

# BREAST CANCER IMAGE CLASSIFICATION ON WSI WITH SPATIAL CORRELATIONS

Jiandong Ye, Yihao Luo, Chuang Zhu\*, Fang Liu, Yue Zhang

Beijing University of Posts and Telecommunications, Haidian District, Beijing, China

## ABSTRACT

As common cancer, breast cancer kills thousands of women every year. It's significant to provide doctors computer-aided diagnosis (CAD) to ease their workload as well as improve detection quality. Patch-level CNNs are usually used to classify the breast tissue slice, and the CNNs classify each patch independently ignoring the spatial correlations, resulting in wrong isolated label map. However, the probability distribution of cancer type is related to their adjacent patches. In this paper, we propose a framework integrating CNN and filter algorithm aimed at extracting spatial information and improving the performance of the classification. The network was trained on a breast cancer dataset provided by ICIAR18. For 4-class classification, compared to CNN methods without using spatial correlations, the proposed method achieved about 10% improvement on accuracy over the validation dataset and get smoother probability maps. Our experiments also show that larger kernel size gets better performance. The code is available at <https://github.com/dong100136/Breast-Cancer-Image-Classification-On-WSI-With-Spatial-Correlations>.

**Index Terms**— CNN, Filter Algorithm, Breast Cancer Detection, Spatial Correlations

## 1. INTRODUCTION

Breast cancer is one of the most common cancers in women of all ages and is the second leading cause of cancer death among women after lung cancer [1]. R.A. Smith et al. [2] found that early diagnosis and treatment of breast cancer can significantly reduce mortality. Currently, for diagnosis of breast cancer types, breast tissue biopsy is required. By surgery or other means, doctors obtain some breast tissue. The collected tissue samples will be stained with hematoxylin and eosin (H&E) and observed via an optic microscope for further analysis. Breast cancer cells have a wide variety of morphology, which can be divided into benign tumors, in situ tumors and invasive tumors. An excellent pathologist needs years of training as well as heavy investment. The number of breast tissue slices is increasing, causing great pressure on doctors. Therefore, it is meaningful to study computer-aided diagnosis (CAD) for breast cancer detection based on artificial intelligence. With high resolution image data and

intelligent algorithm, a computer can automatically analyze the tissue samples and mark suspected lesion in a very short time. Related algorithms greatly ease the diagnosis workload and effectively improve the quality of diagnosis [3].

In recent year, convolutional neural network (CNN) has made a great success on image classification. Patch-level detection on a whole-slide image (WSI) is popular due to the high resolution of medical images with huge data size, which is hard for CNNs to handle directly [4][5]. However, spatial correlation is ignored because CNNs deal with each patch classification independently.

In this paper, we propose an approach to improve the performance of classification of breast cancer. Our approach integrates CNNs and filter algorithms to make better use of spatial correlations in the tissue slice comparing with single patch-level CNNs. The use of spatial information can get smoother probability maps and better performance in the classification of breast cancer.

## 2. RELATED WORK

CNNs, such as VGG [6], ResNet [7], Inception Network [8][9] have a great success on a wide range of computer vision tasks, e.g. image classification, object detection, and semantic segmentation.

Since WSIs usually have high resolution and huge data size, training a CNN on WSIs directly requires a very large memory on GPU, which is impossible in most cases. Several patch-level CNN methods were proposed by [4][5][10]. Because of the high resolution of WSIs, most of the studies extract small patches (e.g.  $224 \times 224$  pixels) from WSIs and predict these patches independently. For each patch, the classification is independent. But the label distribution of neighboring patches is correlated and neighboring patches usually get the same label. There are spatial correlations shared between the small patches and their neighboring patches. Ignoring these spatial correlations, the prediction of CNN may contain the wrong isolated result.

Several researchers have paid attention to spatial correlations between patches. Kong et al. [11] proposed Spatio\_Net, which integrated CNN with the 2D long short-term memory (2D-LSTM). At the same time, some works by Zanjani et al. [12] also used CNN as feature extractor, and they implemented conditional random field (CRF) as a post-processing

\*the corresponding author: Chuang Zhu (czhu@bupt.edu.cn)

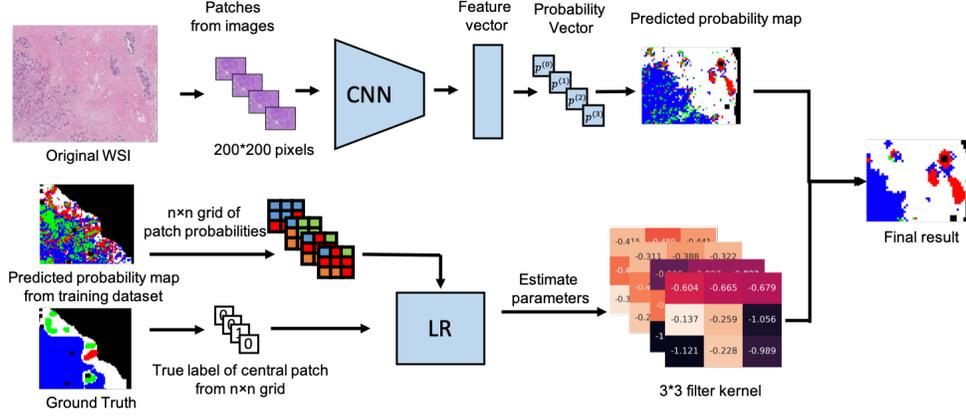


Fig. 1. Illustration of the framework for the use of spatial information

stage. CRF is a probabilistic graphical model, which has been applied to sequence labeling and semantic segmentation. Li et al. [13] combined CNN and CRF and proposed an end-to-end framework by mean-field approximate inference algorithm for better performance. However, the combination of CNN and CRF is complicated and required a lot of computing resources. Our work focuses on proposing a more efficient framework to utilize the spatial correlations.

### 3. METHOD

#### 3.1. Motivation and Overview

Patch-level CNNs, which classify every patch independently, will generate many wrong isolated probabilities, as depicted by the predicted probability map from CNNs in Fig. 1. As shown in the masks, the region of the lesion area is continuous, indicating that the labels of adjacent patches tend to be consistent. From the perspective of image signals, the wrong prediction from CNNs is the noise in the probability map. Inspired by this, we try to apply a filter kernel to remove the noise. On the other hand, the filter kernel takes advantage of the label distribution of neighboring patches to reestimate the output of CNNs. As a consequence, the model considers not only the texture information but also the labels of neighboring patches, obtaining a smoother probability distribution.

In this section, we put forward a two-stage framework for classification of the patches from WSI. Firstly, we trained a CNN to classify the patches and use the network to extract probabilities from patches. Then, we filtered the probability map from the CNN with a filter kernel to get a better analysis by taking advantage of spatial correlations. An overview of this framework is illustrated in Fig. 1.

#### 3.2. Obtain Probability Map from CNN

In the first part, we utilized the CNN as feature extractor. Deep CNNs usually consist of several convolutional lay-

ers, max-pooling layers as well as activation functions like ReLU, Softmax. In fact, CNN can be used as the classifier directly, but it will distinguish the cancer label of the patch independently ignoring the label distribution of the neighboring patches.

#### 3.3. Reestimate with Filter Kernel

For patches from a WSI, we can feed them into the trained CNN model and get a probability map. The probability map that patches belong to class  $r$  represents as  $P^{(r)} = \{p_{i,j} | 0 \leq i, j \leq 1\}$  and  $r \in \{0, 1, 2, 3\}$ . Let  $F^{(r)}$  be the filter kernel for class  $r$  and the reestimate probability map  $\hat{P}^{(r)}$  is given as:

$$\hat{P}^{(r)} = P^{(r)} * F^{(r)} \quad (1)$$

From the perspective of each  $Patch_{i,j}$ , its probability  $p_{i,j}^{(r)}$  is:

$$\hat{p}_{i,j}^{(r)} = \sum_{n=-K/2}^{K/2} \sum_{m=-K/2}^{K/2} p_{i+n,j+m}^{(r)} \times w_{n,m}^{(r)} \quad (2)$$

where  $w^{(r)}$  is the parameter of filter kernel  $F^{(r)}$  and  $K$  is the size of filter kernel, which can be 3 for a grid of  $3 \times 3$  kernel or 5 for a grid of  $5 \times 5$  kernel. If the label of  $Patch_i$  is same as its neighboring patches, it will get a higher probability. In contrast, it will get a lower probability when its label differs from its neighbors.

If we concatenate the probabilities of a patch and its neighboring patches into a single probability vector as  $x_i$ , and let  $y_i^{(r)} \in \{0, 1\}$  be the label of  $Patch_i$  from the grid of patches. The problem can be solved as an optimize problem for each label as below:

$$\arg \min_{w^{(r)}} \sum_{i=1}^N \log(1 + \exp(-y_i^{(r)} w^{(r)T} x_i)) \quad (3)$$

We use logistic regression algorithm as the strategy for parameter estimating of the filter kernel. LR was chosen as our fusion strategy owing to its robustness and efficiency against high-dimensional features and multi-class task.

The probability  $P(y_i = r|x_i)$  that  $Patch_i$  belong to class  $r$  is given as:

$$P(y_i = r|x_i) = \frac{\exp(-w^{(r)T} x_i)}{\sum_{r=1}^R \exp(-w^{(r)T} x_i)} \quad (4)$$

where  $R$  is the number of cancer type.

## 4. EXPERIMENT

### 4.1. Dataset and Environment

The proposed method was applied to the dataset from ICIAR 2018, which consists of 10 high resolution WSIs. Each WSI has multiple normal, benign, in situ carcinoma and invasive carcinoma regions and each WSI has a corresponding list of labeled coordinates that enclose benign, in situ carcinoma and invasive carcinoma regions, and the rest is marked as normal, as depicted in Fig. 2 [14]. With these labeled coordinates we can make masks of these WSIs and extract fixed size patches by sliding a patch window.

The approach we proposed was implemented with keras-2.0.8 and Tensorflow-1.4.1 in python 3.6 with NVIDIA GeForce GTX 1080 Ti GPU.

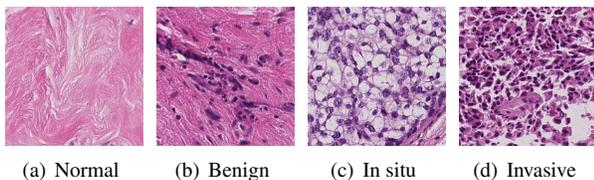


Fig. 2. Examples from the dataset

### 4.2. Implementation Details

We used WSI-3 to WSI-10 for the training dataset and the rest of the slides for the test dataset. We slid a window of  $200 \times 200$  pixels to get patches from WSI and recorded the coordinates of the center point of each patch. Before using these patches as the training dataset, we ran an algorithm to remove the background patches depending on its histogram, which is concentrated in a small range. Then we tagged each patch according to its mask.

The implementation of our proposed algorithm can be divided into two parts.

In the first part, we used VGG16 [6], ResNet50 [7], InceptionV3 [8] and InceptionV4 [9] provided by Keras as feature extractor. For each coordinate of the center points, we take patches of  $200 \times 200$  pixels and then feed them into the CNNs for training. The training dataset is too small for the model,

and it is easy to get overfitting. So, we fine-tuned the model weights of the Imagenet-pre-trained CNNs. We used stochastic gradient descent of learning rate  $1e-7$  and a momentum of 0.01 to optimize the architecture.

Then the second part is a filter algorithm, which is the Logistic Regression in this experiment. The algorithm takes the probabilities extracted by the CNN models as the input features. For each coordinate of the center points, we take a patch of  $600 \times 600$  pixels and then cut it into a grid of patches by  $200 \times 200$  pixels. The training dataset was as same as the first part.

### 4.3. Result

#### 4.3.1. Better Performance with Spatial Information

Table.1 shows patch-level classification accuracies of the CNNs on validation dataset as well as the result of filter algorithm base on them. Compared with the output of the CNNs, the overall accuracies improved by about 10%, suggesting the probability distribution of a patch is related to its neighboring patches. Surprisingly, the VGG16 gets better performance on the baseline than the other networks which is not the same as their performance on Imagenet dataset [15]. We consider that the different data scale and the different characteristics of datasets lead to this phenomenon.

Then we explored the effectiveness of different filter kernel size of  $k \times k$ , where  $k \in 3, 5, 7$ . The result is summarized in Table.1. As the kernel size gets larger, which means the algorithm can get more spatial information from the neighboring patches, the performance becomes better. However, the reward is becoming smaller as the kernel getting larger.

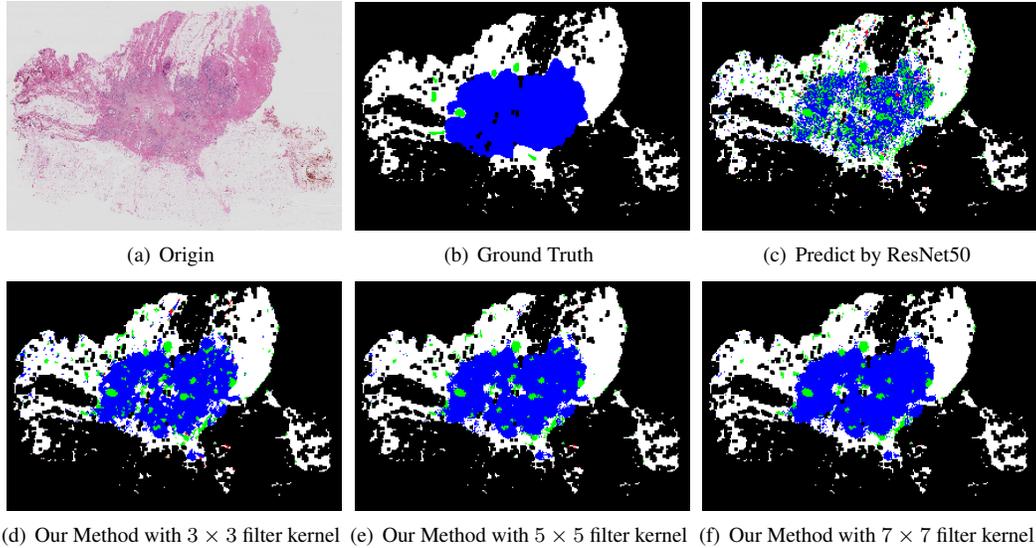
Table 1. Patch-level accuracies of different kernel size

	baseline	$3 \times 3$	$5 \times 5$	$7 \times 7$
VGG16 [6]	0.7759	0.8657	0.8917	0.9046
ResNet50 [7]	0.7509	0.8759	0.8954	0.9074
InceptionV3 [9]	0.7583	0.8907	0.9074	0.9176
InceptionV4 [16]	0.7454	0.8769	0.8898	0.9028

Furtherly, we implemented our method with ResNet18 [7] on 2-class classification for tumor and non-tumor and compared it with NCRF [13], which was trained with our dataset. The result on Table.2 shows that our method achieves higher patch-level accuracies.

Table 2. Patch-level accuracies on 2-class classification

kernel size	baseline	NCRF	Our method
$3 \times 3$	0.8144	0.8220	0.8703



**Fig. 3.** Predicted probability maps of A10 from ResNet50 (a) original WSI (b) ground truth annotation (c) probability distribution from ResNet50 (d) probability distribution from our method with  $3 \times 3$  kernel (e) probability distribution from our method with  $5 \times 5$  kernel (f) probability distribution from our method with  $7 \times 7$  kernel  
 (White) normal ■ benign ■ in situ carcinoma ■ invasive carcinoma ■ background

#### 4.3.2. WSI-level Performance

Finally, we applied our method to WSIs and tested the trained model on WSI-level classification.

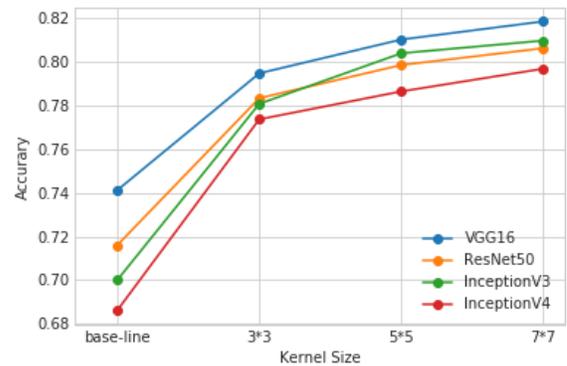
Fig.3 shows the visualization of classification result on a WSI. Table.3 and Fig.4 show the performance of our method compared to the baseline of CNNs. With the filter algorithm, the classification maps get smoother compared to the result from CNN. As we can see, with the spatial information from neighboring patches, the algorithm can reduce the isolated result from the CNNs. The label distribution of each patch is not conditionally independent. The result also suggested that the kernel size plays an import part in the classification.

**Table 3.** WSI-level accuracies of different kernel size

	baseline	$3 \times 3$	$5 \times 5$	$7 \times 7$
VGG16 [6]	0.7412	0.7948	0.8103	0.8186
ResNet50 [7]	0.7160	0.7834	0.7986	0.8063
InceptionV3 [9]	0.7000	0.7808	0.8040	0.8098
InceptionV4 [16]	0.6861	0.7737	0.7865	0.7969

## 5. DISCUSSION

This work focuses on taking advantage of spatial correlations between patches. The CNN can classify the patches from the WSI, but the result from the CNN will contain many wrong isolated predicted labels as it ignores the spatial correlations. Compared to the previous methods without tak-



**Fig. 4.** WSI-level Accuracy of different kernel

ing advantage of spatial correlations, our framework can get smoother label maps. Further experiments show that larger kernel size can achieve better performance in classification but the reward is becoming smaller as kernel size getting larger.

## 6. ACKNOWLEDGEMENTS

This work is supported in part by the Beijing Natural Science Foundation (4182044), Basic scientific research project of Beijing University of Posts and Telecommunications (2018RC11). This work is conducted on the platform of Center for Data Science of Beijing University of Posts and Telecommunications.

## 7. REFERENCES

- [1] Carol E Desantis, Jiemin Ma, Ann Goding Sauer, Lisa A Newman, and Ahmedin Jemal, "Breast cancer statistics, 2017, racial disparity in mortality by state," *Ca A Cancer Journal for Clinicians*, vol. 67, no. 6, pp. 439, 2017.
- [2] R. A. Smith, Mettlin CJ, Davis KJ, and Eyre H, "American cancer society guidelines for the early detection of cancer.," *Ca A Cancer Journal for Clinicians*, vol. 52, no. 1, pp. 8–22, 2002.
- [3] Rachel F Brem, Jocelyn A Rapelyea, Gilat Zisman, Jeffrey W Hoffmeister, and Martin P DeSimio, "Evaluation of breast cancer with a computer-aided detection system by mammographic appearance and histopathology," *Cancer*, vol. 104, no. 5, pp. 931–935, 2005.
- [4] Dan C. Cireşan, Alessandro Giusti, Luca M. Gambardella, and Jürgen Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2013, pp. 411–8.
- [5] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.
- [6] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [8] C Szegedy, W Liu, Y Jia, P Sermanet, S Reed, and D Anguelov, "& rabinovich, a.(2015). going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.
- [9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [10] Teresa Araújo, Guilherme Aresta, Eduardo Castro, José Rouco, Paulo Aguiar, Catarina Eloy, António Polónia, and Aurélio Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PLoS one*, vol. 12, no. 6, pp. e0177544, 2017.
- [11] Bin Kong, Xin Wang, Zhongyu Li, Qi Song, and Shaoting Zhang, "Cancer metastasis detection via spatially structured deep network," in *International Conference on Information Processing in Medical Imaging*, 2017, pp. 236–248.
- [12] Farhad Ghazvinian Zanjani, Svitlana Zinger, et al., "Cancer detection in histopathology whole-slide images using conditional random fields on deep embedded spaces," in *Medical Imaging 2018: Digital Pathology*. International Society for Optics and Photonics, 2018, vol. 10581, p. 105810I.
- [13] Yi Li and Wei Ping, "Cancer metastasis detection with neural conditional random field," 2018.
- [14] "Icifar 2018 : 15th international conference on image analysis and recognition," <https://www.aimiconf.org/icifar18/>.
- [15] "The top-1 and top-5 accuracy refers to the model's performance on the imagenet validation dataset," <https://keras.io/applications/>.
- [16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning.," in *AAAI*, 2017, vol. 4, p. 12.