上海交通大學

Multimodal Model for Predicting Mortality of Lung Adenocarcinoma

Yuxiang Zhang

Biomedical Science

School of Medicine

Multimodal Model for Predicting Mortality of Lung Adenocarcinoma

Yuxiang Zhang

1. Abstract

In order to predict the mortality of patients suffering lung adenocarcinoma (LUAD), we build a multimodal model, which combines the genomic data, transcriptomic data and clinical data from The Cancer Genome Atlas (TCGA) database. Using a series of statistical and machine learning methods such as Cox regression, LASSO, Random Forest and Support Vector Machine (SVM), our model obtained the accuracy of 0.83 for the task of patients' mortality prediction. These results demonstrate the power of using machine learning and big data to produce significant improvements in the prediction of survival state.

2. Methods

The data obtained from the TCGA database are divided into three parts: clinical information, pathological images and omics data. Clinical information contains some clinical attributions about patients, from which we select four interested features: age, survival status, 5-year survival time and tumor stage for prediction modeling. Then, the omics section contains genomic SNP and transcriptomic expression matrix. Pathological image part contains the pathological image of lung tumor tissue of LUAD patients and the pathological image of adjacent normal tissue. The pathological images are stored as Whole Slide Images (WSIs) format, which cannot be directly used for deep neural network training due to their high pixels.

This multimodal model we build can be divided into three sections: the clinical data part, genomic data part and transcriptomic part. The initial idea is to improve the accuracy of model prediction by combining different modal data. Therefore, we selected clinical data, which is closer to phenotype, using statistical method to construct the first part of prediction model. Then, we employ machine learning methods on genomic and transcriptomic data for the second and third part, respectively.

2.1 Clinical Data Part

The data source of clinical data is clinical features from TCGA database. There are two parts of clinical information, a series of clinical symptoms of the patient and the therapies and medications used by each patient. We selected a series of clinical features from the first part (clinical_patient_luad.txt) in TCGA database, including living status, time-to-death, forced vital capacity and pathologic stage. Using Cox regression on all these features, we can get a multi-variable linear model to generate patients’ risk score based on clinical information.

2.2 Genomic Data Part

The genomic data comes from SNP data of TCGA database. In the data matrix, each row represents the genotype of one gene in different patients, and each column represents the genotype of multiple genes in different patients. The resulting genomic data is Single Nucleotide Polymorphism (SNP), and as the result has been preprocessed, no further normalization of the dataset was performed here.

Considering that this data is very sparse and the number of features is much larger than the number of samples, we used SVM for classifying, and try different kernel functions such as linear kernel function and Gaussian kernel function in the prediction.

2.3 Transcriptomic Data Part

The data source of transcriptomics data is expression matrix of different genes from LUAD patients in TCGA database. In the expression matrix, each row represents the expression level of one gene in different patients, and each column represents the expression level of multiple genes in different patients. The resulting transcriptomic data is Fragments Per Kilobase of exon model per Million mapped fragments (FPKM), and as the result has been preprocessed, no further normalization of the dataset was performed here.

We combine the living status in clinical data with the FPKM data to form an expression matrix containing the patients' living status. Then we use the traditional machine learning method Least absolute shrinkage and selection operator (LASSO) to analyze the expression matrix, selecting the appropriate parameters, and find important genes of interest. After finding the genes of interest, we used them to build a prediction function forming probability. Similarly, we also use the same data matrix to build a Random Forest model, predicting the probability of death.

2.4 Single Layer Perceptron

When the final judgment results of the three prediction models are obtained, in order to increase the accuracy of the model, the output results were input into a single-layer perceptron (Fig.1), and different weights are given to different models to further integrate the classification results.

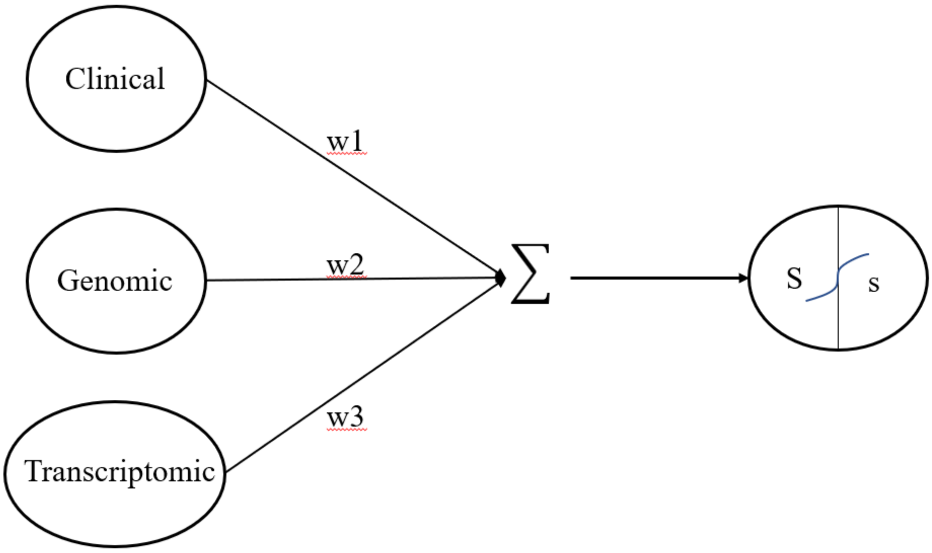

Figure 1: Single Layer Perceptron of Multimodal Model.

Overall, the flow chart of the whole model is shown in Fig.2, the final prediction is determined by combining three modal predictions according to a certain weight.

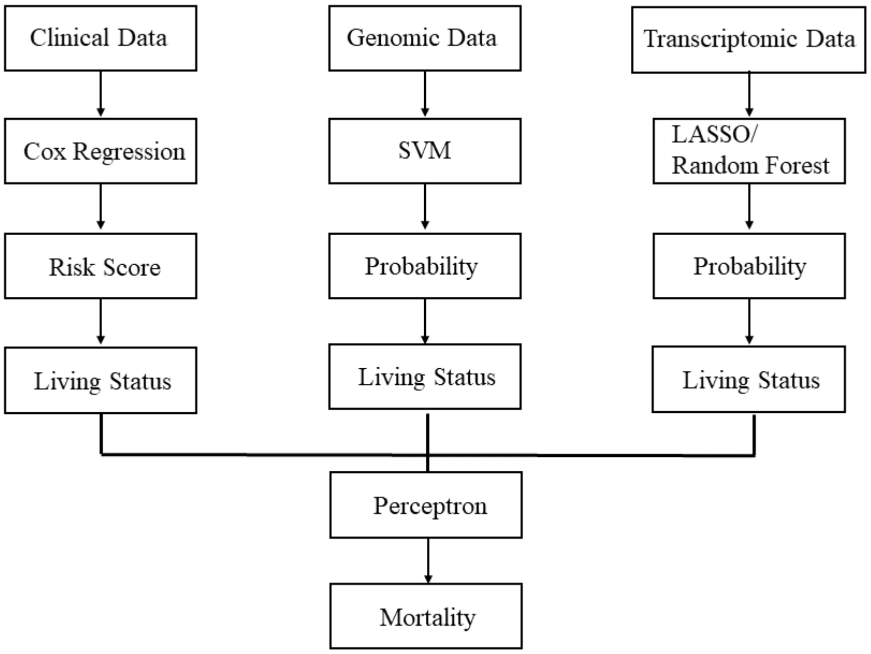

Figure 2: Flow Chart of Mortality Prediction Model.

3. Results and Discussions

In this section, we will show the results of different models and parameter settings, and discuss these results.

3.1 Clinical Model

The clinical features and the corresponding coefficients selected by Cox regression analysis and their statistic values are shown in Table 1.

Table 1: Coefficients of Cox Regression

Features	coefficients	Hazard Ratio	p-value
Stage	0.178	0.89	0.047
FVC	-0.327	0.66	0.036
Time-to-death	-0.794	0.45	0.008

FVC (Forced Vital Capacity)

The prediction function contains three significant coefficients, namely pathologic stage, forced vital capacity and time-to-death. It can be seen from the Hazard Ratio that these three features are protective factors. The concordance index of the function is 0.76 in training set, 0.73 on test set.

3.2 Genomic Model

Linear kernel SVM and Gaussian kernel SVM are compared by statistical test, we constructed 2×2 contingency table of McNemar's test. It is found that the accuracy of Gaussian kernel function is slightly higher than that of linear kernel function, but there is no statistical difference under 95% confidence interval. Therefore, SVM with Gaussian kernel function is used for subsequent analysis of SNP data. The accuracy is 0.73 and AUC is 0.79. (Fig.3)

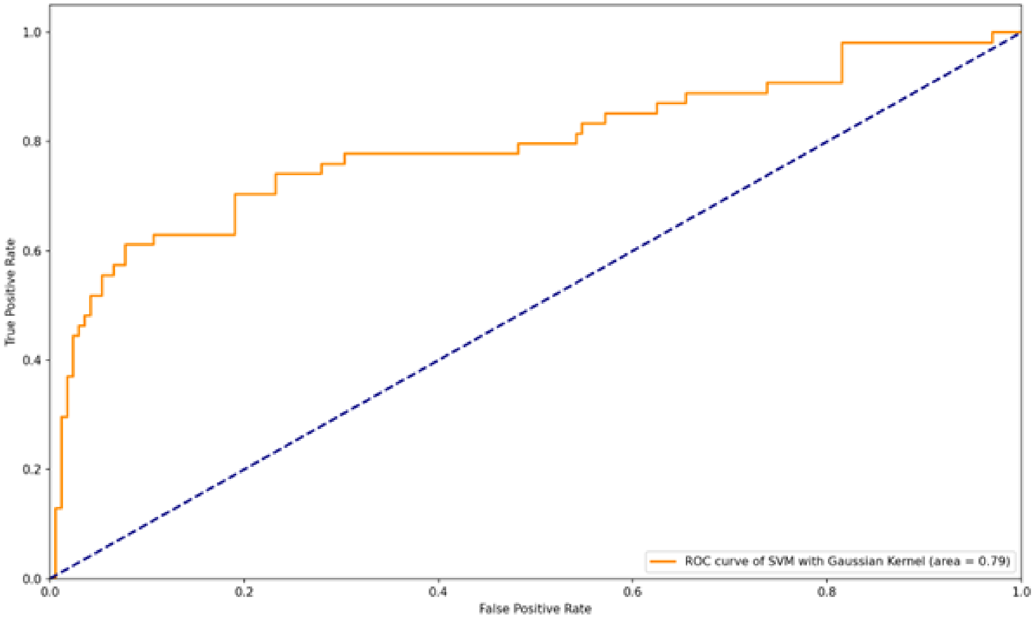

Figure 3: The ROC Curve of SVM with Gaussian Kernel.

3.3 Transcriptomic Model

The genes obtained by lasso and random forest were used to establish models, respectively (Fig.4). It was found that the random forest had higher accuracy (0.77) and AUC (0.83).

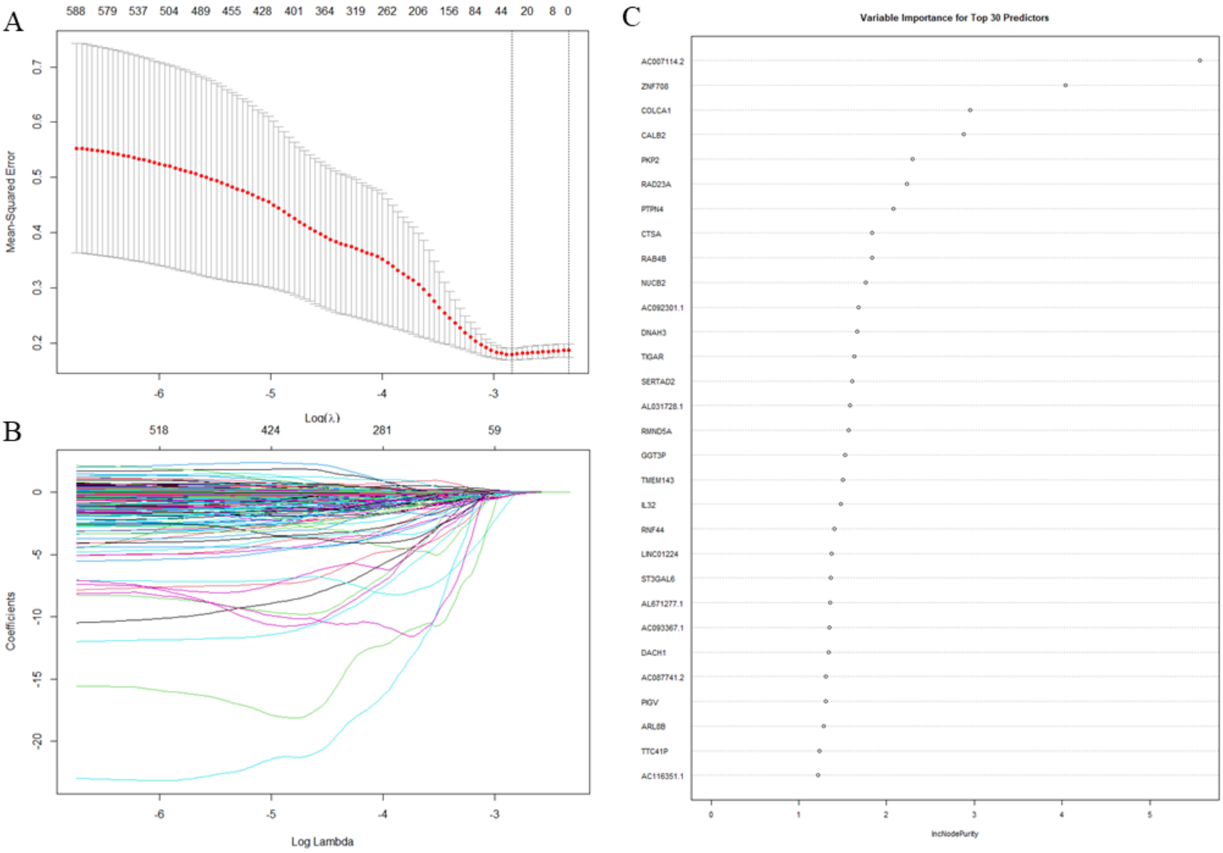

Figure 4: Selection results of LASSO and Random Forest.

(A) The logarithm of the selected number and the mean square error of the corresponding model.

(B) Coefficients corresponding to different genes in different number of λ.

(C) Genes Selected by Random Forest.

The selected genes was listed in Table 2.

Table 2: 30 Genes Selected by Random Forest.

Genes of Interest				
RGS20	HCFC2	LYAR	AL133227.1	FAM207A
MT2A	AL645634.1	RPL38P1	LDLRAD3	RAB21
AC004817.3	LNCAROD	GUSBP12	PAFAH1B2	DKK1
VAX1	KRT18P19	AC090541.1	SMNDC1	AC116917.1
AC004466.2	ENSG00000273689.1	AP000311.1	CTSL	AC013828.1
LINC02649	EIF3I	AC025575.1	APO00721.1	AC024075.3

3.4 Parameter Settings of Single-layer Perceptron

Since the transcriptomic model has high accuracy and AUC, the weight of the transcriptome model is increased to 0.7, and the parameters of the genome and clinical data are set to 0.4. When the weight sum is greater than 1, the patient is judged to be alive; while less than or equal to 1, the patient is judged to be dead. The final model gain accuracy of 0.83, which is slightly better than three single models.

3.5 Discussions

We find that better performance can be achieved by combining different data and machine learning methods. By calculating the recall rate, it is found that the comprehensive three models are also stronger than the single model, which is also of great significance in practical application and has a certain reduction in misdiagnosis.

However, the data used are not the closest to the phenotype. If the data of proteomic or metabolic omic data can be accessed for analysis, the obtained model should be more meaningful. Also, combining pathological pictures, which are the closest to the real situation, with omics and clinical data to increase the model's modal by using deep learning to process pathological sections reasonably, it could be predicted that the model may have higher accuracy.

4. Experimental settings

All the data comes from TCGA database, which includes different modal data of LUAD patients. The patient ID of test set were indexed in the provided files, and we use the rest of the patients for training. We use the entire training dataset to train the model. Noticed that that the features of some samples in the data set are missing, we preprocess the data by deleting the sample with feature missing and filling with the mean value of this type of data. In order to improve the training efficiency of model, we set the maximum number of iterations in Cox regression and Random Forest, the model will stop training when the parameters change to a certain extent. After building three prediction models of different modal data, we combine the results by using perceptron. The environment we use is R 4.2.0, the packages used in model construction are: glmnet, randomForest, Hmisc, survival, My.stepwise and e1071 (A package for SVM).