Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features

Divya Spoorthy, ES15BTECH11001

December 19, 2019

1 Introduction

Human pathologists have been doing the pathology diagnosis for a long time, by observing the stained specimen on the slide glass using a microscope. Recently, WSIs (Whole Slide Image) were being scanned and captured as digital image so we can automate the process of analysis of WSIs.

2 Histopathology Image Sources

HE stained histopathology whole-slide images of lung adinocarcinoma and squamous cell carcinoma were obtained from TCGA webiste. Total 2186 images which included samples from 515 lung adenocarcinoma patients and 502 lung squamous cell carcinoma patients. Magnification is fixed to x 40. WSIs were tiled to 1000 x 1000 pixels, and 10 densest images among them were taken. The inputs to the classification algorithms were the quantitative features extracted from the images as described in the previous section, and the outputs were the predicted diagnoses groups.

3 Evaluation of Quantitative Features from Images

A segmentation and feature extraction pipeline was built using CellProfiler. A streaming application was designed to request and filter the required images from GDC repository and stream the large SVS file. The SVS file obtained is then pre-processed one-by-one in situ. Using Opentools, it is zoomed to 40x and the region is a whole-slide image. This images are tiled into overlapping 1000 x 1000 pixels using bftools.

To reduce computation , only the 10 densest (percentage of non-white pixels in the RGB array representation) per image series were used at tiles. Segmentation was done by Cellprofiler using the UnmixColors module, then identified the tissue foreground from unstained background by a threshold calculated by Otsu Algorithm. A segmentation and feature extraction pipeline was built using Cell Profiler. Features extracted from the batch input of TIFs -¿ export to xls dump used by machine learning models for classification.

4 Machine Learning Methods for diagnosis classification

In this report, applications of digital pathological image analysis using machine learning algorithms are detailed. Typical steps of histopathological image analysis using machine learning are shown is the Figure1. Prior to applying Machine Learning algorithms some pre-processing should be done, which was mentioned in the section3. Machine Learning algorithms are divided into supervised and unsupervised learning. Here we use supervised learning algorithms since we have the labels for our data. The following algorithms were implemented.

1. Naive Bayes Classifiers
2. SVM with Gaussian, Linear and Polynomial kernels
3. Bagging
4. Random Forest

The dataset was randomly partitioned into 75 percent training set and 30 percent test set. Python with Sklearn package was used to implement the above mentioned algorithms. The models were built and selected the features using data only from the training set, in order to rigorously evaluate the performance of the finalized models with the untouched test set. Classification model was designed to classify the images of two different type of lung cancer.

1. Lung Adenocarcinoma
2. Lung Squamous Cell Carcinoma

The inputs to the classification algorithms were the quantitative features extracted from the images as described in the previous section, and the outputs were the predicted diagnoses groups.

5 Evaluation and Results

To evaluate the performance of the classifier, AUCs were calculated.

Classifier	AUC score
Bagging	0.76
Naive Bayes	0.61
Random Forest	0.72
SVMs with gaussian	0.73
SVMs with linear	0.68
SVMs with polynomial	0.70

References

[1] Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features Kun-Hsing Yu, Ce Zhang, Gerald J. Berry, Russ B. Altman, Christopher R, Daniel L. Rubin Michael Snyder Nature Communications volume 7, Article number: 12474 (2016)xtit, (Academic Press, New York, 2003).

[2] Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data David A Gutman,Jake Cobb, Dhananjaya Somanna, Yuna Park, Fusheng Wang,1 Tahsin Kurc, Joel H Saltz, Daniel J Brat, and Lee A D Cooper.

[3] Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care Ugljesa Djuric, Gelareh Zadeh, Kenneth Aldape Phedias Diamandis npj Precision Oncologyvolume 1, Article number: 22 (2017)