Final Project for MS-327 Signal Processing Based on Biomedical Statistics

Mingzhen Tian

518111910021

Zhiyuan College, Shanghai Jiao Tong University

1 Introduction

Lung adenocarcinoma is the most common type of lung cancer, comprising around 40% of all lung cancer cases. In spite of achievements in understanding the pathogenesis of this disease and the development of new approaches in its treatment, unfortunately, lung ADC is still one of the most aggressive and rapidly fatal tumor types with overall survival less than 5 years.[1] It is very important to diagnose as early as possible.

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. LUAD is lung adenocarcinoma (LUAD) of the Cancer Genome Atlas (TCGA) data set including clinical information, pathology image and omics data. The LUAD dataset contains clinical information on 522 individuals, 1067 pathology images, LUAD mutation and tumor expression matrix.

In this report, we built different models for prediction of mortality and compared their performance. We used Pytorch to build a neural network for classification of pathological images, used R to perform logistic regression, random forest, and survival analyses to predict clinical data. Finally, combining these models, we got an accuracy rate of about 80% and analyzed the results.

2 Classification of pathological images using neural network

We used Pytorch package to build a neural network for classification and prediction of pathological images. The model is ResNet-18 including 17 convolutional layers and 1 FC layer, which is simple and practical. The epoch is 200 and the final accuracy was around 80%. The details of the model are as follows.

2.1 Preprocessing of medical images

LUAD dateset contains histopathological images of the patient's lungs, stained by the H&E method. Ther are stored in SVS format, which is a common way to store pathological images. SVS format pathological images are mostly at 100,000 resolution, too large to be used directly, and common programs cannot recognize this format, so they need to be converted to common format and reduced in size.

In this project, we used an open source C library OpenSlide with a Python interface openslide-python to section the pathological images. Openslide was used to open the SVS file and slice the picture. The second layer of the slice, which is of appropriate size, was adopted and converted it into PNG format using numpy and PIL. Figure1 is an example of converted pathological images.

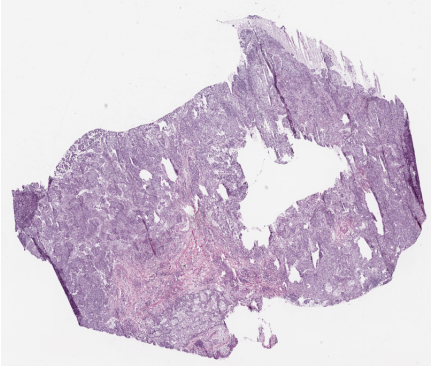

Figure 1: example of converted pathological images

Later, due to the clutter of file storage location, we used OS package to traverse the SVS file, converted the above format and stored in a unified location. Then, we divided the images into test sets and training sets according to the file *sample_index(train&test).xlsx* and stored them in two folders respectively.

2.2 Data loading and augmentation

Thanks to the TorchVision package, data loads easily in PyTorch. We started by creating *MyDataset*, which used OS to import the images and labels in it. We created the dictionary based on two columns *bcr_patient_barcode* and *vital_status* in the file *luad.csv*. We compared the file name of the images with the dictionary to get the survival state, and stipulated that *Alive* is 1 and *Dead* is 0.

Torchvision.transforms is an image preprocessing package in Pytorch that contains a variety of functions that transform the image data, which is necessary in the image data reading step. So we used some transformation functions in *Torchvision.transforms* to augment the image.

Firstly, since the left and right lungs may be different, we used the *transforms.RandomHorizontalFlip* to flip the images horizontally at random. The probability of flipping is 50%. In the second step, in

order to reduce the computation of training, we used *transforms.RandomCrop* to randomly crop the image to 224×224. Compared to the normal cells, lung adenocarcinoma cells are larger, with larger nuclei, higher nucleolar ratios, and marked eosinophilic nucleoli. Thus, LUAD can be identified based on local cell morphology, not overall characteristics, which is the basis for the rationality of random cropping. Finally, use *transforms.ToTensor* to convert the PIL format to tensor.

2.3 Model construction and training

The model used in this project is ResNet 18 which is simple and practical. The ResNet 18 model was constructed using BasicBlock as basic unit, which involves two 3 by 3 convolution operations, two BatchNorm2d operations and a ReLU activation function. The ResNet 18 model has 4 convolution layers each layer containing 2 BasicBlocks, 1 pooling layer and 1 FC layer. The in_channel is 3 because of the RGB format.

After the model was built, we began to train it. The total epoch was 200. The optimizer used was Adam with decaying learning rate. The initial learning rate was 0.1, with a StepLR attenuation of 80% per 10 epochs. Figure2 is the attenuation curve. The loss function used was *torch.nn.functional.cross_entropy* to determine how close the actual output is to the expected output. In order to improve the efficiency of GPU and reduce time, we adopt checkpoint mechanism. Model parameters are stored every 50 epochs, and the model parameters can continue to be used for subsequent training.

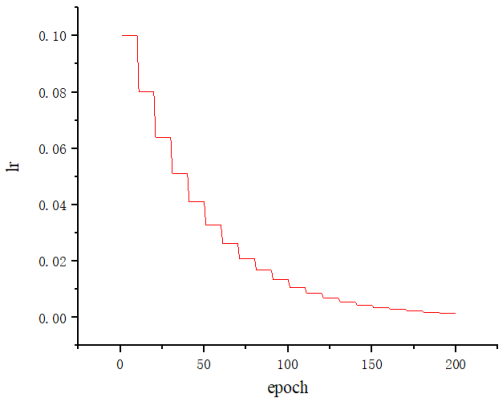

Figure 2: StepLR attenuation curve

2.4 Results and discussion

The results of the test set and training set are as Figure3. The accuracy of test set converges to 79.838%, the accuracy of the training set converges to 70.935%, the average loss of the training set converges to 0.594. It can be seen from the results that the training effect of the model is good.

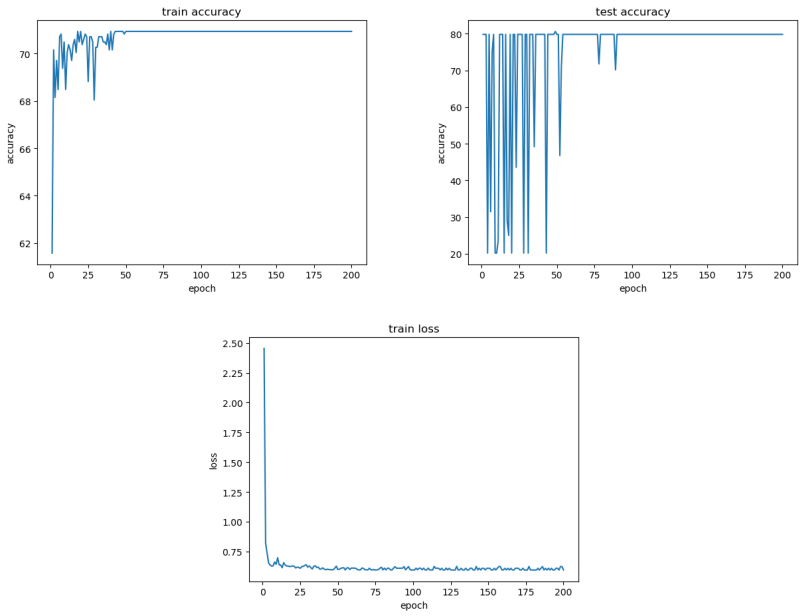

Figure 3: Results of training

The deficiency of the model is that it doesn't recognize images with a lot of white space as Figure4. In the process of random cropping, there is a high probability of clipping to a blank image, which leads to a large jump in training accuracy. So the direction of model improvement is to increase the ability to identify areas with color range.

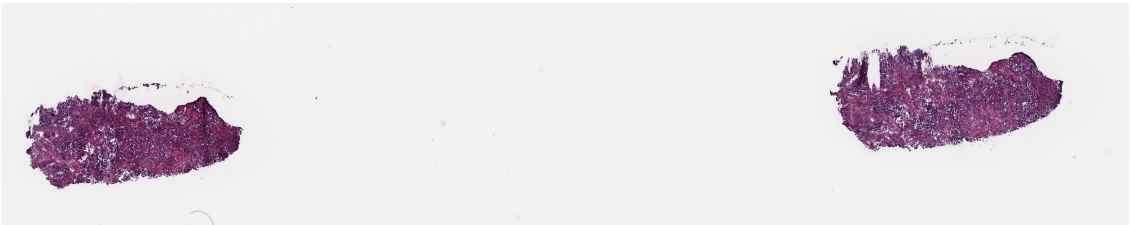

Figure 4: image with a lot of white space

3 Prediction of mortality using clinical data

In LUAD dataset, clinic information is an important part, including 123 attributes of 522 individuals. This project uses the following important characteristics for prediction: vital status, race, gender, age at initial pathologic diagnosis, days to last followup, days to death, tobacco smoking history, number pack years smoked, dlco predictive percent, pathologic stage and icd 10(classification of diseases). Wherein, tobacco smoking history is a categorical variable, divided into five categories(1-5) based on smoking history. Number pack years smoked is a continuous variable defined as the number of cigarettes smoked per day divided by 20 and then multiplied by the number of years smoked.

In this project, we constructed several models including logistic regression, random forest, and Cox proportional hazards model to predict mortality. When we compared the results of different models, the best accuracy was over 75%.

3.1 Data cleaning and preprocessing

Because the raw data is often incomplete, noisy, and inconsistent, we need to clean the raw data. In raw data, there are several names meaning it is not available, such as Not Evaluated, Unknown, Not Available and NA, so we replaced them all with Na. After that, the data set was then divided into training sets and test sets using Python. In addition, the raw data contains more than a hundred indicators, many of which are useless, so we picked only 10 important of these variables.

In the raw data, many variables have missing values, so we will deal with these missing values next. Because there are too many missing values, if deleted, the data will be greatly affected, so we adopt the interpolation method to deal with the missing values. Since most of the variables are discrete variables, we use the mode of this attribute to interpolate the missing value. For a dead individual, *death_days_to* is the survival time. For a alive individual, we assume *last_contact_days_to* is the survival time, and right censored. In addition, we also convert the data types of categorical variables into factors.

3.2 Logistic regression model

Logistic regression analysis, a generalized linear regression analysis model, is often used in data mining, automatic disease diagnosis, economic forecasting and other fields. In this project, we used e1071 R package to perform logistic regression. In this model, the variables *age_at_initial_pathologic_diagnosis*(p=0.0383) and *tobacco_smoking_history_indicator4*(p=0.0920) are significant. Figure5 is the ROC curve of this model using pROC R package. The final accuracy is 72.27%.

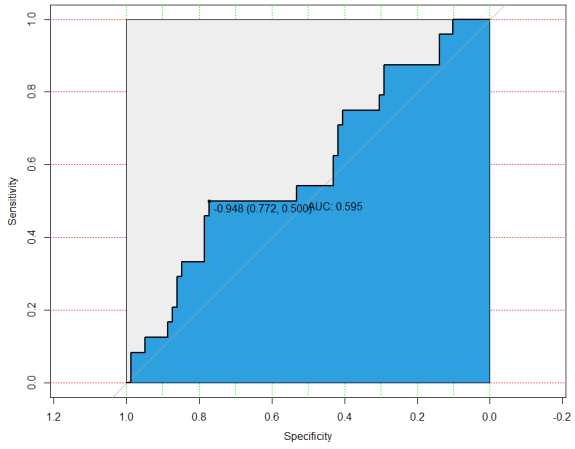

Figure 5: ROC curve

3.3 Random forest model

In machine learning, a random forest is a classifier containing multiple decision trees. In this project, we used e1071 R package to perform random forest model with ntrees=500. The final accuracy is 75.50%. Then we compared the two classification models. According to the results, we can see the random forest model is better.

3.4 Cox proportional hazards model

The cox proportional hazards model is a semi-parametric regression model. Taking survival outcome and survival time as dependent variables, this model can simultaneously analyze the influence of many factors on survival time, and can analyze data with censoring. In this project, we use survival R package and survminer R package to perform Cox regression. Figure6 is the hazards ratio. According to the results we can see that *tobacco_smoking_history_indicator5* and *ajcc_pathologic_tumor_stageStage II* have higher risk of death.

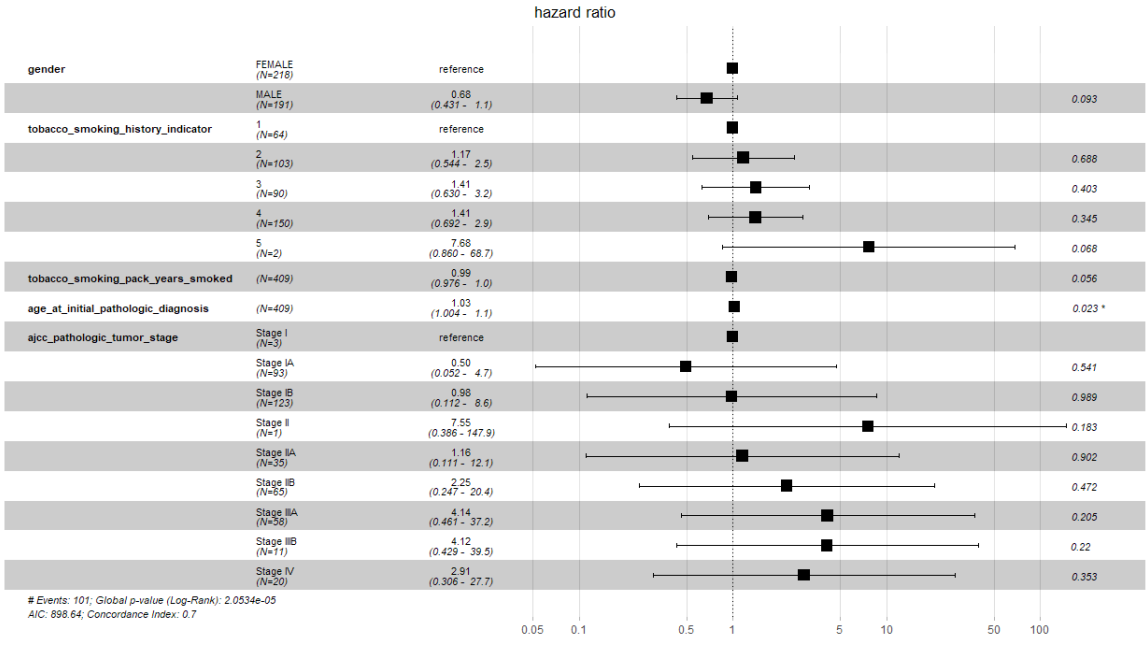

Figure 6: hazards ratio

3.5 Discussion of clinical prediction model

Due to the absence of many clinical data and the simplicity of processing the missing values, the accuracy of the model is relatively low. We can use the machine learning interpolation method to deal with missing

values, such as k-Nearest Neighbor, decision tree to improve the accuracy of the model.

4 Conclusion

In this project, we analyzed cancer pathological images and clinical data using neural network algorithms, logistic regression, random forest, and survival analyses to construct a prediction model for death with an accuracy rate of about 80%. Through the combination of several models, the accuracy of prediction has reached a relatively good level. However, there are still many defects in the model, especially in the prediction model of clinical data. The next direction is to increase the avoidance of the white space in the image and to process the clinical data more precisely.

References

1. Denisenko, T. V., Budkevich, I. N., & Zhivotovsky, B. (2018). Cell death-based treatment of lung adenocarcinoma. *Cell death & disease*, 9(2), 117. https://doi.org/10.1038/s41419-017-0063-y

Acknowledgement

In the process of finishing my project, the teachers and teaching assistants gave me a lot of help. I would like to thank them for their patient teaching and answering. My team members are also very enthusiastic to answer my questions, I also want to thank them.