Exploring the Determinants of Lung Cancer: A Machine Learning-Based Approach

Hanxuan Ye

Computer Science and Technology, Jinan University, Guangzhou, 510000, China
yehanxuan@stu.sdp.edu.cn

Abstract. In the current field of medical research, identifying and evaluating the key factors affecting the onset of cancer and especially the lung cancer's onset is of great significance to improve the early management and diagnosis of lung cancer. This study employed a comprehensive approach, utilizing both the random forest algorithm and logistic regression, to analyze and predict the risk factors associated with lung cancer. Logistic regression algorithm can provide a model for the probabilistic relationship between features and lung cancer risk. Pearson correlation analysis's feature importance scoring method and random forest algorithm can select the most influential features from numerous potential risk factors to build an efficient lung cancer risk prediction model. The study began with a preliminary analysis of multiple variables in the data set to determine their relevance to lung cancer development. Pearson correlation analysis was employed to assess the magnitude of the linear relationship between each feature and the risk of lung cancer, and random forest algorithm was further used to score and rank the importance of the features. On the basis of feature selection, specific features were selected as input variables for model training, and a lung cancer risk prediction model was constructed by machine learning algorithm. By comparing and analyzing the baseline model constructed with all the features, the selected feature model maintains comparable or even higher prediction accuracy while reducing the model complexity. This result proves that feature selection plays a crucial role in enhancing model efficiency and accuracy.

Keywords: Lung Cancer; Logical Regression, Pearson Correlation; Random Forest

1 Introduction

Lung cancer remains the primary cause of cancer-related fatalities worldwide, presenting a significant public health challenge. Historical records show a stark increase in lung cancer cases, from fewer than 22 documented cases in the late 1840s to 1.8 million newly diagnosed cases globally in 2012, resulting in 1.6 million deaths in that same year alone [1-3]. Despite advancements in diagnostics and treatment, early detection and management of lung cancer are hindered by the complex interplay of genetic, environmental, and lifestyle risk factors. A comprehensive understanding of how these

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_2

factors collectively impact lung cancer risk is lacking, as many studies focus on isolated factors without assessing their interaction and relative importance.

Magnetic Resonance Imaging has been called "Since the X-ray found that the development of the most important medical diagnosis field" 100 years ago [4]. Throughout its brief history, starting from the pioneering research by Lauterbur et al. and subsequent foundational studies, MRI has evolved into a fundamental cornerstone of contemporary medical care [5]. However, the principle of MRI is subtle and less intuitive compared to more traditional medical imaging, and its effectiveness in imaging air-filled lungs is limited. Moreover, statistical analysis plays a vital role in revealing the complexity of disease incidence, progression, and treatment outcomes. Random Controlled Trails, multiple regression analyses, etc. have greatly improved the understanding of lung cancer, using random effects models to assess the likelihood of publication bias for all prior lung conditions and all subcategories by funnel plots and to assess sources of heterogeneity by meta-regression [6, 7]. The field of lung cancer research and diagnosis has undergone profound changes through the integration of Machine Learning (ML) technologies, with major leaps in the ability to detect, classify, and predict the outcomes of this complex disease, and as these technologies continue to evolve [8], they have the capacity to transform the diagnosis, therapy, and outlook of lung cancer, ultimately enhancing results and standard of living.

In order to solve such problems that may exist in traditional methods, this study used a dual analysis method, combining Pearson correlation analysis with the Random Forest algorithm's feature importance score, to isolate those factors that contribute to lung cancer risk. The study sought to go beyond statistical analysis to gain a deeper understanding of the biological and environmental mechanisms of lung cancer. By identifying the most influential risk factors, the aim is to shed light on the pathways by which these factors influence lung cancer development. By improving the specificity and accuracy of lung cancer risk prediction, this study aims to promote a paradigm shift in methods for lung cancer prevention and early detection.

This study meticulously applies Pearson correlation analysis to evaluate the linear relationship between various factors and lung cancer risk. Subsequently, the Random Forest algorithm is utilized to perform a feature importance analysis, allowing for the ranking and selection of the most influential factors. This dual-analytical approach facilitates the identification of key risk factors and their incorporation into a predictive model. Through dividing the data into smaller segments, a corresponding decision tree is gradually built [9]. By comparing the performance of this model against those built with a comprehensive dataset, this study aims to demonstrate the efficacy of focusing on significant predictors, thereby contributing to the optimization of lung cancer risk prediction models.

2 Method

2.1 Data Preparation

The study utilizes a dataset sourced from the Kaggle website, encompassing medical records of over 300 patients [10]. This dataset includes 15 features, such as age, smoking status, and alcohol consumption, aimed at predicting a binary outcome regarding lung cancer presence (whether a patient has lung cancer or not). Through thorough analysis of these features, this study identified the most impactful factors for the model's predictive performance. These selected features were then used to train the proposed prediction model.

In the data preprocessing stage of this study, standardized data processing is conducted to eliminate the influence of different measurement scales and ensure the effectiveness of model training. Second, this study converts the string-type features in the data to numeric features, a step that is critical to making the data fit into the algorithmic model.

After data preprocessing, this study performed Pearson correlation analysis to identify linear relationships between features in the dataset. By calculating the Pearson correlation coefficient between the features, which features had a significant linear correlation with the binary classification outcome of lung cancer can be determined. This analysis not only helps to understand the correlation between different features and the likelihood of developing lung cancer, but also provides the scientific basis for feature selection.

2.2 Machine Learning-based Lung Cancer Prediction

The Workflow of Machine Learning. From this study, a series of key steps shown in Fig. 1. are typically followed to build and evaluate the model. First, the problem definition phase defines the project's objectives and problem statement. The data collection phase then involves acquiring and collating data sets for training and testing the model. The data preprocessing stage includes steps such as cleaning data, processing missing values, and feature engineering to ensure data quality and availability. The model building phase then involves selecting the appropriate model type and model training to learn the data patterns. In the branch of model construction, the model selection phase involves selecting the model type that best fits the problem, while the model training phase involves fitting the model parameters using the training data. The parameter tuning phase involves adjusting model parameters to optimize performance. Finally, the model evaluation phase evaluates the model's performance utilizing test datasets. These steps constitute a complete machine learning project flow, helping to ensure that efficient and accurate models are built to solve specific problems.

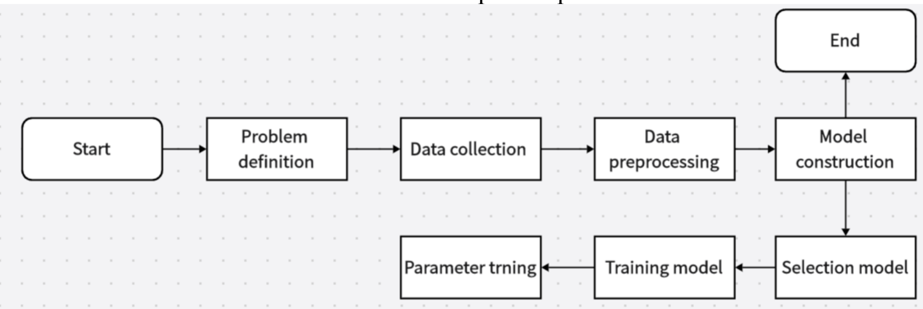

Fig. 1. Process flow of machine learning.

Logistic Regression. The potential impact of data science and machine learning in healthcare is garnering growing attention [11-15]. Machine learning includes a spectrum of techniques, spanning from practical statistical methods like supervised logistic regression models to more complex computational models, for example, a variety of neural networks [16].

Logistic Regression, a commonly employed algorithm in statistics and machine learning, is primarily utilized for addressing binary classification tasks. Despite its name containing "regression," logistic regression functions as a classification technique. It makes predictions about the likelihood of an event happening by mapping the output of linear regression between 0 and 1 using a logical function (usually the Sigmoid function).

The basic form of logistic regression model can be expressed as: y=σ(θTX), y is the predicted output (0 or 1) given the input X (eigenvector),θ is model parameters (including weights and bias)，σ is sigmoid function.

Sigmoid function, also known as the logistic function, is the crucial part of logistic regression. Its formula is:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

This function corresponds to any real number z to the interval (0,1), making it interpretable as a probability. Within the realm of logistic regression, z is typically the linear combination of features and parameters, i.e., $z=\theta TX$.

In logistic regression, this study uses the dot product of the model parameters θ (including weights and bias) with the feature vector X to represent the linear combination of the input data:

$$\theta^T X = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + ... + \theta_n X_n \tag{2}$$

Here, θ0 is the bias term (also known as the intercept), θ1,θ2,...,θn are the weights, and X1,X2,...,Xn are the input features.

By passing the linear combination θTX through the sigmoid function, this study obtains the predicted probability y:

$$\mathrm{y} = \sigma(\theta^T X) \tag{3}$$

This probability represents the likelihood of the target variable y being of the positive class (usually encoded as 1) given the input features X.

To train the logistic regression model, this study needs a loss function to assess the variance between the model's predictions and observed outcomes. The commonly used loss function in logistic regression is the log loss (also known as logistic loss or cross-entropy loss):

$$\mathrm{J}(\theta) = -\frac{1}{m} + \sum_{i=1}^{m} [y^{(i)} log(y^{(i)}) + (1 - y^{(i)}) \log (1 - y^{(i)})] \tag{4}$$

Here, m represents the number of samples, $y^{(i)}$ denotes the true label of the i-th instance, and $y^{(i)}$ represents the predicted probability by the model for the i-th instance.

Finally, training the logistic regression model involves minimizing the loss function J(θ), typically achieved through gradient descent or other optimization algorithms. The update rule for gradient descent is:

$$\theta j = \theta j - \alpha \frac{\partial J(\theta)}{\partial \theta j} \tag{5}$$

Here, α signifies the learning rate, and $\frac{\partial J(\theta)}{\partial \theta j}$ signifies the partial derivative of the loss function concerning the parameter θj.

By iteratively updating the parameters θ, it can find the minimum of the loss function, thereby training an optimal logistic regression model.

Random Forest. This study initially explores tree-based models as they serve as the foundational elements of random forest algorithm. The model entails iteratively partitioning the provided dataset into two subsets based on a specific criterion until a predefined stopping criterion is satisfied. The end nodes of decision trees are referred to as leaves [17].

The two-dimensional input Spaces are all segmented into the direction aligned with one of the axes. Fig. 2 indicates the graphical representation of divided subspaces.

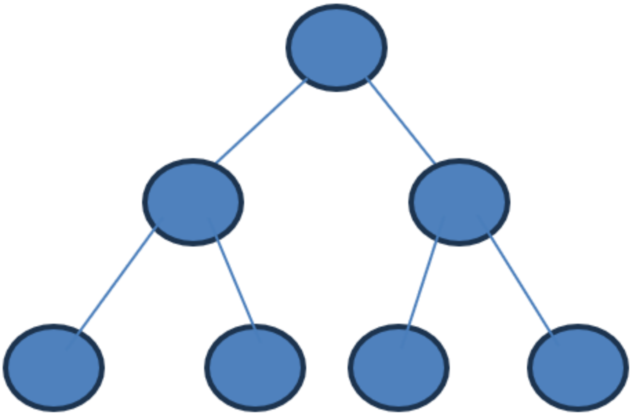

Fig. 2. Graphical representation of divided subspaces.

Random Forest, an ensemble learning technique, is predominantly employed for classification and regression tasks, demonstrating strong performance across a wide array of datasets. Additionally, Random Forest is robust against statistical assumptions and preprocessing complexities, capable of effectively managing extensive datasets characterized by high dimensionality and missing values [18]. It improves model accuracy and stability by constructing multiple decision trees and aggregating their predictions. A key advantage of Random Forest is its ability to handle nonlinear data and its robustness against overfitting, especially in datasets with a large number of features. Random Forest comprises numerous decision trees, each trained autonomously with randomness introduced in two ways: In ensemble learning methods like Random Forest, two key techniques are employed to enhance the diversity and robustness of individual decision trees within the ensemble. Firstly, bootstrap sampling involves training each decision tree on a random sample of the original dataset selected with replacement. This process creates multiple subsets of the data, allowing each tree to learn from slightly different perspectives. Secondly, random feature selection is implemented at each decision node, where the algorithm chooses a subset of features randomly to determine the best split. By introducing randomness in both data sampling and feature selection, Random Forest promotes model variance and reduces overfitting, leading to more reliable and accurate predictions across diverse datasets.

In a random forest, two key techniques are used to improve the diversity and robustness of the individual decision trees in the integration. First, through self-sampling, each decision tree is trained on stochastic samples of original data, using a put-back sampling method. This process creates multiple subsets of the data, enabling each tree to learn from a slightly different perspective. Secondly, at each decision node, the algorithm randomly selects a subset of features to determine the best split. By introducing randomness into data sampling and feature selection, random forests promote model variability and reduce overfitting, resulting in more reliable and accurate predictions on different data sets.

In classification tasks, Random Forest predictions are determined through a majority vote or averaging of all decision tree predictions. The ultimate class is the one that garners the most votes. In regression tasks, predictions are calculated as the average of all decision tree predictions.

Random Forest model development typically involves several key steps. Firstly, model parameters, such as the quantity of trees, maximum tree depth, and minimum samples for splitting, are selected to configure the ensemble. Subsequently, the model is trained using training data, where each decision tree learns from bootstrapped samples and random feature subsets. Following training, the model's performance is evaluated using separate test data to assess its predictive accuracy. To optimize model performance, parameters are tuned based on evaluation results, and the training-evaluation cycle is iterated until the optimal configuration is achieved. This iterative process of parameter selection, training, evaluation, and tuning is crucial for building a Random Forest model that effectively generalizes to new data and delivers robust predictions.

Random Forest also offers insights into feature importance, which is invaluable for understanding the data and the model's decision-making process. The evaluation of feature importance is primarily conducted by observing the contribution of each feature to model performance at the decision tree split nodes. In particular, the significance of a feature can be assessed by evaluating the reduction in model performance when the values of that feature are randomly shuffled. A higher feature importance score indicates a greater influence of that feature on the predictive outcome of the model.

3 Result and Discussion

3.1 The Feature Selection Based on Pearson Correlation and Random Forest

Table 1. Pearson Correlation with target variable.

SMOKING	-0.058179
SHORTNESS OF BREATH	-0.060738
AGE	-0.089465
CHRONIC DISEASE	-0.110891
ANXIETY	-0.144947
FATIGUE	-0.150673
YELLOW_FINGERS	-0.181339
PEER_PRESSURE	-0.186388
CHEST PAIN	-0.190451
COUGHING	-0.248570
WHEEZING	-0.249300
SWALLOWING DIFFICULTY	-0.259730
ALCOHOL CONSUMING	-0.288533
ALLERGY	-0.327766

Table 2. Random Forest Feature Importance.

AGE	0.248064
ALLERGY	0.086116
ALCOHOL CONSUMING	0.074769
YELLOW_FINGERS	0.071846
SWALLOWING DIFFICULTY	0.065598
PEER_PRESSURE	0.058936
CHRONIC DISEASE	0.055938
ANXIETY	0.052924
COUGHING	0.052341
WHEEZING	0.051886
FATIGUE	0.051380
CHEST PAIN	0.050294
SHORTNESS OF BREATH	0.045077
SMOKING	0.034832

Table 3. Random Forest Accuracy.

	precision	recall	f1-score	support
1	0.95	0.99	0.97	113
2	0.83	0.45	0.59	11
accuracy			0.94	124
macro avg	0.89	0.72	0.78	124
weighted avg	0.94	0.94	0.94	124

The Pearson correlation coefficient shown in Table 1 shows the linear correlation between each feature and the target variable, which may be a certain disease or other outcome. A negative value signifies a negative correlation meaning the magnitude of target variable tends to decrease as the eigenvalue increases. Here, the feature that affects the target variable least is "ALLERGY" (-0.327766) and the feature that affects most is "COUGHING" (-0.248570).

The feature importance of a random forest model shown in Table 2 shows how important each feature is to the model's predictions. Here, "AGE" is the most important characteristic (0.248064), followed by "ALLERGY" (0.086116) and "ALCOHOL CONSUMING" (0.074769).

The accuracy of the random forest model trained using the selected features is 0. 94, while the accuracy using all features is 0.91. This suggests that in this case, the model using feature selection performs better.

Combining the above analysis, this study can conclude that COUGHING and ALLERGY may be key features of the predicted target variable. The random forest model considers AGE as the most important feature, which has a significant impact on the prediction of target variables.

According to Table 3, the model trained using the selected features performed well, with an accuracy of 0.927. In category 1 predictions, the model showed high accuracy is 0.94, the recall is 0.98 and the F1-score is 0.96, but was weaker in category 2 predictions, with the precision of 0.67, 0.36 recall, and the F1-score is 0.47. This model excels in the context of overall data set with weighted avg accuracy, recall rate, and F1-score of 0.92. Although there is room for improvement in the category 2 predictions, overall the model is robust on the dataset, but further optimization is recommended to improve the prediction accuracy for category 2.

3.2 Logistic Regression

Table 4. The Performance of Logistic Regression.

	precision	recall	f1-score	support
1	0.95	0.99	0.97	113
2	0.83	0.45	0.56	11
accuracy			0.94	124
macro avg	0.89	0.72	0.78	124
weighted avg	0.94	0.94	0.94	124

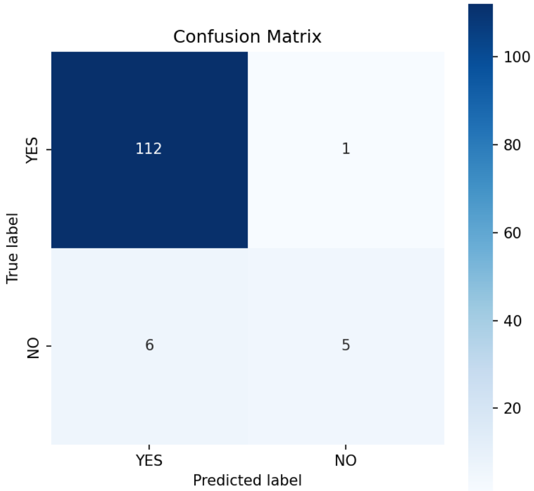

Fig. 3. The Confusion Matrix.

According to the confusion matrix shown in Fig. 3, the model accurately predicted 112 samples in the "YES" category, but incorrectly predicted 1 sample in the "YES" category as the "NO" category. For the "NO" category, the model correctly predicted 5 samples, but mistakenly predicted 6 samples of the "NO" category as the "YES" category. This means that the model has some difficulty in making predictions about the "NO" category.

According to the above model evaluation metrics shown in Table 4, predictions for category 1 (1) performed well, achieving the 0.95 of accuracy, the 0.99 in recall rate, the F1-score of 0.97, with sample size of 113. This indicates that the model has high precision and recall rate when identifying category 1, and the comprehensive evaluation index also performs well. However, for category 2 (2) predictions, the model's performance is more average, with the accuracy is 0.83, the recall rate is 0.45, and F1-score is 0.59, with a supporting sample size of 11. This means that the model has some challenges in recognizing category 2, especially in terms of recall rates.

In total, the model's accuracy stands at 0.94, with weighted average accuracy, recall, and F1-score also at 0.94, respectively, indicating that the model performs well on the overall dataset. However, the macro average indicator shows the accuracy is 0.89, the recall rate is 0.72, and the F1-score (0.78), indicating that the model is lacking in balance between different categories. It is suggested that future research should focus on improving the prediction ability of category 2 to further improve the overall performance and balance of the model.

4 Conclusion

This study analyzed and predicted the risk factors of lung cancer by using logistic regression and random forest algorithm. Logistic regression algorithm can provide the modeling of the probabilistic relationship between features and lung cancer risk. Pearson correlation analysis and random forest algorithm are used in this study to select the most influential features from numerous potential risk factors and build an efficient lung cancer risk prediction model. Comprehensive experiments were carried out to assess the proposed approach. The experimental results demonstrate the efficacy of logistic regression model is close to the performance of selected features and all features, while the performance of random forest model after feature selection is slightly better than that when all features are used. In the future, feature selection can be further optimized to improve prediction accuracy across all categories. Continuous improvement in the feature selection process can potentially improve the overall prediction performance of all categories, continuously improving the accuracy of predictions by optimizing the model.

References

1. Hasse, C.: Cancerous tumors in the respiratory organs. In: Swaine, W. (ed.) An Anatomical Description of the Diseases of the Organs of the Circulation and Respiration, pp. 370-375. Sydenham Society, London, England (1846).

2. Debakey, M.: Carcinoma of the lung and tobacco smoking: a historical perspective. Ochsner Journal 1, 106-108 (1999).
3. Ferlay, J., Soerjomataram, I., Ervik, M., et al.: GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. International Agency for Research on Cancer, Lyon, France (2013).
4. Hashemi, R. H., Bradley, W. G., Lisanti, C. J.: MRI: the basics. Lippincott Williams & Wilkins (2012).
5. Plewes, D. B., Kucharczyk, W.: Physics of MRI: a primer. Journal of Magnetic Resonance Imaging 35(5), 1038-1054 (2012).
6. Berkey, C. S., Hoaglin, D. C., Mosteller, F., Colditz, G. A.: A random-effects regression model for meta-analysis. Statistics in Medicine 14, 395-411 (1995).
7. Brenner, D. R., McLaughlin, J. R., Hung, R. J.: Previous lung diseases and lung cancer risk: a systematic review and meta-analysis. PLoS One 6(3), e17479 (2011).
8. Qiu, Y., Wang, J., Jin, Z., Chen, H., Zhang, M., Guo, L.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control 72, 103323 (2022).
9. Kumar, G. K., Viswanath, P., Rao, A. A.: Ensemble of randomized soft decision trees for robust classification. Sadhana - Academy Proceedings in Engineering Sciences 41(3), 273-282 (2016).
10. Kaggle: Lung Cancer Analysis Accuracy 96.4%. Available at: https://www.kaggle.com/code/hasibalmuzdadid/lung-cancer-analysis-accuracy-96-4/input.
11. Toh, C., Brody, J. P.: Applications of machine learning in healthcare. In: Kheng, T. Y. (ed.) Smart Manufacturing - When Artificial Intelligence Meets the Internet of Things, IntechOpen (2021).
12. Bhardwaj, R., Nambiar, A. R., Dutta, D.: A study of machine learning in healthcare. Proceedings - International Computer Software and Applications Conference, pp. 236-241 (2017).
13. Chen, P. C., Liu, Y., Peng, L.: How to develop machine learning models for healthcare. Nature Materials 18, 410-414 (2019).
14. Saleem, T. J., Chishti, M. A.: Exploring the applications of machine learning in healthcare. International Journal of Sensor Wireless Communications and Control 10, 458-472 (2019).
15. Health Education England: Topol Review. Accessed: March 25, 2019, Available at: https://www.hee.nhs.uk/ourwork/topol-review.
16. Cleophas, T. J., Zwinderman, A. H.: Machine Learning in Medicine - A Complete Overview. Springer, Switzerland (2020).
17. Schonlau, M., Zou, R. Y.: The random forest algorithm for statistical learning. The Stata Journal 20(1), 3-29 (2020).
18. Zhu, T.: Analysis on the applicability of the random forest. Journal of Physics: Conference Series 1607(1), 012123. IOP Publishing (2020).