AUTOMATIC CLASSIFICATION ON PATIENT-LEVEL BREAST CANCER METASTASES

Sanghun Lee, Sangjun Oh, Hyoeun Kim, Changdae Lee, Kyuhyoung Choi, Sun Woo Kim

Deep Bio Inc.

ABSTRACT

Automatic diagnosis of breast cancer is a challenge that promises more accessible healthcare. In this paper, we describe the process of predicting slide-level cancer metastasis with machine learning techniques. First, a whole slide image is split into smaller patches which are classified for cancer by DenseNet, a Deep Neural Network with established performance. Next, the patch-level results are aggregated into a confidence map, which then goes through DBSCAN, a clustering algorithm, to reveal morphological features of cancerous regions. Finally, the extracted features from individual slides are gathered to train XGBoost, a boosting algorithm, to predict slide-level diagnosis. The resulting slide-level results determine the pN stages of individual patients.

***Index Terms*—** Breast cancer, pN stage, deep neural network, DenseNet, DBSCAN, XGBoost

1. INTRODUCTION

Diagnosis of pathological whole slide images is not a trivial task; not only is the process time-consuming, but the results are also highly dependent on the pathologist's skill level, and even the most experienced pathologists may have discrepancies on viewing the same slide. Such circumstances may lead to unexpected human errors, which an automated diagnosis assistance system may be able to help prevent. 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.

Recently, Convolutional Neural Networks (CNNs), one of the Deep Neural Network models, have been showing revolutionary results on image recognition and classification tasks.[2] In ImageNet Large Scale Visual Recognition Competition (ILSVRC), advanced CNN model proves the state-of-the-art performance on classification of 1000 classes such as cats, dogs, etc. Since CNNs are not limited to a specific image domain, it can also be applied to the field of digital pathology. As an example, in a pathological study of breast cancer, CNN was used to achieve state-of-the-art results.[3]

In this paper, our automated diagnosis process consists of 3 steps. The first step is to classify patches whether these contain cancer or not. This is necessary due to the size of the whole slide images, considering that average size of whole slide image is about 200,000 by 100,000 pixels. Here, the patch classifier is trained on patches of 240 by 240 pixels, randomly generated from the annotated cancerous and non-cancerous regions. The second step is the hard example mining process. Upon observing the initial performance of generated heatmaps, additional normal patches were extracted from the training slides into the training dataset. The final step is to extract morphological features from each heatmap, and train XGBoost [4] , a boosting algorithm, to classify slide-level metastases. The patient-level pN stages are determined based on 5 slide metastases.

2. METHODS

The overall procedure, as illustrated in Figure 1, is as follows:

- Preprocess datasets, including patch extraction and color normalization.
- Train patch-level classifier with a pretrained version of DenseNet-121.
- Optimize the training data distribution with hard example mining.
- Extract morphological features from individual heatmaps clustered with DBSCAN algorithm. [5]
- Train slide-level classifier with XGBoost

2.1. Preparation of patch-level data set

The data set for training patch classifier comes from both Camelyon16 and Camelyon17 dataset (abbreviated to the '16 and the '17 dataset, respectively). Assuming that the '16 dataset contains two different medical centers, total seven different stain styles are included. Cancerous and non-cancerous

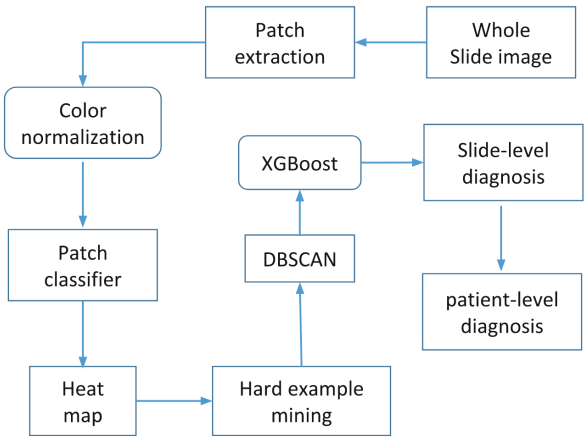


Fig. 1. whole procedure

patches in '17 dataset are extracted from the slides containing annotations. Patches with cancer are generated inside annotation areas, while normal patches are extracted outside annotation areas, randomly without intersection. To examine various statistics, we balance the number of patches almost equally between cancer and normal patches. The resulting total number of patches in training, validation, test are 95,149, 58,000, and 48,000 respectively.

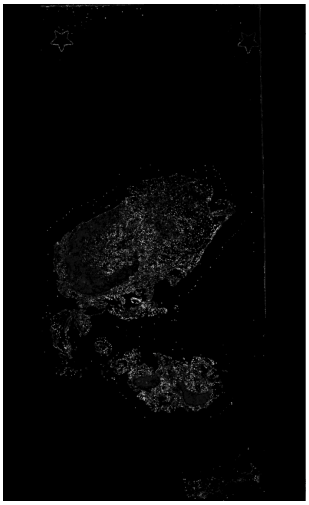
As color tone affects the performance of CNNs, the difference of staining styles between institutes poses a problem. [6] In this regard, the seven different stain styles can affect prediction accuracy, so we normalize the color with Generative Adversial Network (GAN). This makes different stain styles into similar color tones with same morphological distribution.

2.2. Train patch-level classifier

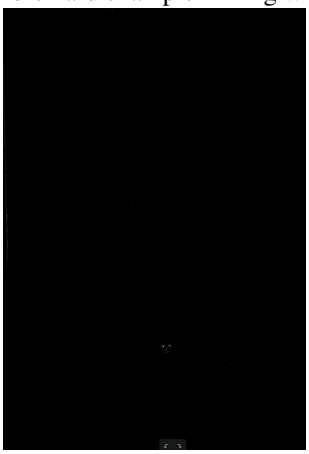
Patch classifier is trained by DenseNet-121 model pretrained with 1000-class ImageNet dataset. [7] Since our aim is to predict among two classes in patch-level prediction, a fully-connected layer from 1000 to 2 is added at the end of the original version. Initial learning rate is 0.1 and reduced by one tenth per 10 epochs, and the optimizer is SGD with 1e-4 decay.

2.3. Hard example mining with heatmap

By using the patch classifier, each whole slide image is transformed to a heatmap which considers a 240-by-240 pixel patch as a single pixel, as shown in Figure 2. The bright white space marks regions with high probability of cancer. Since normal patches may not be sufficient to represent the entire distribution of the normal types, additional normal patches are chosen from the heatmap regions that disagree the most with the reference annotation. Finally, the same patch classifier is trained again with the dataset with the additionally extracted normal types.



(a) Heatmap before hard example mining without threshold



(b) Heatmap after hard example mining above threshold

Fig. 2. Hard example mining shows significant decrease of false alarms.

2.4. Extracting morphological features

To classify slide-level metastases, morphological features from heatmap are extracted by DBSCAN algorithm as shown in Figure 3. Per each of the three largest clusters within a slide, features such as the major axis, minor axis, area, density, mean probability, max probability, and min probability are extracted.

2.5. Training and predicting slide-level classifier with XGBoost

Slides with 24 features are trained by XGBoost. Since each heatmap may or may not be represent the status of the metastases, 400 random slides from the given set of 500 are trained,

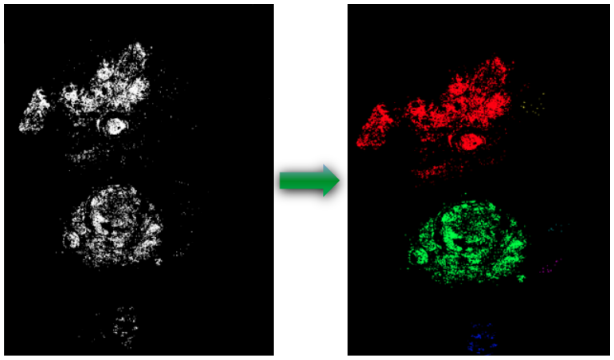


Fig. 3. The original heatmap(left) and the clusters after applying DBSCAN(right)

while the other 100 slides left out for validation. The well-formed criteria for hyperparameter tuning is to show high accuracy with both the entire 500 slides and the 100 validation slides which contains each types of metastases distributions similar to entire 500 slides.

The remaining task of predicting patient-level pN-stage is automatically determined by slide-level metastases predictions.

3. RESULTS

Patch-level classifier shows 0.99 and 0.98 ROC, PR-AUC respectively in both validation and test patches. Optimal threshold in the validation patch set is 0.58, as chosen for the highest F1 score. With this optimal threshold, accuracy, recall, specificity, and precision are 0.99, 0.98, 0.99 and 0.99, respectively, in the validation set.

In the heatmap level, many false alarm cases are improved by comparing original annotated slide to the heatmap after the hard example mining step.

Slide-level accuracy was 0.92 and 0.924 in the validation slides and the entire 500 slides. The kappa score for the entire 500 slides is 0.96. Note that the validation slides in XGBoost differs from 500 test slides used for the submission.

4. DISCUSSION

Our and usual method of approach to predict pN-stage is basically patch-level approach. This is because the whole slide image is too big to be trained in the memory restrictions of hardware, while the number of whole slide image is too small to train. A crucial limitation in this approach is that, although patch classifier reaches 0.99 accuracy, false alarm cases still occur in a certain ratio, which makes it difficult to catch ITC cases sensitively in slide-level.

In the process of training XGBoost, since the optimal 400 slides for training does not come from the same 80 patients and the number of 20 patients samples is too small, patient-level prediction is calculated on the entire 500 slides. To prevent overfitting, we set to balance accuracy between 100 validation slides and the entire 500 slides. In order to check pN-stage of the 100 validation slides, it is ideal to set 100 slides from 20 patients. However, as it is still important to have high accuracy in slide-level prediction and optimal 400 training set does not always come from 80 patients, we had to use whole 500 slides to calculate total weighted kappa score.

Annotation process is still time-consuming and expensive. In the field of pathology, it is costly to get a sufficient number of annotated data as the golden standard. Therefore, considering the limited number of annotated slides, unsupervised or at least semi-supervised learning is necessary to overcome the problem of annotation cost.

5. CONCLUSION

Patch-level to slide-level approach works generally well, but still many problems remain. Our team also tried end-to-end approach with limited size of samples, but due to small number of slides the result accuracy was significantly lower than that of patch-level approach.

In future work, we will try unsupervised and semi-supervised learning to overcome the limit number of annotated slides.

6. REFERENCES

[1] Oscar Geessink, Péter Bándi, Geert Litjens, and Jeroen van der Laak, "Camelyon17: Grand challenge on cancer metastasis detection and classification in lymph nodes," 2017.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[3] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep Learning for Identifying Metastatic Breast Cancer," *ArXiv e-prints*, June 2016.

[4] Tianqi Chen and Carlos Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, KDD '16, pp. 785–794, ACM.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, "A density-based algorithm for discovering

clusters a density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, KDD'96, pp. 226–231, AAAI Press.

[6] A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. M. Schlitter, A. Sethi, I. Esposito, and N. Navab, "Structure-preserved color normalization for histological images," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, April 2015, pp. 1012–1015.

[7] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.