Automatic Breast Cancer Tumor Detection and pN Staging using Deep Learning

Vinay Singh, Prakash Gupta

ABSTRACT

This paper presents and evaluates automatic breast cancer metastases detection in lymph nodes whole-slide images (WSIs). The detection is performed in slide-level and patient-level processing. The pN-stage for every patient is determined by the number of positive lymph nodes that consists of 5 categories. We used convolutional neural networks for slide level detection of tumor cells. The pN-stage evaluation has been done as post processing stage for detected positive regions by using contour analysis and trained random forest.
This method is trained and validated on Camelyon17 challenge datasets. The method consists of two parts: tumor region detection using CNN and cancer metastasis stage prediction using Random Forest.

Index Terms— ResNet-18, convolutional neural networks, Camelyon17, breast cancer detection

1. INTRODUCTION

The focus of the challenge is on designing an automatic way to detect and classify breast cancer metastases in lymph nodes. To achieve the goal, whole slide images (WSI) of the hematoxylin and eosin stained lymph nodes tissue were provided. This is one of the ends that the Camelyon17 challenge [1] aims to provide for by predicting pN-stages of 100 patients. The
challenge dataset consists of 5 hematoxylin and eosin stained slide images of different lymph nodes from each subject. Five patient-level classes of pN stage, namely pN0, pN0(i+),
pN1mi, pN1, and pN2, are automatically determined by 5 slide-level metastases, which are negative, micro-metastases, macro-metastases, isolated tumor cells (ITCs). Therefore, it is critical to classify each slide correctly.

2. METHOD

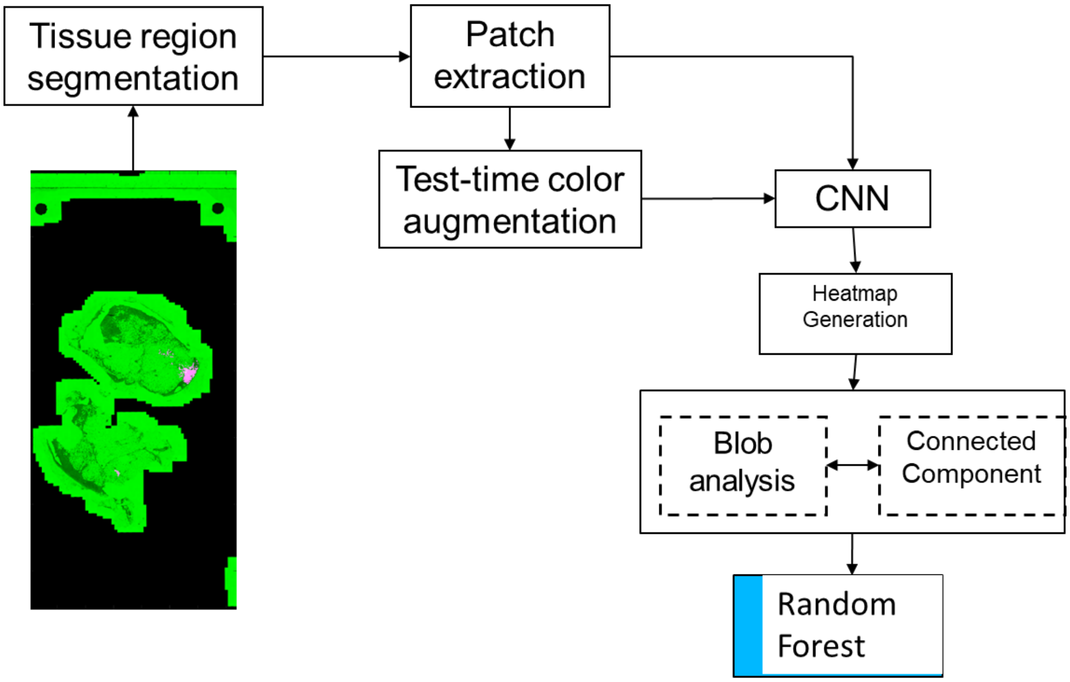
A deep learning based typically consist of training and testing. We have used camelyon’16 and 17 data to ResNet-18 model.

2.1 Pre-Processing

We have extracted valid tissue regions from whole slide imaging using Otsu methods [1][2].

2.2 Patch Extraction

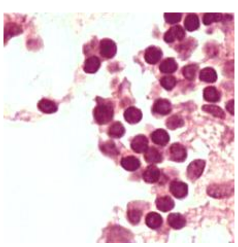
Task was to get label for every patch extracted from valid tissue region. We have used ground truth given in training dataset to get appropriate label. We have taken a special care for the patches at the boundary.

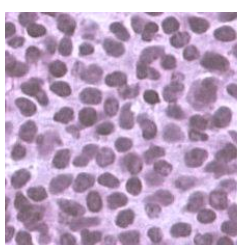


Used a threshold to decide a patch is tumor or normal.

2.3 Data Augmentation

We have use very diverse set of data augmentation to incorporate all possibilities. This includes- Random Image Rotation, Random Image left right flip, Random Image Following are few data augmented images:





Up and Down Flip, Random Hue, Random Saturation and Random Brightness.

2.3 Convolutional Neural Network

The extracted patches after going through data augmentations are fed into Convolutional Neural Network, we have used a popular variant called ResNet – 18. We initialized this with pretrained weights from imagenet data.

2.4 Heatmap Generation

For each whole slide image, after extracting the tissue region from a lower magnification, we take the corresponding patch at highest magnification and run through our trained model, using the obtained probability we generate a probability heatmap for the whole slide image.

2.5 Feature Extraction

We extract 15 features by thresholding at 3 levels, in total we extract 45 features. After each thresholding we extract features like area of the largest tumor region, eccentricity, entropy, highest probability etc.

2.6 Whole Slide Classification

Using the extracted features from the previous step we train a random forest classifier for all 500 training slides, using five 5 fold cross validation and ensemble all 5 classifiers using a max vote method.

3. RESULTS

We have trained our model first only on CAMELYON16 Training dataset, to evaluate how good is it. We achieved a FROC score of 0.72

Using the same model we trained it on the remaining datasets (Camelyon16 Test as well as Camelyon17 Training). Our whole slide classifier was trained using the features extracted from the heatmaps generated by above model, our five-fold cross validation results were the following:

Model	**Score**	**Metric**
ResNet-18	0.72	FROC
Random Forest (1st Model)	0.52	Quadratic Kappa
Random Forest (2nd model)	0.48	Quadratic Kappa
Random Forest (3rd Model)	0.41	Quadratic Kappa
Random Forst (4th Model)	0.51	Quadratic Kappa
Random Forest (5th Model)	0.50	Quadratic Kappa
Random Forest (overall)	0.49	Quadratic Kappa

We have achieved competitive performance on our tumor classification model and whole slide image classification model.

4. References

[1] Nobuyuki Otsu, "A threshold selection method from gray-level histograms," IEEE transactions on systems, man, and cybernetics, vol. 9, no. 1, pp. 62–66, 1979

[2] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al., "Detecting cancer metastases on gigapixel pathology images, arXiv preprint arXiv:1703.02442, 2017.