Lung Cancer Risk Analysis and Prediction Using Machine Learning Techniques

Hongyi Ding [1, †], Qi Tong [2, †], Hongran Wang [3, †, *], Zhan Zheng [4, †]

[1] China World Academy, Changshu, China

[2] Mian Yang Dong Chen international school, Mianyang, China

[3] Qingdao No.2 High School, Qingdao China

[4] Whittle School & Studio Shenzhen campus, Shenzhen, China

* Corresponding author email: wangxiaolin@qdu.edu.cn

[†]These authors contributed equally

Abstract. In this work, the main challenges are to find the factors for lung cancer and to use machine learning techniques to analyze the risk of lung cancer. Lung cancer is a malignant tumor, usually arising from the bronchial mucosa or glands of the lungs. The death rate of patients is very rapid. The incidence and death rates of lung cancer are increasing year by year in many countries. Over the past 50 years, many countries have reported significant increases in lung cancer morbidity and mortality. The incidence and mortality of lung cancer in men rank first among all malignant tumors, and the incidence and mortality in women rank second. The random forest and logistic regression are used to predict lung cancer risk based on patients' symptomatic and behavioral features.

Keywords: Lang Cancer; Random Forest; Logistic Regression; Machine Learning.

1. Introduction

Lung cancer is a malignant tumor, usually arising from the bronchial mucosa or glands of the lungs. The death rate of patients is very rapid. The incidence and death rates of lung cancer are increasing year by year in many countries. Lung cancer can be divided into two types, primary lung cancer, and secondary lung cancer. Primary lung cancer refers to cancers originating from the bronchial lungs, most of which originate from the bronchial mucosal epithelium and less frequently in the bronchial glands and alveolar epithelial cells. Primary lung cancer is also called bronchogenic carcinoma. Primary lung cancer is different from metastatic lung cancer [1]. Metastasis is the main reason when cancer has metastasized to other parts of the lung. Secondary lung cancer is metastatic lung cancer that develops from cancer elsewhere in the body [2]. Today, the cause of lung cancer has not been clearly studied, but several important influencing factors are as follows. Smoking can cause lung cancer: cigarettes contain various carcinogens, such as benzopyrene. Inhalation of cigarette smoke or tar can induce respiratory and skin cancer in laboratory animals. Smokers are ten times more likely to develop lung cancer than non-smokers, and the incidence of lung cancer is higher in heavy smokers, 20 times higher than non-smokers. Moreover, chronic lung diseases can cause lung cancer: Some people are unaware of chronic lung diseases, which may also cause lung cancer, such as pneumoconiosis, silicosis, tuberculosis, etc. If one has these lung diseases, they have a high chance of developing lung cancer. The family environment can cause lung cancer as well. House decoration materials and cooking fumes contain many ulcer-causing substances, passive smoking caused by smoking among family members, and the cause of lung cancer in women may be related to long-term inhalation of burns. The cooking fumes are related. 4. Occupational disease factors can cause lung cancer: This is because some jobs require daily exposure to radioactive substances such as radium and uranium, as well as mustard gas, asbestos, petroleum, asphalt, coal tar, arsenic, chromium, nickel, copper, tin, and other carcinogenic substances are the cause of lung cancer. 5. Air pollution can cause lung cancer: the prevention of lung cancer is to improve air pollution. Due to air pollution, the incidence of lung cancer in many industrialized countries is extremely high, and the urban area is higher than the rural area, and the factory and mining area is higher than the residential area [3]. In

the traditional sense, an important criterion for doctors to diagnose lung cancer is to obtain the patient's lung tissue for pathological examination through various examination methods. Case specimens can be obtained by looking for cancer cells in the patient's sputum, taking a biopsy under a bronchoscope, or a biopsy of the chest wall, or even need to open the chest to remove the lung tissue for pathological examination. Patients with pleural effusion can extract the patient's pleural effusion, and after centrifugation, remove the sediment to do a picture examination to find cancer cells. In addition to the pathological examination, CT examination is an important means of diagnosing lung cancer. Through chest CT examination, the cross-section of the chest wall can be displayed. Hidden parts that are not easily found under X-ray can be found, such as the interpulmonary area, above the diaphragm, next to the spine, Early lung cancer lesions behind the heart, and help differentiate from tuberculosis, pneumonia, etc. [4].

At present, there are many difficulties in diagnosing lung cancer: 1. Lung cancer is often asymptomatic in the early stage, especially for tumors that grow around the lungs. About one-third of lung cancers are discovered during routine physical examinations or other diseases, and some of them are not at an early stage. 2. There are various symptoms of lung cancer, none of which is unique to lung cancer and is easily covered by other acute and chronic lung diseases. 3. There is no simple and easy method for lung cancer screening, such as checking a few drops of blood as people hope to determine whether there is a tumor. Lung cancer is very harmful to all patients, even life-threatening [5]. Therefore, this work aims to explore the factors that induce cancer and make a lung cancer prediction model, which may help people prevent lung cancer and maintain good health.

Prediction is a data science task that is at the heart of many activities in an organization. Forecasting is a very important part of every industry. It can participate in project planning, effectively allocate resources and set goals, thereby improving performance. We used a logistic regression prediction model. Logistic regression, also known as logistic regression analysis, is a generalized linear regression analysis model often used in data mining, automatic disease diagnosis, economic forecasting and other fields [6]. We also used random forest. Random forest refers to a method that uses multiple decision trees to train, classify and predict sample data. While classifying the data, it can also give the importance of each variable (gene). The sex score evaluates the role of each variable in the classification [7]. We adopted a system that records people's attributes to predict lung cancer and collected and recorded 16 attributes and 284 attribute information. The data contains data on living habits, such as smoking and drinking. It also collects 10 features of related symptoms, such as having a yellow finger, anxiety, coughing, and shortness of breath.

2. Dataset

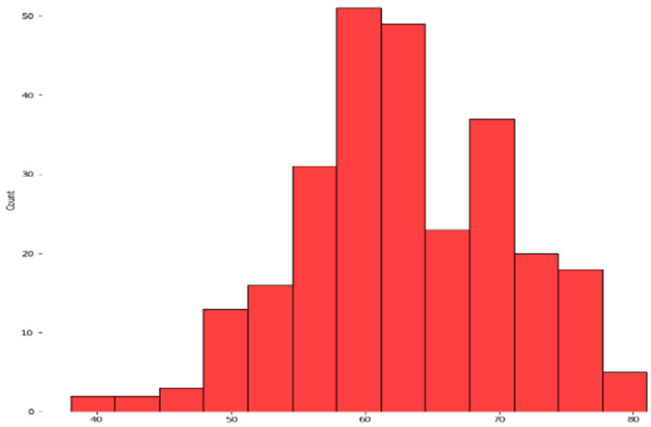

Fig 1. Age distribution of positive patients

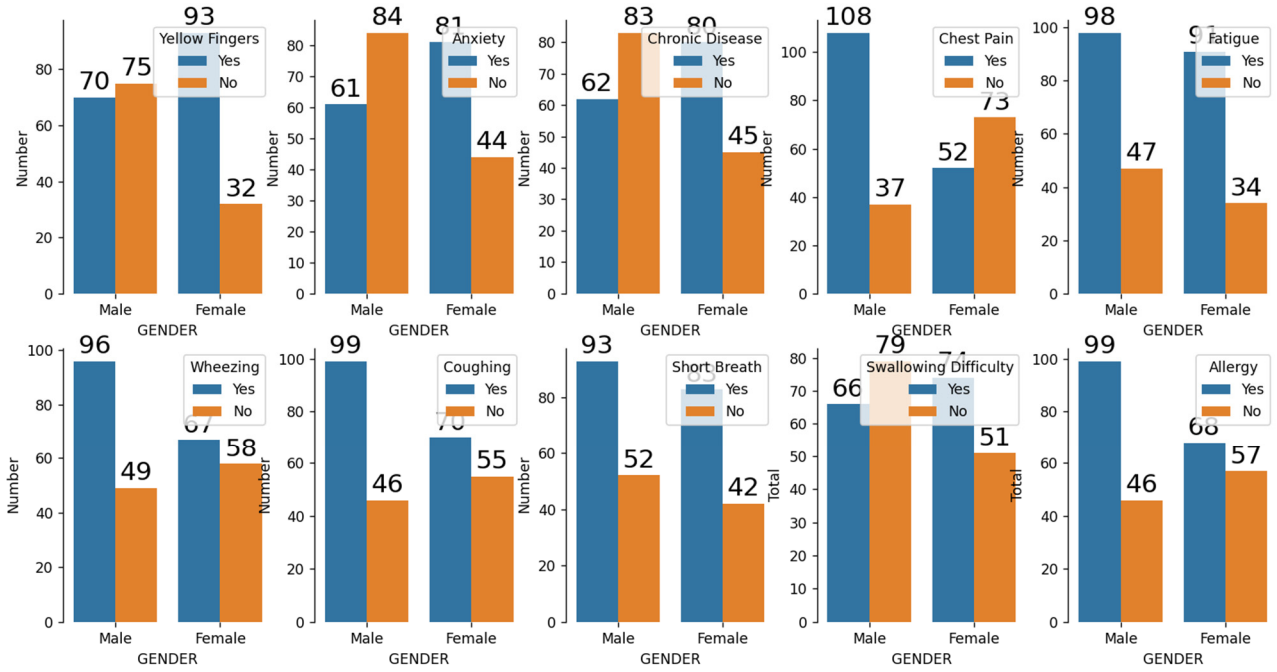

Fig 2. Data visualization of lifestyles and symptoms

We use the dataset collected by a website online lung cancer prediction system [8]. There are 309 samples in this dataset. There are 270 positive samples, with 39 negatives. Thus, the positive rate among the data is 87.4%. Regardless of positive or negative, there are other 15 columns, which are mainly separated into two parts, symptoms and lifestyle.

In the dataset, the ratio between men and women is 52.52:47.48. As shown in Fig. 1, the age distribution is mainly concentrated in the interval of age 55-75, which means that the citizens who are at this age have a higher probability of taking the survey.

3. Method

3.1 Cancer Factor Analysis

There are two types of features: symptom features and lifestyle features. We visualize the correlation between the features and the cancer risk in Fig. 2. It is noteworthy that smoking is an important indicator that follows common sense because it is widely recognized that a smoker is much easier to have lung cancer. Here, the ratio of males is 71 to 54, and the ratio of females is 60 to 53. The next part about lifestyle is the effects of alcohol. Drinking alcohol in positive cases is much more common in males rather than females, which is 101:24(M) and 44:69(F).

Coming to symptoms parts, the first choice for the survey-taker is the yellow fingers. In this case, the man's number is very close, and the amount of yes and no is 62:63. But in the women's part, people who answer yes increase rapidly to 84, and there are a few women who do not have a yellow finger. Fatigueness is another common symptom among positive cases. Both genders have a similar output, which is 84:41(M) and 81:32(F). Despite the many symptoms discussed above, there are still five more left. The wheezing ratio among males is 81 to 44, but for females, it is 61:52. Supplementary to this, coughing happened in a ratio of 85:40 in males and 64:49 in females. Besides, the short breath data ratio seems close, with males 80 to 45 and females73 to 40. Not only the shallowing difficulty ratio among gender (M:59: 66 and F: 65:48) but also the allergy (M:85: 40 and F: 61:52) is collected by the survey.

In the conclusion, 55-65 aged people were likely to get lung cancer. Bad lifestyles such as smoking and alcohol consumption clearly increase the possibility of getting lung cancer. Based on the high accuracy and the large base of data in Kaggle, it is useful to use logistic regression in predicting these

kinds of '0' or '1' situations. Without any of these, the prediction would not be very accurate, and it would be useless to study lung cancer. Using this method, we could predict if one patient would get lung cancer based on what symptom do, they got and how many symptoms the patients get.

3.2 Cancer Prediction

3.2.1 Logistic Regression

Logistic regression is a model which mostly used to analyze binary problems where two different outcomes are labeled as "0" and "1" [9]. It assumes a linear line with the relevance between the independent and dependent variables [10]. The logistic regression coefficients include the intercept (a_0) and slope (a_1) of this line. For most models, it is core to estimate a coefficient, especially the slope. We need to create a link function that illustrates the relevance between independent and dependent variables.

$$\ln(\frac{q}{1-q}) = a_0 + a_1 x \tag{1}$$

When finding the solution (probability q) for function, the possibility has a sigmoidal relevance with the independent variable, with the estimated possibilities properly curbed between 0 and 1[11].

After getting the probability, we could easily define if the independent variable is more likely to be "0" or "1" according to a baseline set with research and experience to make the regression fit to the need to do classification.

As for medical research, logistic regression has an evident advantage as it can provide the probability of getting a certain disease. Thus, we could give some practical advice based on the results.

3.2.2 Random Forest

Random forest [12] is a tree-based model, and it is popular in statistical analysis. As a tree-based model, the decision trees can solve classification tasks precisely by changing the partition and stopping criteria [13]. In random forest, the huge improvement in classification accuracy results from growing an ensemble of trees to vote for the most popular class [14]. The class chosen by the most trees will be the result for classification whilst the average prediction of each tree is the output for regression tasks.

For classification problems, we select parts of the predictor variables to separate an internal node, depending on preset criteria that are formulated as an optimization problem. One of those criteria in classification problems is entropy. At each internal node of the decision tree, entropy is given by the formula

$$E = -\sum_{i=1}^{a} p_i \times \log(p_i) \tag{2}$$

where a is the number of individual classes and pi is the prior possibility of each class which is given above. To get the most information at every decision-tree split, we need to maximize the value(E) [15]. After doing that, we could get the result for classification and apply it to another research.

4. Experiments

4.1 Implementation Details

We implement our models using the scikit-learn library from Python, and we perform the feature correlation analysis using the seaborn library. We use 80% percent of the data for training and test 20% of the data for testing. For the random forest, we choose n estimators, a parameter about the number of trees in the.

4.2 Result and Analysis

For the results, we measure the performance using precision, recall, f1-score, and support for each class. The precision is given by $\frac{true\ positrives}{true\ positrives + false\ positives}$, which indicates the ability of the

classifier not to consider a negative sample as a positive one. The recall is given by $\frac{true\ positives}{true\ positives + false\ negatives}$, which indicates the capability of the classifier to find the positive samples. F1-score is a weight-considered mean of recall and precision, where an F-beta score reaches its best score at 1 and worst value at 0. We also compute macro avg, which counts each label's metrics and calculates their mean (here is unweighted). However, this average does not consider the imbalance, and thus, we compute the weighted avg, which counts each label's metric and gets their average value, weighted by the support data.

In the first experiment, we used logistic regression to create a model and fit the dataset we got from Kaggle. It is surprising to see the high accuracy of the model according to table 1.

Table 1. Classification Report for Logistic Regression with the Accuracy is 96.77%

	precision	Recall	F1-score	Support
0	0.50	0.50	0.50	2
1	0.98	0.98	0.98	60
Accuracy			0.97	62
Macro avg	0.74	0.74	0.74	62
Weighted avg	0.97	0.97	0.97	62

After that, we generate a random forest classification to solve the problem and compare the results. We first set n-estimators to be 1, and the result did not go as well, as shown in Table 2.

Table 2. Classification Report for Random Forest with the Accuracy is 91.94%

	precision	Recall	F1-score	Support
0	0.20	0.50	0.29	2
1	0.98	0.93	0.96	60
Accuracy			0.92	62
Macro avg	0.59	0.72	0.62	62
Weighted avg	0.96	0.92	0.94	62

We then try to get the parameter to be bigger than 100, which means we put more resources and time into getting the model to fit the data. According to Table 3, we get the same result as logistic regression.

Table 3. Classification Report for Random Forest with the Accuracy is 96.77%

	precision	Recall	F1-score	Support
0	0.50	0.50	0.50	2
1	0.98	0.98	0.98	60
Accuracy			0.97	62
Macro avg	0.74	0.74	0.74	62
Weighted avg	0.97	0.97	0.97	62

We get a model that could provide if a person gets lung cancer based on the questionnaire. Based on the high accuracy, it is useful to use logistic regression in predicting these kinds of '0' or '1' situations. Random forests require great training resources to produce satisfying results. So, if the dataset is not large or we have enough resources to train, the random forest will be more precise. However, because the dataset is small, we could not get a more precise model.

5. Conclusion

This work discusses the prediction of lung cancer using logistic regression and random forest methods based on 10 features. The analytic results show that bad lifestyles, such as smoking and alcohol consumption, are the most important factors for cancer.

From this research, since bad lifestyles could increase the possibility of getting lung cancer, people should keep a healthy lifestyle and do more exercise to reduce the chance of getting lung cancer. Also, it is essential to develop data visualization techniques so the researchers can overview the data quickly and save time on research by providing a direct view of the database. In conclusion, lung cancer is a dangerous disease that causes 18.5% of cancer deaths, and it is very hard to cure once cancer has developed for more than 5 years. This work can reduce the doctor's pressure by quickly determining the cancer possibility and thus serves as a pre-screening test.

References

[1] Mayo Clinic, "Lung cancer - Symptoms and causes," Mayo Clinic, Mar. 23, 2021.

[2] B. E. Johnson, "Second lung cancers in patients after treatment for an initial lung cancer," JNCI: Journal of the National Cancer Institute, vol. 90, no. 18, pp. 1335–1345, 1998.

[3] American Cancer Society, "What Causes Lung Cancer?," Cancer.org, 2010.

[4] American Cancer Society, "How to Detect Non-small Cell Lung Cancer | Lung Cancer Tests," www.cancer.org, Jun. 01, 2021.

[5] S. M. Farber, M. A. Benioff, and J. D. Smith, "Diagnostic problems of cancer of the lung," California Medicine, vol. 76, no. 5, p. 328, 1952.

[6] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, Logistic regression. Springer, 2002.

[7] R. E. Wright, "Logistic regression." 1995.

[8] Z.-Q. Hong and J.-Y. Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane," pattern recognition, vol. 24, no. 4, pp. 317–324, 1991.

[9] Wikipedia Contributors. "Logistic Regression." Wikipedia, Wikimedia Foundation, 12 Apr. 2019.

[10] P. Schober and T. R. Vetter, "Logistic regression in medical research," Anesthesia and analgesia, vol. 132, no. 2, p. 365, 2021.

[11] "Lung Cancer," www.kaggle.com. https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer (accessed Sep. 10, 2022).

[12] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, Applied logistic regression. John Wiley & Sons, 2013, vol. 398.

[13] S. J. Rigatti, "Random Forest," Journal of Insurance Medicine, vol. 47, no. 1, pp. 31–39, 2017.

[14] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5–32, 2001.

[15] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," The Stata Journal, vol. 20, no. 1, pp. 3–29, 2020.