BREAST CANCER METASTASES PREDICTION FROM WHOLE-SLIDES IMAGES USING DEEP SEMI NON-NEGATIVE MATRIX FACTORIZATION

Kaixian Yu, Rongjie Liu, Maomao Ding, and Hongtu Zhu

Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, 77030

ABSTRACT

The study aims at automated detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections. The whole-slides images, as a high-resolution image, have attracted more and more interests from the medical image analysis community. In this study, we present an analyzing framework for establishing a predictive model based on WSI of histological lymph node sections. The proposed workfolw consists of several key steps: i) metastases feature extraction, ii)foreground extraction, iii) classifier establishing, learning and prediction. In the modeling step, we explored dictionary learning and deep learning models. The performance is evaluated and compared based on AUC values. Our final results indicate that deep learning model can extract textual features of the metastatic regions. Specifically,

***Index Terms*—** Breast cancer, metastases, whole-slides images, dictionary learning, deep learning

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.

In our analysis, the goal is to combine the detection and classification of metastases in multiple lymph node slides to predict a patient's pN-stage, using machine learning algorithms. The main challenges lies in the large image size for a single slide.

We have designed a deep semi non-negative matrix factorization based pattern classification to classify patches of WSI. Eventually using the patches information to reconstruct the tumor size information in each WSI.

1.1. Metastases

Pathologists typically examine the sentinel lymph nodes excised from patients with invasive breast cancer more thoroughly than they have historically those from axillary lymph node clearance specimens. This, it is thought, increases the chances of detecting small metastatic foci (i.e. macrometastases (> 2 mm), micrometastases ($0.2 - 2$ mm), or isolated tumor cell clusters (< 0.2 mm or < 200 cancer cells in one section)). However, the clinical significance of these small metastatic deposits remains unclear.

2. MATERIAL AND METHOD

2.1. Imaging Data

Whole-slides images for 200 patients in total. 100 patients as training data and another 100 as testing data. Each patinets have 5 WSI for 5 lymph nodes. Among 500 WSI, there were 50 WSI with various tumor status annotated by pathologists. There are 4 status for the WSIs, Macro/micro/itc/normal; they differ in terms of the tumor region size/cells. One important assumption we made was that the textures of the local environment were the same except the size. Thereforem we could analyze the image of macro/micro/itc regions interchangeably.

2.2. Pre-processing

We used the Automated Slide Analysis Platform(ASAP) to view the WSI, and manually segmented the meaningful tissue regions for downstream analyzing. In addition, we converted the RGB image to a gray scale image for its computational convenience.

2.3. Image Feature Extraction

To collect abundant training samples, we chopped the WSI with annotated region into 256×256 smaller non-overlapping patches. And mark those with annotated tumor region as tumor patches. From here on, our analysis will focus on the patches.

We did the same chopping procedure to all training and testing WSIs, so that we could work on a much smaller image yet still keep the resolution so we will not lose valuable information.

2.4. Statistical Analysis

The feature map of the patches (tumor and normal) were obtained using deep semi non-negative matrix factorization

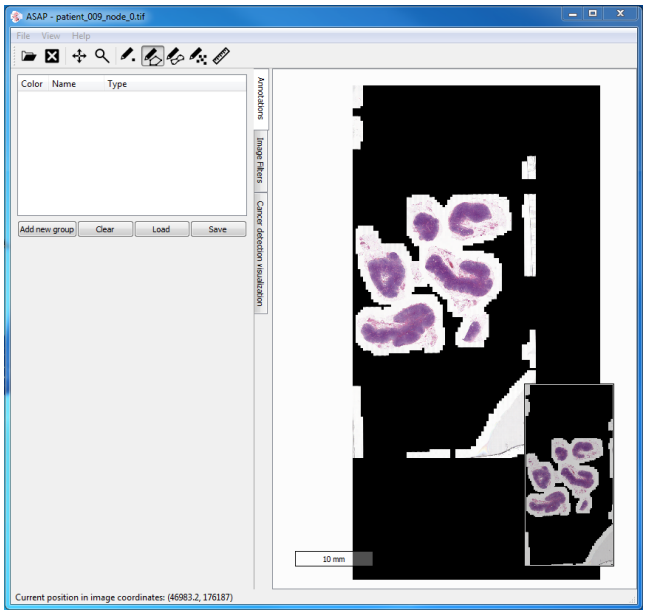

Fig. 1.

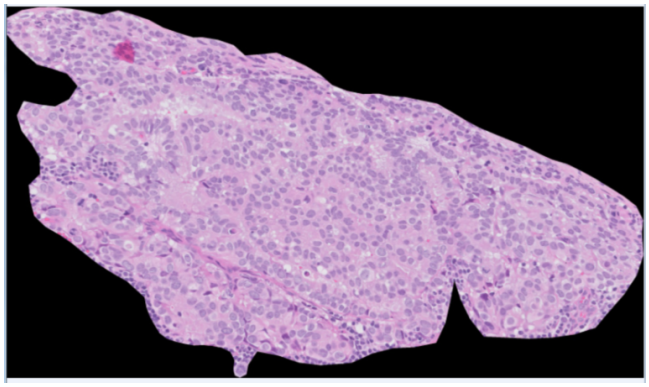

Fig. 2.

(DSNMF) [1], We set 4 layers of hidden states 400, 300, 200, 100 in this deep learning based DSNMF (Fig. 3). The output at the last layer was further used to build classification model, XGBoost [2].

The tuning parameters of XGBoost were fine tuned by 10 fold random split, where 50% of the training patches were randomly selected to train the model and the performance was accessed using the other 50% of the training patches. A grid searching for *learning rate, depth,* and *regularity parameters* were performed.

Every patch of a specific WSI were processed using the trained DSNMF model to extract features, and given a probability of being tumor texture according to the XGBoost and extracted features. Then the probabilities were mapped back to WSI to capture neighboring information to determine the size of the malignant site.

The pixel level size of macro/micro/itc sites were further learned from the given training WSIs where the status of each one is given but lack of the actual location site. The reason we did not use the pixel size is that it is likely in a patch only a certain percentage were tumor; therefore, the raw pixel size may be biased in determining the actual site characteristics.

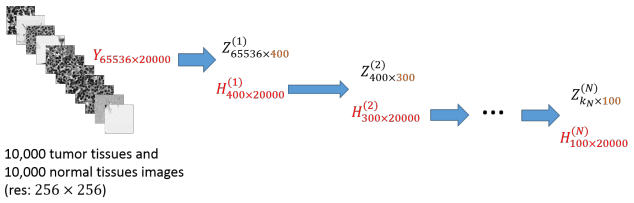

Fig. 3. Illustration of the deep semi non-negative matrix factorization approach.

3. DISCUSSION

Our work was based on a deep learning based semi non-negative matrix factorization, this approach adaptively project the images into lower dimension space to reduce the noise and achieve better discriminate results.

In traditional image analysis, one always start with certain alignment (registration) of the images to a common template. But unfortunately, nowadays a lot of images could not be registered either due to the large variability/noise or lack of well-defined registration criterion. The later case is especially popular in cancer image analysis, since tumor happens every where in the body, and the shape and texture are not exactly alignable. Thus, the alignment of such images involving tumor textures are not well defined. Due to this fact, most of traditional approaches do not work or do not work well.

The deep learning based method has enabled the analysis of such images that can not be aligned by extracting low level features. In the DSNMF frame work, the registration was done implicitly and internally in one of the layers, and the registration was a quite low level feature matching. And it turns out this kind of "registration" is quite robust to tumor included images.

The projected features sit in a much lower dimension space (100 compare to 256×256), and the discriminative ability are much higher than the original images. A random split evaluation showed that the accuracy of predicting tumor of normal patches reaches 0.88 and the AUC goes up to 0.94.

Although it seems promising, there are still some issues with this approach. One is that the layers and hidden states of DSNMF is hard to determine, and it is not possible to access the best factorization in any term. But the local optimal obtained in our approach still yields reasonable results.

In this challenge, we worked on these much smaller patches of the whole image in stead of the original image.

When splitting, it is likely we lost some information; thus, it would be ideal to have some overlapped split. The ultimate approach would be using the whole image, but our resource are limited to such split and conquer approach.

4. REFERENCES

[1] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjoern Schuller, "A deep semi-nmf model for learning hidden representations," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, Tony Jebara and Eric P. Xing, Eds. 2014, pp. 1692–1700, JMLR Workshop and Conference Proceedings.

[2] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016.