Lung cancer risk prediction and feature importance analysis with machine learning algorithm

Xinzhu Li

School of Information Science and Engineering, Shandong University, Qingdao, Shandong, 266237, China

202000120040@mail.sdu.edu.cn

Abstract. The morbidity and mortality of lung cancer are high and the detection is difficult, which poses a great threat to people's health. Early detection and accurate diagnosis can significantly improve survival rates. In recent years, machine learning models have been widely applied to classify and predict the risk of getting lung cancer based on clinical features and certain environmental factors. In the research, a decision tree model and a random forest model are developed and validated to predict lung cancer risk using a dataset that contains patients' living environments and clinical symptoms. The research also compared the amount of key features which are used in both decision tree and random forest model during prediction. Additionally, several key features in the prediction period are identified and are applied to a winform application designed for patients to test their risk level of getting lung cancer. The accuracy of the model used in winform application shows that the application can effectively predict lung cancer according to the clinical symptoms and living environment of patients, which proves that it has good application value.

Keywords: machine learning, winform, feature importance analysis.

1. Introduction

According to the survey, lung cancer is among the most commonly diagnosed cancers with an indispensably high mortality rate [1,2]. It is also one of the leading causes of cancer-related death due to late detection and the lack of effective treatment options. Therefore, identifying potential patients with a high risk of developing lung cancer is of vital importance for early detection. Besides the well-known risky factors such as smoking, age, and genetics, other factors such as alcohol use and air pollution should also be taken into account. Moreover, certain symptoms such as the coughing of blood, wheezing, and clubbing of fingernails are all important indicators of lung cancer.

Machine learning methods have already become an important tool for classification and prediction. By applying machine learning models, researchers can classify and predict potential cancer patients into high, medium, or low-risk groups and detect key features that can be used in prediction from complex datasets [3,4]. By applying machine learning methods, an application for patients to test their risk of getting lung cancer can be developed. Both patients and healthcare providers can benefit from it and take timely measures to control progression of lung cancer.

The predictive models discussed are based on supervised machine learning techniques. By applying machine learning methods, the importance of factors indicating lung cancer will be presented

2. Method

2.1. Dataset

The research is based on a dataset from Kaggle [5]. The dataset focused on several risky causes and several symptoms may indicate lung cancer. The number contained in the dataset is 1,000 and all the attributes (1 as patient Id, 23 as inputs to machine learning models and 1 as level that can be used as labels) are described as follows. All the features are described categorically except for the age of patients.

Patient Id: This feature refers to the Id of patients, but it is useless in machine learning methods. ***Age:*** This numerical feature refers to the age of surveyed patients. ***Gender:*** This feature is a categorical description of the gender of the patient. ***Air pollution:*** This feature refers to the level of air pollution of the environment around patients. ***Alcohol use:*** This feature refers to the level of alcohol use of the patient. ***Dust Allergy:*** This feature reflects patients' degree of dust allergy. ***Occupational Hazards:*** This feature indicates the the potential risk of the patient's occupation. ***Genetic Risk:*** This feature refers to the level of genetic risk of the patients, taking into account whether there were other patients in the potential patients' family. ***Balanced Diet:*** This feature refers to how healthy the patients' diet is. ***Obesity:*** This feature refers to the level of obesity of the patient. ***Smoking:*** This feature refers to the frequency of smoking. ***Passive smoker:*** This feature indicates whether there are other smokers around patient. ***Chest Pain:*** This feature refers to the the severity of the patient's chest pain. ***Coughing of Blood:*** This feature refers to the level of the symptom of coughing of blood of the patient. ***Fatigue:*** This feature refers to whether the patient always feels tired. ***Weight Loss:*** This feature refers to the level of weight loss of the patient. ***Shortness of Breath:*** The feature indicates the frequency of whether the patient is often short of breath. ***Wheezing:*** This feature refers to the frequency of the patients' wheezing symptom. ***Swallowing Difficulty:*** This feature refers to the severity level of patients' swallowing difficulty. ***Clubbing of Finger Nails:*** This feature refers to whether the patients have the symptom of finger nails clubbing.

2.2. Data preprocessing

Before applying machine learning models, data cleaning, checking for null values, replacing "level" with integers should be achieved.

For data cleaning, the column which refers to patient Id will be dropped. After checking, it can be found that there are no null values in the dataset. As for the "level" column, it will be replaced with integers. 2 will represent high level, 1 will represent medium level and 0 will represent low level. Then they can be used as labels to be applied in machine learning models.

After the above procedures, the dataset will be randomly divided into training set and development set in an 8:2 ratio considering the small scale of the dataset. The models will be evaluated on the development set using cross-validation.

2.3. Machine learning models

Considering the scale of the dataset isn't big enough and may not contain linear relationship, Decision Tree and Random Forest are employed. By applying these models, ranking importance of the features which indicate lung cancer can also be achieved.

2.3.1. *Decision tree.* Firstly, Decision Tree (DT) is considered. Decision tree is a common method for establishing systems used for classification based on various features and for prediction for a target variable [6,7]. It works by recursively partitioning the input space into small regions, using a set of decision rules based on input features. The algorithm starts with a root node that represents the entire input space, and then selects the feature the best separates the input data based on some measures of purity such as entropy or information gain. The selected feature is then used to split the input space into more regions. The process is repeated for each child node, creating a tree-like structure.

2.3.2. *Random forest.* Random forest (RF) model is a popular machine learning method for prediction and is useful in reducing the number of variables needed for prediction in order to lighten the burden of data collection [8,9]. Since the goal is to build an application for the potential patients to predict their risk of developing lung cancer, it is a suitable model. Random forest is a kind of ensemble machine learning method that gathers multiple decision trees to improve the accuracy and stability of predictions. It works by generating a group of decision trees, each of them is trained on a randomly selected subset of the input features and training data. During prediction, each tree independently gives an output, and the final prediction result is obtained by taking the average mean or taking the majority vote of the individual tree predictions.

2.3.3. *Grid search cross validation.* In order to find out proper hyperparameters, grid search cross validation (GridSearchCV) is employed. According to the research, GridSearchCV helps to produce the best hyperparameters for testing accuracy [10]. GridSearchCV works by specifying a set of hyperparameters to be tuned and a range of values for each hyperparameter. The method then generates all possible combinations of hyperparameters and trains and evaluates the model for each combination using cross-validation. The hyperparameter combination that yields the highest performance score is then selected as the optimal hyperparameters for the model. By using cross-validation, the accuracy of the model is also evaluated on the test set.

3. Result

According to the dataset, feature correlations are calculated. It is clear that certain features such as obesity, coughing of blood, and alcohol use have significantly high correlation with lung cancer, which correlations are 0.83, 0.78, and 0.72 respectively. On the other hand, features such as age, gender and wheezing have low correlation with lung cancer, which correlations are 0.06, -0.16 and 0.24 respectively.

Besides data, performance of model is also an important factor in lung cancer prediction, which results are demonstrated in Table 1.

Table 1. Evaluation of the models.

Model	Train set accuracy	Test set score
Decision Tree Model	100%	1.0
Random Forest Model	100%	1.0

Both of the models performed well in this dataset. By applying decision tree model and random forest model, the key factors can also be found out. The factors and their importance during prediction will be listed in Table 2 and Table 3, showing the result of random forest model and decision tree model respectively.

Table 2. Feature importance from random forest model.

Feature	Importance	Feature	Importance
Coughing of blood	0.3734	fatigue	0
Wheezing	0.2579	age	0
Obesity	0.1290	smoking	0

Table 2. (continued).

Snoring	0.1131	Balanced diet	0
Air pollution	0.0713	Chronic lung disease	0
Clubbing finger nails	0.0553	Genetic risk	0
Chest pain	0	Occupational hazards	0
Dry cough	0	Dust allergy	0
Frequent cold	0	Alcohol use	0

Swallowing difficulty	0	gender	0
Shortness of breath	0	Passive smoker	0
Weight loss	0		

It is clear that compared with random forest, decision tree model used significantly less features in prediction, which may lead to unstable results. Therefore, random forest model will be given priority in application.

The following Figure 1 shows importance of features used in random forest prediction. In the picture, x axis represents the rank of the features.

Table 3. Feature importance from decision tree model.

Feature	Importance	Feature	Importance
Coughing of blood	0.1692	Shortness of breath	0.0325
Wheezing	0.0895	Occupational hazards	0.0282
Genetic risk	0.0758	Balanced diet	0.0270
Fatigue	0.0745	Chest pain	0.0267
Clubbing finger nails	0.0698	Dust allergy	0.0266
Passive smoker	0.0545	Frequent cold	0.0155
Air pollution	0.0534	Weight loss	0.0145
Smoking	0.0524	Dry cough	0.0094
Swallowing difficulty	0.0513	Chronic lung disease	0.0049
Obesity	0.0479	Age	0.0033
Alcohol use	0.0379	Gender	0.0001
Snoring	0.0351		

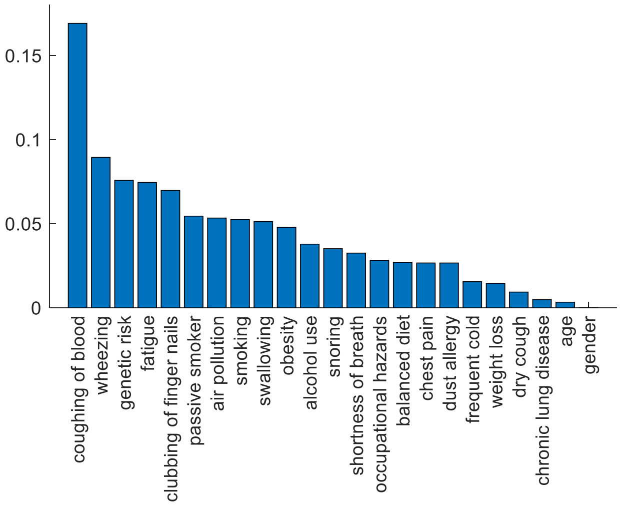


Figure 1. Importance of features in random forest prediction.

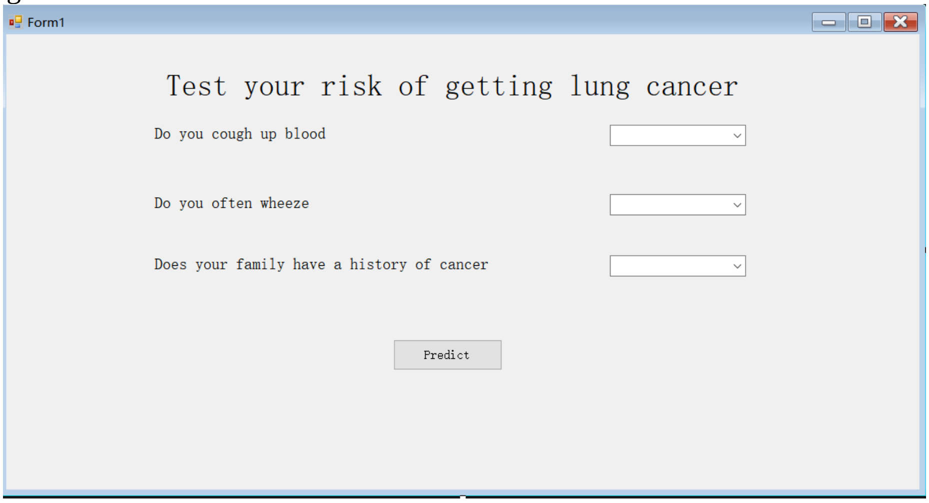
It is easy to find that the importance of features used in prediction is different from the correlation rank. It is because that the feature importance obtained from random forests and the magnitude of the correlation coefficient are not directly related. The importance of features in the random forest is assessed by calculating the contribution of each feature during prediction period to the accuracy of the model, which measures the effect of the feature on the classification results in the model, while correlation coefficient measures the linear relationship between two variables. It's possible for a feature to have high importance in a random forest model but low correlation with the target variable, such as the wheezing feature.

By extracting the top three important features which represents coughing of blood, wheezing and genetic risk, a winform application is built to help patients to predict their risk of getting lung cancer. The random forest model is retrained using only these features. Moreover, since is difficult for patients to classify their levels in these features, this work also compressed the data in the original dataset. The original level of 1 to 3 is replaced with 0, 4 to 6 is replaced with 1, and 7 to 9 is replaced with 2. Compared with the original model, Table 4 displays the performance of the new model.

Table 4. Result comparison of the new model.

Model	Training accuracy	Testing score
Decision Tree	100%	1.0
Random Forest (original)	100%	1.0
Random Forest (new)	87%	0.9

Although it performs worse than the previous model, it reduces 20 attributes and reduces the complexity as well. This new model is more practical in clinical application. The python script of the model will be run by calling the cmd window using C#. The winform application is shown as the following Figure 2.



Test your risk of getting lung cancer

Do you cough up blood

Do you often wheeze

Does your family have a history of cancer

Figure 2. The winform application for patients.

By selecting their level of the features, they have in the combo boxes and press the predict button, patients can get a prediction result indicating high risk, medium risk, and low risk through a message box.

4. Conclusion

Lung cancer is a serious disease, and predicting the probability of getting it is important for early prevention and treatment. Random forest and decision tree are two common but useful machine learning algorithms that can both be used to predict the probability of lung cancer. By applying random forest and decision tree, the importance of certain factors in predicting lung cancer is obtained and a winform application is built. By accurately predicting the likelihood of lung cancer, doctors and patients can take necessary measures such as regular screening or lifestyle modifications to reduce the risk of developing lung cancer. Additionally, accurate lung cancer prediction models can help clinicians make more informed and timely decisions about the treatment and management of lung cancer, potentially leading to better patient outcomes.

In this research, random forest and decision tree both perform well in predicting lung cancer using the same dataset. It can also be discovered that certain symptoms can predict lung cancer better than environment factors. However, factors such as whether the data set contains a set of features that is comprehensive enough is still under question. The method still requires further research.

References

[1] Schabath, M. B., & Cote, M. L.: Cancer progress and priorities: lung cancer. Cancer epidemiology, biomarkers & prevention, 28(10), 1563-1579 (2019).

[2] Bade, B. C., & Cruz, C. S. D.: Lung cancer 2020: epidemiology, etiology, and prevention. Clinics in chest medicine, 41(1), 1-24 (2020).

[3] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I.: Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17 (2015).

[4] Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., & Waddell, N.: Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Medicine, 13(1), 1-17 (2021).

[5] Kaggle-Lung Cancer Prediction, https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

[6] Song, Y. Y., & Ying, L. U.: Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130 (2015).

[7] Charbuty, B., & Abdulazeez, A.: Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28 (2021).

[8] Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E.: A comparison of random forest variable selection methods for classification prediction modeling. Expert systems with applications, 134, 93-101 (2019).

[9] Schonlau, M., & Zou, R. Y.: The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29 (2020).

[10] Ahmad, G. N., Fatima, H., Ullah, S., & Saidi, A. S.: Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. IEEE Access, 10, 80151-80173 (2022).