with small convolutions and high-level features with larger convolutions. Second, the computational efficiency and low parameter count advantages of Inception have made it feasible to utilize Inception networks in high-resolution scenarios.

It is well known that deep learning methods are heavily dependent on the size of the training dataset, with a network structure of higher complexity requiring more data to avoid overfitting and generalizing. Meanwhile, the breast pathological images provided are very large spanning 2048 × 1536 pixels. To address the problems of large image sizes and insufficient data, we extract patches from each image and augment the image by applying varying degrees of rotation and flipping the extracted patches. This mode of data augmentation is consistent with a real-world scenario, as there is no fixed orientation adopted by pathologists when observing and analyzing pathological images under a microscope. The label for each patch is inherited from the class assigned to the original image.
Objects in pathology images possess various scales and high complexity, therefore, learning richer hierarchical representations is critical for image feature representation. CNNs have proven effective for image feature representation. Nevertheless, the convolutional features in the CNN gradually become coarser with increasing receptive fields. To make the feature representation of pathological image patches more representative, we efficiently combine features from multiple convolutional layers. Thus, we can retain richer multilevel and complementary information such as local textures and fine-grained details lost by higher levels. In practice, as shown in Fig. 3, we use the standard pretrained Inception-V3 from the TensorFlow slim distribution for feature extraction. We removed fully connected layers from the model to allow the networks to consume images of arbitrary size. Here, unlike most of the previous models that converted the last convolutional layer consisting of 2048 channels via global average pooling into a one-dimensional feature vector with a length of 2048, we use richer multilevel convolutional features to fit the characteristics of this mission. Specifically, we use average pooling on the final output of the last three inception modules. Then, we concatenate the three vectors into a 5376D (1208D + 2048D + 2048D) dimensional vector, which is used as the richer multilevel feature representation of the pathological image.

### 4.3. Image-wise method

When deep learning is applied to process natural images, a complete image is directly used as input for end-to-end training. However, the size of the pathological image is too large, limited by memory size, and the original image is inevitably divided into several small patches. The accompanying problem is how to integrate the results of each small patch and obtain the image-wise classification result. The general methods are majority voting or SVM. These two kinds of methods are simple and direct, but they also have achieved good results. However, such a simple and direct method loses considerable contextual information about the whole image. How to keep the contextual information of each small patch has been the recent research focus. The two methods proposed in recent works by Awan et al. [24] and Nazeri et al. [32] were also dedicated to retaining contextual information and achieved good performance. The method proposed by Awan et al. [24] bound four feature vectors extracted from the spatially close patches and flattened into one. However, this simple flattened approach has difficulty integrating spatially close features. Another method to preserve the contextual information of each small patch was proposed by Nazeri et al. [32]. In their method, the first patch-wise CNN acts as an autoencoder that extracts the most salient features of image patches, while the second image-wise CNN performs classification of the whole image. However, they retained only the information of the top, bottom, left and right of a patch, and the remote context information still was not retained.

In response to the above disadvantages, we propose utilizing an RNN to fuse the contextual information of features that is directly incorporated on top of a CNN feature extractor to make the final image-wise classification decision. In our method, the CNN captures patch features, while the RNN captures short-term and long-term dependencies between the patches to retain contextual information. LSTM is one of the most common variations of RNNs. The connections between RNN units form a directed loop, which creates the internal state of the network, giving the network the ability to remember inputs at a distance. To further improve our model, we use the bidirectional long short-term memory network (BLSTM) [33], which is an extension of LSTM. BLSTM combines the output of two LSTMs, one to process input data from left to right, and the other to process input data from right to left. This structure provides the output layer with complete past and future contextual information for each position in the input sequence, which is also consistent with our problem’s characteristics that the context of each patch in a pathological image has no difference between up and down or left and right. In our proposed method, 12 feature vectors (12 × 1 × 5376) can be extracted from one pathological image by the CNN. These 12 feature vectors are inputted into a 4-layer bidirectional LSTM. Finally, we add a fully connected layer in the last layer of the LSTM (as shown in Fig. 4). Because we are performing 4 classifications, the output of this fully connected layer has 4 nodes.
5. Results

In this section, we present the performance of our proposed algorithm on our released dataset. All experiments in this paper are finished on NVIDIA Tesla K40 GPUs using the TensorFlow [34] framework. We mainly use accuracy and sensitivity to evaluate the performance of our method. If an image belongs to the category, it is classified as positive; otherwise, it is classified as negative. The sensitivity and accuracy can be defined as follows:

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]

where TP, TN, FP and FN are the true-positive, true-negative, false-positive and false-negative predictions, respectively.

5.1. t-SNE Visualization

T-distributed stochastic neighbor embedding (t-SNE) [35] is a nonlinear dimensionality reduction method that is excellent for embedding high-dimensional data for visualization in a low-dimensional space. Specifically, it models each high-dimensional object by a two- or three-dimensional point such that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. The core idea behind t-SNE is to find a two-dimensional representation of the data, maintaining the distance between the data points as much as possible. Fig. 5a and b show the two-dimensional (2D) representation of 12 feature vectors extracted from 12 patches from a breast cancer pathological image using t-SNE. Each data point in Fig. 5a and b represent the feature vector extracted from the corresponding patch in Fig. 5c and d.

The reason that the data points represent some spatially remote patches that are very close to the final 2D representation is that tumor features are diverse in pathology and irregularly distributed. Therefore, in the routine pathological diagnosis, spatially remote patches with key features for the classification of breast tumors are also an important basis for the final classification decision. As shown in Fig. 5c, Patches 2, 3, 9 and 10 all have key pathologic features of invasive carcinoma, such as the tumor cells showing diffuse or sheet distribution, small adenoid structure or cell nests with tubular structure. Jointly considering patch 2 with the spatially remote patches 9 or 10 can make the classification decision about invasive breast cancer as accurately as jointly considering patch 2 with the spatially close patch 3.

Moreover, in some cases, distant dependencies play a more important role in making the final classification decision than close dependencies. As shown in Fig. 5d, patches 6, 8 and 9 all have key pathologic features of ductal carcinoma in situ, such as the tumor cells in the duct arranged in solid nests with nuclear hyper-chromatism, heterogeneous pathologic condition in which malignant epithelial cells are confined within the ducts of the breast without evidence of invasion, etc. Joint considering patch 6 with the spatially remote patches 8 or 9 can make the classification decision as ductal carcinoma in situ more accurately than joint considering patch 6 with the spatially close patches 7 or 10.

Thus, by jointly considering the short-term and the long-term spatial correlations between patches, our proposed method not only deep mines the pathologic features of breast cancer but also simulates a real-world scenario in which a pathologist analyzes the pathological images.

5.2. Accuracy comparison with previous methods

The performance of our proposed method on patch-wise and image-wise accuracy is shown in Table 3. We compared the average classification accuracy with most of the advanced methods. Because some of

Fig. 5. Visualization of two pathological images using t-SNE. Fig. 5a and b show the two-dimensional (2D) representation of 12 feature vectors extracted from 12 patches from a breast cancer pathological image using t-SNE. Each data point in Fig. 5a and b represent the feature vector extracted from the corresponding patch in Fig. 5c and 5d.
the previous work used the Bioimaging2015 dataset with 249 training images and others used the ICIAR2018 dataset with 400 training images, and since 249 and 400 images are not considerably different, for convenient and concise comparison, we only compared the accuracy of the method in the case of the 400 training images.

For the 4-class pathological image classification, our method achieved 82.1% average accuracy in patch-wise and 91.3% average accuracy in image-wise. There are mainly two reasons for the good performance on the patch-wise. We use a pretraining model that allows for better generalization on a smaller number of pathological image datasets. Additionally, in contrast to previous work that used only the original CNN architecture, we use the more advanced Google’s Inception-V3, which ensures the model’s better learning ability. In addition, we analyzed the reasons for achieving image-wise good performance. Because we use richer multilevel feature representation to represent patches, image-wise information fusion can be more complete. At the same time, we use an integration method of a deep neural network to preserve the short-term and long-term spatial correlations between patches.

5.3. Accuracy comparison with different combinations of methods

We compared the average classification accuracy of different combinations of patch-wise and image-wise methods. It should be noted that to be comparable to previous work, the dataset we use is the same size as the dataset (ICIAR2018) that has been used by most current work. Therefore, rather than experimenting on our complete dataset, we randomly select 400 images for the training set and test them on another 100 images. To select the CNN model suitable for pathological images, we fixed the image-wise phase with the method of majority voting and then used different patch-wise CNN models for training. We tried some of the most mainstream approaches. The three most representative are listed in Table 4. The results showed that the model proposed by Oxford University’s Visual Geometry Group (VGG) had general performance, which may be related to the VGG model being proposed very early, and the follow-up work made many improvements on its foundation.

Moreover, the complexity of the models proposed later is relatively high, which has a great promoting effect on the learning ability of the algorithm. From the experimental results, Google’s Inception-V3 and ResNet [40] achieved almost the same results, but considering the computational efficiency and low parameter count advantages of Google’s Inception, it may be suitable for high-resolution pathological image classification tasks. Therefore, we finalized the use of Google’s Inception as our method for the model of image feature extraction.

Next, to compare the influence of different image-wise methods on the whole model, we fixed the patch-wise phase to use Google’s Inception-V3 and started training from scratch. The experimental results show that the SVM method achieves better results than the majority voting method.

Then, we compare the influence of the patch-wise pretraining CNN model on the overall result. In the patch-wise phase, we fixed the image-wise phase using the SVM method and used Google’s Inception-V3 model pretrained on ImageNet and the model trained from scratch. As seen from the experimental results, the pretrained model achieved better results. Because our dataset is relatively small compared to the natural image dataset, the pretrained model can help us better initialize and converge. This result has been shown many times in other areas of medical imaging [41].

Furthermore, we verify the overall effectiveness of our proposed algorithm. In the patch-wise, we used Google’s Inception-V3 pair with fine-tuning, and in the image-wise phase we used BLSTM with 4 layers. This model achieved the best average accuracy of 90.5% in the test set. Finally, we used richer multilevel features in the patch-wise phase, and the average classification accuracy was further improved by 0.8%.

5.4. Confusion matrix and AUC

The confusion matrix of the predictions on the test set is presented in Fig. 6 using a model trained on the dataset that contains a total of 400 images. As with the experimental setup in section 5.3, we randomly selected 400 images for the training set and tested them on another 100 images from our released dataset for comparability with the previous method. It can be seen from the confusion matrix in Fig. 6 that the categories of normal, benign, in situ and invasive all obtain high classification accuracy. Specifically, the in situ and invasive categories obtain classification accuracy of 95% and 97% respectively. However,
in comparison, the classification accuracy of normal and benign categories is only 86% and 87%. What is more significantly is that 10% of normal categories are misclassified as benign categories and 10% of benign categories are misclassified as normal categories. The same phenomenon can be found in the figure of Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC). Fig. 7 shows the mean AUC value of 89.25%, corresponding to 85%, 86%, 92% and 94% for the four classes based on receiver operating characteristic analysis.

In general, from the two experimental results shown in Figs. 6 and 7, we can see that the classification result of benign and normal is relatively lower than in situ carcinoma and invasive carcinoma. The reason for this phenomenon is that the subclass of benign and normal is not only diverse but also closely related to the age of the patient. Therefore, in the case of a limited number of pathological images, it is difficult to cover enough features of pathological images of benign and normal. For this reason, the final classification result is relatively low. To alleviate this problem, our dataset deliberately collected different subclasses of benign pathological images spanning different age groups. Because the majority of patients who go to the hospital for pathological examination are abnormal, there are very few clinical normal records. Therefore, we focused on only the benign categories. We will show the advantages of our dataset in the following experiment section.

5.5. Sensitivity comparison between different datasets

To illustrate the advantages of our proposed dataset, especially the diversity of benign pathological images, we performed experiments on different datasets using the same method. The comparison of the average sensitivity of image-wise results using ‘Google’s Inception-V3 + SVM’ method between the Bioimaging2015 dataset and our dataset is shown in Fig. 8. From the figure, it can be seen that after using a larger dataset, each class of sensitivity is improved, especially the classification sensitivity of benign images was relatively lower. For example, the method proposed by Araújo et al. describes that the image-wise sensitivity of benign images is only 66.7%, but the image-wise sensitivity of normal, in situ and invasive is 77.8%, 77.8% and 88.9%, respectively, because the characteristics of benign images are not salient, they can be subdivided into many subcategories. Moreover, their characteristics show greater diversity with age.

Therefore, to accurately classify benign images, more adequate dataset volumes and data diversity are needed to train the algorithm.

6. Conclusion

In this article, we proposed a new method for breast cancer pathological image classification using a hybrid convolutional and recurrent deep neural network. Based on the richer feature representation of the pathological image patches, our method considered the short-term and the long-term spatial correlations between patches through a RNN, which is right behind a richer multilevel CNN feature extractor. Thus, the short-term and long-term spatial correlations between patches were both considered. Through extensive experiments and comparisons, it was shown that our new method outperforms the state-of-the-art method. Additionally, we released a larger and more diverse dataset of breast cancer pathological images to the scientific community. We hope that the dataset can serve as a benchmark to facilitate a broader study of deep learning in the field of breast cancer pathologic images.

For the future work, to improve the accuracy of classification, outstanding deep learning algorithms and large enough as well as diverse dataset are indispensable. In terms of algorithms, the use of attention mechanisms in deep learning algorithms is a direction that can be tried, because it has achieved outstanding performance in natural image processing. In terms of dataset, larger dataset should be opened like ImageNet to provide a benchmark for the research community. Of course, advances in hardware are equally important. After all, it is ideal to directly use a complete high-resolution image as input to a deep neural network. At the same time, we are trying to extend this approach to whole slide images which will more difficult but will produce greater value in clinical practice.

Acknowledgements

This research is supported by Peking University International Hospital Research Grant (No. Y2018ZD005), the National Key Research and Development Program of China (No. 2017YFE0103900 and 2017YFA0504702), the NSFC projects Grant (No. U1611263, U1611261, 61502455 and 61672493), Beijing Municipal Natural Science Foundation Grant (No. L182053) and Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase).
References


