Prediction and Visualization Analysis of Lung Cancer Risk by Machine Learning

Ziqi Wan *

Guangdong Experimental High School, Guangdong, China

* Corresponding author email: JikyWan2006@163.com

Abstract. One of the most lethal cancer types in the world is lung cancer. It gravely threats human health and causes more than one million deaths annually. Current research demonstrate that an early and accurate determination of lung cancer is the key element in reducing the mortality rate, which motivates us to predict cancer at an early stage by exploiting the capacity of machine learning. In this work, both qualitative visualization analysis and quantitative experiments are conducted to reveal the relevant factors and the effectiveness of lung cancer risk prediction. Visual analyses are implemented to display the distribution and correlation between characteristics of patients and lung cancer risk. The visualizations demonstrate that smoking has a relatively large impact on lung cancer. Moreover, various types of machine learning algorithms are implemented and compared, including the logistic regression, K-nearest neighbors, random forest, and decision tree. These methods achieve satisfactory prediction results, with an accuracy of about 84% to 90%.

Keywords: Machine Learning; Lung Cancer; Visualization; Cancer Prediction; Random Forest.

1. Introduction

One of the most lethal malignancies to human health and life is lung cancer. Based on information provided by the World Health Organization's international agency for research on cancer (IARC), lung cancer has become the malignant tumor with the second incidence rate and the first mortality among all malignant tumors [1]. According to the latest data of the National Cancer Center's 2019 annual report of cancer registration in China, the incidence rate of lung cancer in China is 60.40 among a hundred thousand people, and the standardized rate of the world population is 36.02 among a hundred thousand people; the mortality was 48.42 among a hundred thousand people. The world population standardization rate is 27.87 among a hundred thousand people [2, 3].

The early stage of lung cancer lacks salient symptoms, so it is difficult to observe. When patients are found to have lung cancer, most of them are in the middle and late stages of lung cancer, which means that they miss the best period of surgical treatment. At the same time, cancer cells will transfer to other parts of the human body, which makes treatment more difficult. That's the reason why the 5-year survival rate significantly decreases, which brings significant disease burden and economic pressure to patients and their families [4, 5, 6]. Therefore, early and accurate detection and treatment of patients with potential risks can help them better cure cancer.

According to the data statistics, if the disease can be found in the early stage and the operation can be carried out in time, Patients with stage 1A lung cancer had a 5-year survival rate of more than 80%, compared to just 5% for patients with intermediate and advanced stages. [7].

At present, tertiary prevention of cancer has become an important measure of prevention and control. Among them, the primary prevention is mainly aimed at etiological prevention, such as injecting the HPV vaccine and changing living habits [8]. Secondary prevention is mainly early detection, early diagnosis, and early treatment, mainly through screening and other means [9].

With the advancement of information technology and the growth of big data in medicine, computer-aided data analysis and modeling have become the focus of current medical diagnosis and treatment technology [10]. In this study, a dataset with 15 different symptoms related to lung cancer was accessed from the Internet, and the data were learned and predicted by machine learning to explore the relationship between the incidence rate of lung cancer and the symptoms of patients, and to find out which features have a greater impact on the diagnosis of lung cancer.

2. Method

2.1 Dataset Description

The dataset used in this paper is downloaded from Kaggle. It records a total of 309 patients, and a total of 16 symptoms of patients. And Table 1 shows the data attribution.

Table 1. Data structure Description

No.	Attribute	Description
1	Yellow fingers	"1" if they have yellow fingers and "0" if they don't.
2	Wheezing	"1" if they wheeze and "0" if they don't.
3	Swallowing difficulty	"1" if they have problem swallowing and "0" if they don't.
4	Smoking	"1" if they smoke and "0" if they don't.
5	Shortness of breath	"1" if they experience shortness of breathing and "0" if they don't.
6	Peer pressure	"1" if they are influenced by peer pressure and "0" if they're not.
7	Lung cancer	"YES" if they are positive with lung cancer and "NO" if they're not. This is the column that the author is trying to predict. It is called the target.
8	Gender	The gender of the subjects, M or F
9	Fatigue	"1" if they're easily fatigued and "0" if they're not.
10	Coughing	"1" if they cough regularly and "0" if they don't.
11	Chronic disease	"1" if they have a chronic disease and "0" if they're don't.
12	Chest pain	"1" if they have chest pain and "0" if they don't.
13	Anxiety	"1" if they are suffering from anxiety and "0" if they're not.
14	Allergy	"1" if they have an allergy and "0" if they don't.
15	Alcohol consuming	"1" if they consume alcohol and "0" if they don't.
16	Age	The age of the subjects

The dataset is analyzed through 5 parts: Prevalence, non-prevalence, male, female, and the proportion between these characteristics. Figure 1 shows the results using two pie charts. Finally, in the 309 samples, it is found that 39 people did not suffer from lung cancer; 12.62% did not suffer from lung cancer; 270 people suffered from lung cancer; 87.38% from lung cancer; 47.57% from women, and 52.43% from men. At the same time, some distribution graphs of the number of people with cancer corresponding to these 15 characteristics and the graphs of cancer with age are also plotted. In this way, the general overview of the database can be shown in the form of charts.

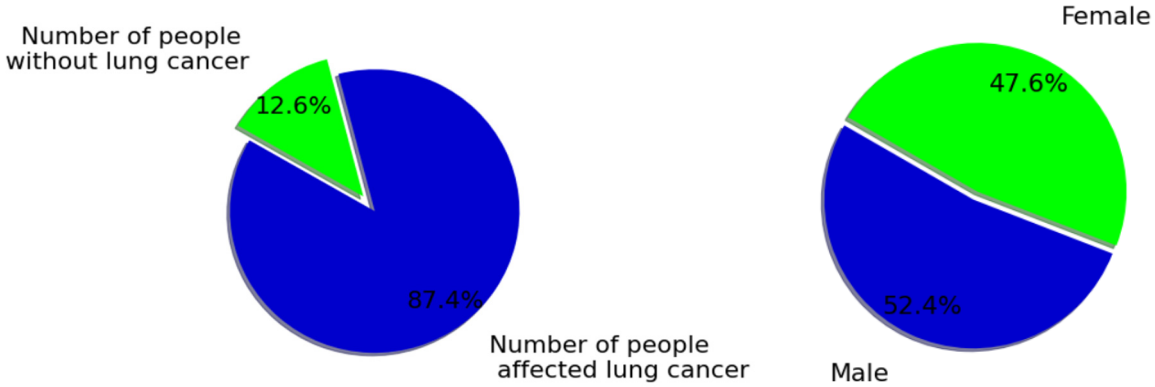


Fig 1. The general overview of the database

The charts reveal the proportion of male patients and female patients is nearly the same, and the proportion of patients who are affected by lung cancer is significant.

2.2 Feature Correlation

After preprocessing the data, an in-depth exploration of data is conducted. The first task is to explore whether all of these 15 symptoms are related to lung cancer. The Chi-squared test is used for feature selection here. It is a nonparametric test. Relatively speaking, nonparametric testing has relatively loose requirements on data distribution, and does not require too much data. The Chi-squared test is a hypothesis test method for counting data, which mainly compares the coincidence degree of theoretical frequency and actual frequency. It is often used for feature selection. For example, in this test, it is necessary to check whether there is any difference between age and whether there is lung cancer. If there is any difference, it indicates that age is related to whether there is lung cancer. In the subsequent analysis, it is necessary to put the classification variable age into the model training. Generally speaking, the basic data of the Chi-squared test includes R rows and C columns. Therefore, it is generally called RC contingency table, or RC table for short. Through the frequency table, when the data are classified by two or more attributes (qualitative variables), it can be determined whether the correlation between the two features is significant by calculating the p value of each chi square test. If the p-value exceeds 5%, the correlation between the two characteristics is not significant. After this series of tests, it was found that these 15 characteristics were all related to lung cancer. Next, the differences will be compared between these features.

2.3 Feature Visualization Analysis

2.3.1 The Influence of Single Independent Variable to Lung Cancer

The second thing is to calculate the correlation between these characteristics. Two methods are used to do the calculation. The first method is to use the function “data.corr” for calculation. At the same time, a thermal map is plotted to show the correlation between various characteristics. According to the calculation, these five symptoms are highly positively correlated with cancer. They are Allergy: 0.33, Alcohol Consuming: 0.29, Swallowing Difficulty: 0.26, Wheezing: 0.25, Coughing: 0.25.

This is the first method that is used to calculate the correlation between 15 symptoms of lung cancer data; in Figure 2, negative numbers represent the symptom is negatively related to lung cancer. The larger the number is in the map, the higher correlation the symptom is with lung cancer.

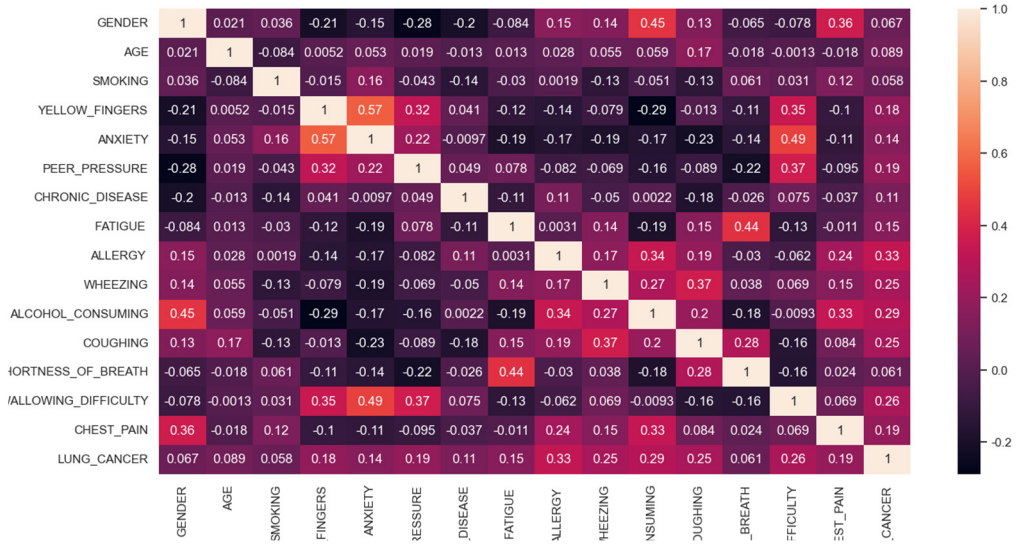


Fig 2. Correlation the symptom with lung cancer

Secondly, Lasso Regression is used to determine the probability of the characteristics that lead to cancer. The advantage of Lasso Regression is that it can make up for the shortage of Least Square Estimation and Local Optimal Estimation of Stepwise Regression, and it can select features well and

effectively solve the problem of multicollinearity among features. For the selection of alpha parameter in Lasso Regression, Grid Search and Cross-Validation can help select the parameter in an ordinary way. In simple terms, Grid search is a method of adjusting parameters, and it can be used to select appropriate hyperparameters. It uses a very simple and violent technique: Exhaustive Search. Of course, the candidate parameters should be given artificially. The Cross-Validation is the CV in GridSearchCV, which refers to cross validation. Cross validation is to further divide the training data into training sets and validation sets, and use the validation sets to adjust the model's hyperparameters and to preliminarily evaluate the model's ability. A 5-fold cross validation is used in the selection of the hyperparameter. The principle is to divide the original training set into k groups, and let each subset be a validation set. The remaining k-1 subassemblies are used as drive as training sets. In this way, K models will be obtained, and the performance index of the classifier under k-cv will be determined by the average of the classification accuracy of the final verification set of K models. K is generally 2 or more. In real operation, it is usually taken to 3. It is not until the data volume of the initial dataset is small that it will attempt to take 2. Through these two methods, the probabilities between these 15 characteristics and cancer are successfully ranked. Their probabilities from high to low are 'alcohol consulting', 'allegory', 'fatigue', 'swallowing difference', 'yellow fingers', 'coughing', 'peer pressure', 'chronic disease', 'wheezing', 'anxiety', 'smoking', 'chest pain', 'age', 'shortness of breathing', and 'gender'.

The results calculated by the two methods are very similar, which also proves the reliability of the two methods and the accuracy of the results calculated by the two methods.

2.3.2 The Influence of Muti-independent-variables to Lung Cancer

Based on the correlation calculation above, several features with high-positive correlation with cancer are obtained. It is essential to analyze the top three features from multiple angles. The method of multivariate analysis of variance is employed to analyze these characteristics. In this experiment, 3 highest ranked symptoms which are arranged by the two methods above are selected as independent variables for the Multivariate analysis of variance. Then each of the symptoms will be combined with the other ones to compare with lung cancer. Results demonstrate that all the combinations have remarkable correlation with being affected lung cancer. (P-value of all combinations are less than 0.05)

2.4 Model Prediction with Four Algorithms

4 algorithms are selected to predict the lung cancer risk using the aforementioned dataset. They are logistic regression, KNN, random forest, and decision tree. The dataset is divided into 4 parts, which are x-train, y-train, X-test and y-test. Then, 80% of all data are divided into training sets and 20% into data sets. The 15 feature items are divided into x parts, and whether the patients' affected cancer is divided into y.

2.4.1 Logistic Regression

The first model is logistic regression. Logistic regression is a generalized linear regression analysis model, which mainly solves the dichotomous problem and is used to express the possibility of something happening. It is often used in data mining, auto-diagnosis of diseases, economic forecasting, and other fields. Lung cancer analysis is a case in point. Two groups of individuals were chosen; one group was lung cancer and the other group was not lung cancer. Both groups of individuals exhibit different physical characteristics and lifestyles. Thus, the dependent variable is lung cancer, and the value is either "yes" or "no". Independent variables may be age, gender, dietary habits, etc. Independent variables may be continuous or categorized. Then, thanks to logistic regression analysis, it is easy to obtain the weight of independent variables, so that the factors that are the risk factors for lung cancer can be broadly understood. At the same time, depending on the weight, the likelihood of an individual with cancer can be predicted based on risk factors.

2.4.2 K Nearest Neighbor (KNN)

The second approach is based on the KNN model. The fundamental principle of KNN is that the target sample is similar to the K samples of the dataset. In other words, when determining how to make a classification decision, if the KNN model is used, the intelligence determines which class the target sample belongs to according to the category of one or more samples nearest to the sample. The benefits of KNN are numerous. This method is simple and effective, with a low cost of collection and a low complexity of the algorithm. Automatic classification of cross samples and large samples of domain classes is required.

2.4.3 Random Forest

The third model is random forest. The random forest is made up of numerous decision trees, and there is no correlation between the various decision trees. When a new sample is input into the classification task, each decision tree in the forest will judge and classify separately, and the classification results of each decision tree will be obtained. The randomized forest considers the part with more similarity in the decision tree classification result as the end result.

2.4.4 Decision Tree

The last of these is the decision tree. The decision tree algorithm employs the principle of probability theory and the graphic method of probability analysis. Furthermore, this is a typical classification methodology. A method for analyzing decisions that is often used to assess risks and assess their feasibility. By forming a decision tree, the data is processed, and the inductive algorithm is used to generate readable rules and decision trees, and then the decision is used to analyze the new data. In essence, the decision tree is a data classification process based on a set of rules.

These 4 algorithms are used to train models of whether a patient will affect cancer with these symptoms, and the model is evaluated by the system.

3. Results

This section mainly shows the results of this study. 3.1 mainly shows the correlation between 15 symptoms and lung cancer and their feature selection. 3.2 mainly shows the distribution of the relationship between different characteristics and lung cancer. 3.3 analysis of the relationship between univariate and multivariate and lung cancer. 3.4 mainly displays the prediction results of the model.

3.1 Correlation Analysis

Table 2. 15 symptoms' P-value

No.	Attribute	P-value
1	GENDER	0.003
2	FATIGUE	0.0029
3	COUGHING	0.0029
4	ALCOHOL_CONSUMING	0.0028
5	AGE	0.0024
6	SMOKING	0.0024
7	ANXIETY	0.0012
8	ALLERGY	0.0012
9	SHORTNESS OF BREATH	0.00082
10	DISEASE	0.00079
11	CHRONIC	0.00053
12	YELLOW_FINGERS	0.00015
13	PEER_PRESSURE	0.00015
14	SWALLOWING	0.00015
15	WHEEZING	0.00013

The results of this experiment will be explained in Table 2. Firstly, the relationship between these 15 symptoms and lung cancer is calculated by Chi-squared test.

The P-value of all features is less than 0.05, which means that all the features have strong correlation between lung cancer.

Figure 3 is a bar chart which can visualize the ranking. It can be seen from the chart that the smaller the y value of each feature is, the stronger dependency it is with affecting lung cancer.

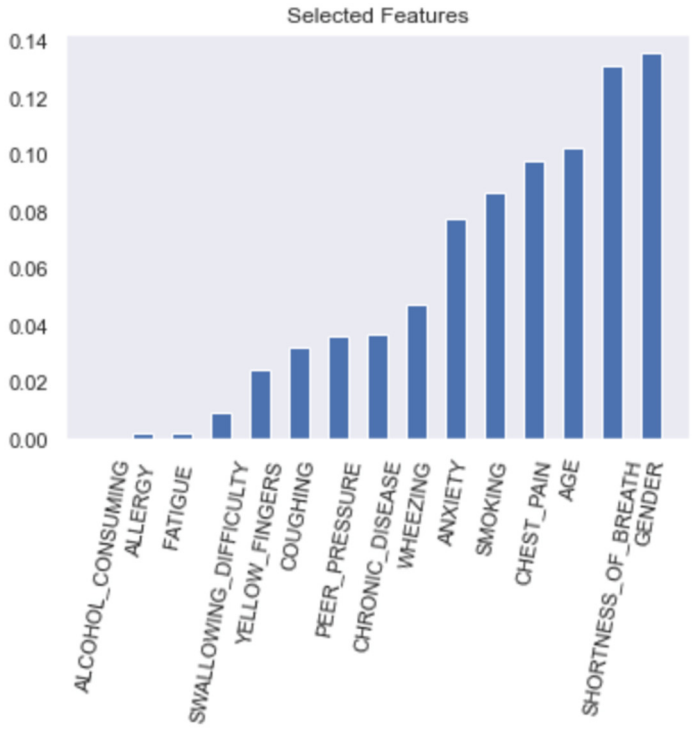


Fig 3. Selected Features

3.2 Distribution of Lung Cancer Related Features

Figure 4 displays a bar chart which reveals the distribution of lung cancer with age. In these charts, blue bars represent patients without lung cancer, and orange bars represent patients who affected lung cancer. Through the bar chart, it is easy to find out that the incident rate of patients in the age range of 56-70 is very high.

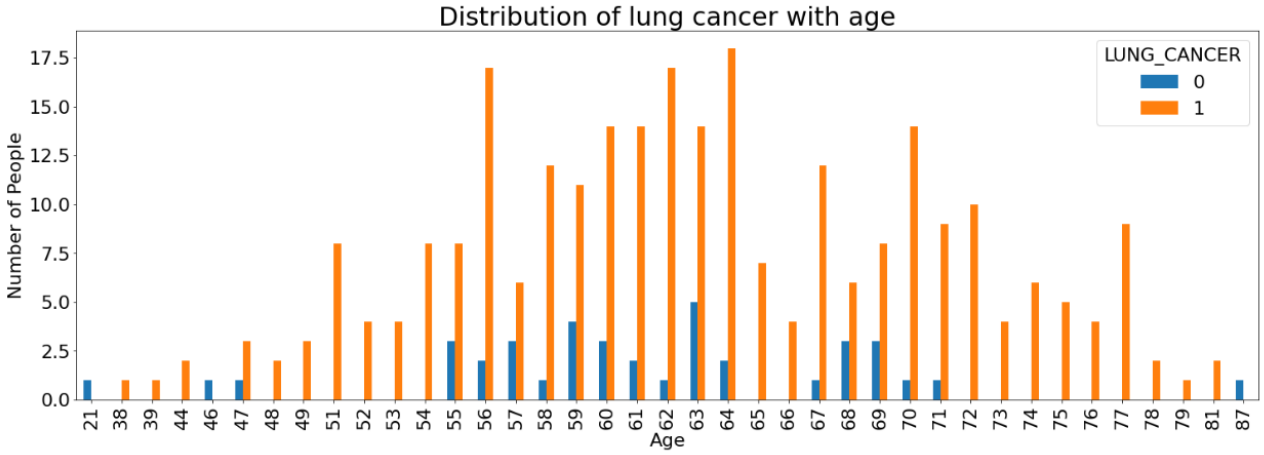


Fig 4. Distribution of Lung Cancer with Age

These bar charts in Figure 5 below are the charts that reveal the general condition of the rest of the 14 features and distribution of lung cancer. Blue bars represent patients without lung cancer, and orange bars represent patients who affected lung cancer. “1” represents male and “0” represents female. From these charts, it is simple to identify that there are some symptoms that the incidence rate of female patients is higher than male patients; such as fatigue, allergy, shortness of breath.

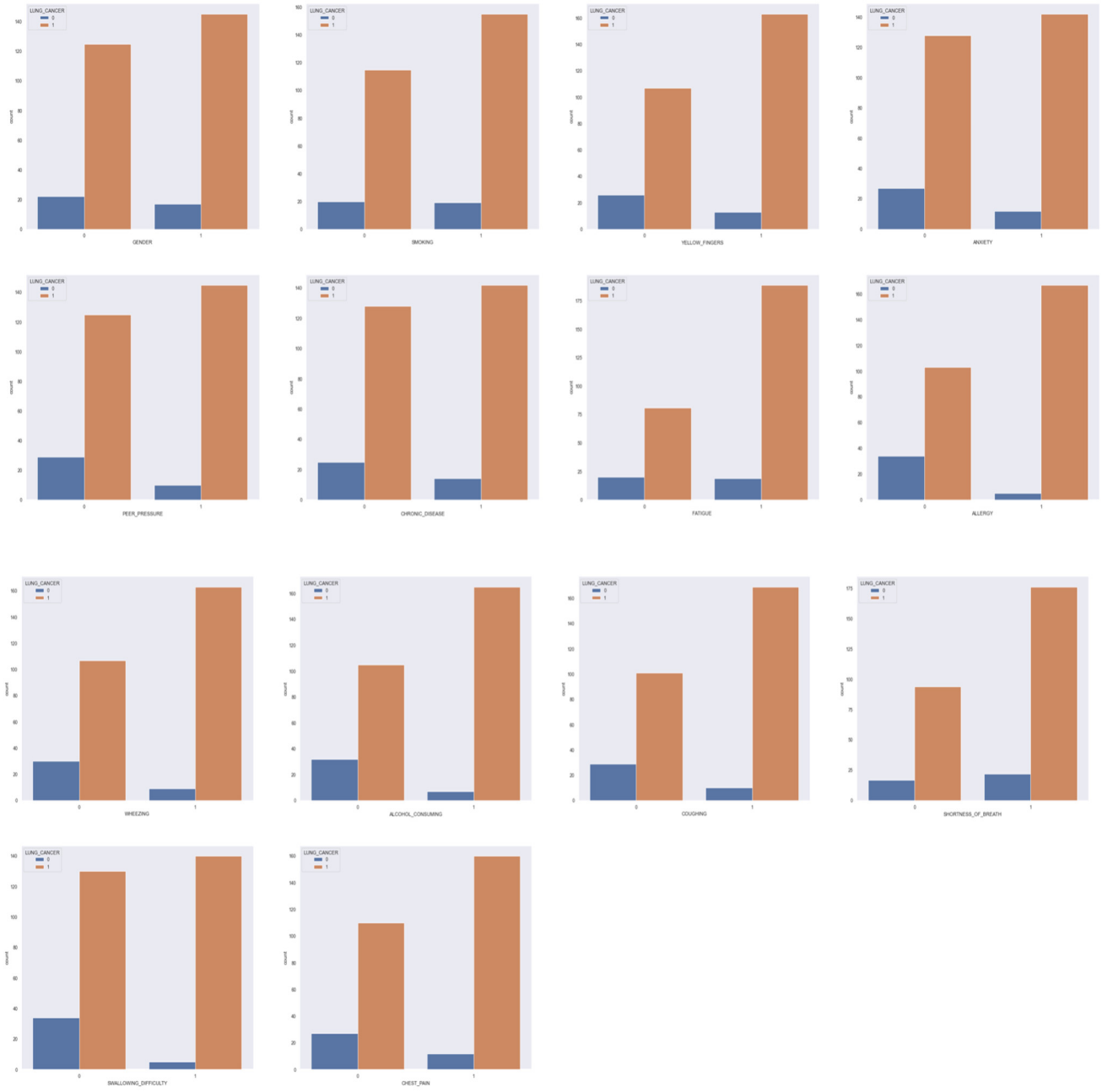


Fig 5. 14 features and distribution of lung cancer

3.3 Analysis of Variance

Based on the above correlation calculation, several features that are highly positively correlated with cancer are obtained. Now they will be analyzed from multiple perspectives.

Through single independent variable to lung cancer

Highly correlated features from the first method: Allergy, alcohol consuming, swallowing difficulty, wheezing, and coughing.

Highly correlated features from the first method: Alcohol consuming, Allergy, fatigue, swallowing difficulty, and yellow fingers.

Through muti-independent-variable to lung cancer

It uses a method called a two-way ANOVA (Analysis of Variance), it considers the influence of the interaction effect of the probably that the patients will affect cancer under two different factors.

Relationship between allergy, alcohol consuming and lung cancer:

From the results, it can be concluded that after the cross comparison between allergy and alcohol consuming and lung cancer, the F value is 6.43, and the p value is 1.17e-2 (less than 0.05, that is, it has a significant impact on the experiment results)

Relationship between alcohol consuming, swallowing difficulty, and lung cancer:

From the results, it can be concluded that after the cross comparison between alcohol consuming and swallowing difficulty and lung cancer, the F value is 15.91, and the p value is 8.31e-5 (less than 0.05, that is, it has a significant impact on the experiment results)

Relationship between swallowing difficulty, allergy, and lung cancer:

From the results, it can be concluded that after the cross comparison between swallowing difficulty and allergy and lung cancer, the F value is 20.26, and the p value is 9.65e-6 (less than 0.05, that is, it has a significant impact on the experiment results)

3.4 Model Prediction

The results in Table 3 are the accuracy of the models that are predicted by 4 different algorithms.

Table 3. the accuracy of 4 predicted models

Algorithm name	Test Accuracy	F1 score
Logistic Regression	0.9032	0.7799
KNN	0.8710	0.6891
Random Forest	0.8387	0.7015
Decision Tree	0.9032	0.6774

4. Discussion

More than 87% of the patients in the data set used in this study had lung cancer. According to various data, smoking has a relatively large impact on lung cancer. However, according to the results of this experiment, the correlation between smoking and lung cancer is only 0.058. There are many factors leading to this result. After carefully studying and searching the information on the Internet, it was concluded that the smoking time of the patients might not be considered in this data set, which led to the deviation between the conclusions and the facts. In future experiments, it is necessary to use a large data set to avoid some small probability events.

5. Conclusion

With the development of big data and the emergence of machine learning, mining medical big data can improve the efficiency of disease diagnosis, reduce the cost of disease discovery, and improve the survival rate. The four machine learning algorithms used in this study still have some defects. For example, in the algorithm using logistic regression, when the number of features is large, the performance of logistic regression is relatively poor, and the fitting is easy to be insufficient. In addition, the amount of data in the database used is still slightly insufficient. In the future, the experiment can be improved by using more different algorithms and collecting more data to eliminate the impact of some special data in the experiment, so as to predict more accurately and provide scientific basis for medical diagnosis decision-making.

References

[1] Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin, 2021, 71(3): 209-249.

[2] Zheng R, Zhang S, Zeng H, et al. Cancer incidence and mortality in China. J Natl Cancer Cent, 2022, 2 (1): 1-9.

[3] He J, Wei WQ. 2019 China cancer registry annual report. Beijing: People's Health Publishing House, 2021: 59-98.

[4] Zeng H, Chen W, Zheng R, et al. Changing cancer survival in China during 2003-15: a pooled analysis of 17 population-based cancer registries. Lancet Global Health, 2018, 6(5): e555-e567.

[5] Zeng H, Ran X, An L, et al. Disparities in stage at diagnosis for five common cancers in China: a multicentre, hospital-based, observational study. Lancet Public Health, 2021, 6(12): e877-e887.

[6] Cressman S, Peacock SJ, Tammemgi MC, et al. The cost effectiveness of high-risk lung cancer screening and drivers of program efficiency. J Thorac Oncol, 2017, 12 (8): 1210-1222.

[7] Siegel RL, Miller KD, Jemal A. Cancer statistics. CA Cancer J Clin, 2017, 67(1): 7-30.

[8] Simms KT, Hanley SJB, Smith MA, et al. Impact of HPV vaccine hesitancy on cervical cancer in Japan: a modelling study. Lancet Public Health, 2020, 5(4): e223-e234.

[9] Li N, Li J, Chen WQ, et al. Research progress of quality assessment of cancer screening guidelines and consensus. Chinese Journal of Epidemiology, 2021, 42(2): 211-214.

[10] Yuan Wenqian, Prediction of gastric cancer survival based on SEER database. Shanghai Normal University, 2021.