

Lung Cancer Prediction using Machine Learning Methods

Qijun He*

School of Queen Mary Hainan, Beijing University of Posts and Telecommunications, Beijing,
100000, China

*Corresponding author: jp2022213740@quml.ac.uk

Abstract. Lung cancer, being a major global health issue, requires not only effective treatments but also robust predictive methods to enable early diagnosis and preventive measures. Detecting the disease at an early stage increases survival chances and reduces the necessity for aggressive treatments with harsh side effects. This paper seeks to contribute to the ongoing research by utilizing machine learning techniques to analyze a variety of factors that influence the severity of lung cancer, such as smoking habits, air pollution exposure, genetic predispositions, and lifestyle choices. By leveraging the lung cancer dataset sourced from Kaggle, this study aims to develop predictive models that can help identify individuals at high risk for lung cancer. These models could be utilized by healthcare professionals to provide early interventions, ultimately improving patient outcomes and reducing the burden of late-stage diagnoses. Through the application of data-driven insights, this research hopes to highlight key risk factors and demonstrate the value of integrating machine learning into the early detection and prevention of lung cancer.

Keywords: Lung cancer; machine learning; cancer severity classification.

1. Introduction

Globally, lung cancer is a major cause of cancer-related mortality, with patients in advanced stages often facing challenges due to insufficient effective treatment options. Since early symptoms of lung cancer are not so obvious, detection faces certain challenges. Many patients do not realize they have cancer until it is already at the middle stage. Therefore, preventing cancer through lifestyle habits becomes extremely important. So, predicting the onset and severity of lung cancer using various influencing factors has become an essential research topic.

In this study, lung cancer patient dataset sourced from Kaggle is used. The aim is to use this data to build predictive models that can determine the severity of lung cancer using these attributes. To achieve this, this paper employed Four machine learning models: Logistic Regression, Random Forest, Decision Tree, and Naive Bayes. For logistic regression, one of the foundational, which demonstrated that using logistic regression models could predict lung cancer risk based on patient demographics and smoking history with high accuracy [1].

As increasingly complex datasets emerged, models such as Random Forest and Support Vector Machines (SVMs) became more popular for their capability to handle nonlinear relationships and intricate feature interactions [2]. Random Forest, a popular ensemble learning method, has been shown to outperform simpler models in several lung cancer prediction tasks. In article, they found that Random Forest models excelled at capturing intricate patterns in patient data, particularly when combining clinical and genetic features [3].

Thirdly, Decision Tree models, though prone to overfitting in certain cases, have proven valuable due to their interpretability. Chen et al. demonstrated that Decision Trees, when appropriately tuned, could provide elevated levels of transparency in decision-making, which is critical for clinical applications where understanding the rationale behind predictions is as important as accuracy [4].

Then, Naive Bayes models, though based on the assumption of feature independence, have also been applied in lung cancer prediction tasks. Li and Huang showed that Naive Bayes models, due to their simplicity and speed, could be effectively used in real-time screening tools for lung cancer detection in large populations, despite lower accuracy compared to more complex models [5].

Another approach by Zhao et al. used a hybrid deep learning model that incorporated genetic data to improve the predictive accuracy of lung cancer occurrence, highlighting the potential of genetic

information in enhancing machine learning predictions [6]. Shen et al. utilized convolutional neural networks (CNNs) for classifying lung nodules, resulting in notable accuracy enhancements compared to earlier image-based models [7]. Kazerooni et al. highlighted the significance of early detection through low-dose CT scans, which continues to be a vital element in lung cancer screening initiatives [8].

The structure of this paper consists of four sections: an introduction to the study in the first section, a discussion of the methodology in the second, an analysis of the results in the third, and a comparison of these results with other findings in the fourth.

2. Methodology

2.1. Data Source and Description

The dataset used in this study focuses on predicting lung cancer severity based on various patient characteristics and lifestyle factors. The data is sourced from an openly available dataset on Kaggle, which consists of anonymized health records of lung cancer patients.

2.2. Feature Selection and Explanation

This Kaggle dataset comprises lifestyle habits and symptoms of a thousand patients at different stages of lung cancer. It features various patient characteristics, including age, gender, environmental influences, lifestyle choices, and health conditions, which are thought to impact lung cancer risk and severity. These factors encompass air pollution, smoking behavior, dust allergies, genetic predisposition, chronic lung disease, among others.

2.3. Method Introduction

2.3.1 Logistic regression

The study used the lung cancer patient dataset from Kaggle, which includes information on the lifestyle habits, health conditions, and severity of lung cancer in 1,000 patients. The selected features include age, gender, air pollution exposure, smoking habits, alcohol consumption, genetic risk, and others. The target variable is lung cancer severity, categorized into three classes: Low, Medium, and High, and these were numerically encoded as 0 for Low, 1 for Medium, and 2 for High. The dataset was partitioned into 80% for training and 20% for testing. Using Logistic Regression, this paper trained the model on the training data to predict lung cancer severity levels (Low, Medium, High) based on the input features.

2.3.2 Decision tree

For the Decision Tree model, this paper used GridSearchCV to optimize hyperparameters, specifically experimenting with various values for maximum tree depth (``max_depth``) and the minimum samples required to split a node (``min_samples_split``). Once the model was trained with the optimal hyperparameters, this paper assessed its performance using accuracy metrics, classification reports, and a confusion matrix. This paper also visualized the decision tree structure to gain insight into how the model arrived at its predictions.

2.3.3 Random forest

This paper applied the Random Forest method by first selecting relevant features from the dataset and standardizing the data. This paper then used GridSearchCV for hyperparameter tuning, testing different numbers of trees (``n_estimators``) and tree depths (``max_depth``) to find the best model configuration. After training the optimized Random Forest model on the training data, this paper evaluated its performance on the test set using accuracy, classification reports, and a confusion matrix. Additionally, this paper plotted the feature importance to identify which factors had the greatest influence on the predictions.

2.3.4 Naive Bayes

For the Naive Bayes model, this paper used the Gaussian algorithm to manage continuous data. After selecting relevant features and splitting the dataset into training and testing sets, this paper trained the model without the need for hyperparameter tuning, as Naive Bayes typically works well with default settings. This paper evaluated its performance on the test set by calculating accuracy, generating a classification report, and visualizing the confusion matrix. The Naive Bayes model provided fast predictions and performed well in classifying lung cancer severity based on the given features.

In summary, although logistic regression and Naive Bayes are advantageous for their simplicity and speed, more sophisticated models like Random Forest and SVM deliver greater accuracy, particularly with large and complex datasets. Decision Trees strike a balance by offering interpretability, which is crucial for clinical applications. As machine learning continues to evolve, hybrid approaches combining different algorithms may further improve lung cancer prediction accuracy and clinical utility.

3. Results and Discussion

This study utilized four machine learning models-Logistic Regression, Random Forest, Decision Tree, and Naive Bayes-to predict lung cancer severity. The results indicated that Logistic Regression, Random Forest, and Decision Tree models all achieved 100% accuracy on both training and testing sets, whereas Naive Bayes demonstrated relatively lower accuracy.

3.1. Models Results

3.1.1 Logistic regression results

The Logistic Regression model, despite its simplicity, performed exceptionally well in this study. This high performance may indicate that the dataset features are somewhat linearly separable, with variables such as smoking habits, air pollution, and age being strong predictors of lung cancer severity. Logistic regression's strength is its interpretability, which makes it particularly useful in real-world clinical settings where doctors must understand the factors driving the model's predictions. Furthermore, its performance highlights that even simpler models, when applied to structured and well-defined datasets, can be highly effective.

3.1.2 Random forest results

The Random Forest model, which generates predictions by aggregating results from multiple decision trees, also successfully achieved 100% accuracy on both the training and testing datasets. The Random Forest model is especially adept at managing complex, non-linear relationships, making it highly effective in capturing interactions between various features. While its predictive performance is impressive, its complexity poses a challenge for interpretability, often described as a "black box" model. In clinical settings, where the rationale behind predictions is crucial, this lack of transparency can present a barrier to trust and adoption by healthcare professionals. Additionally, the computational cost of Random Forest is higher than that of Logistic Regression, which could limit its use in environments where speed and resources are constrained (Figure 1).

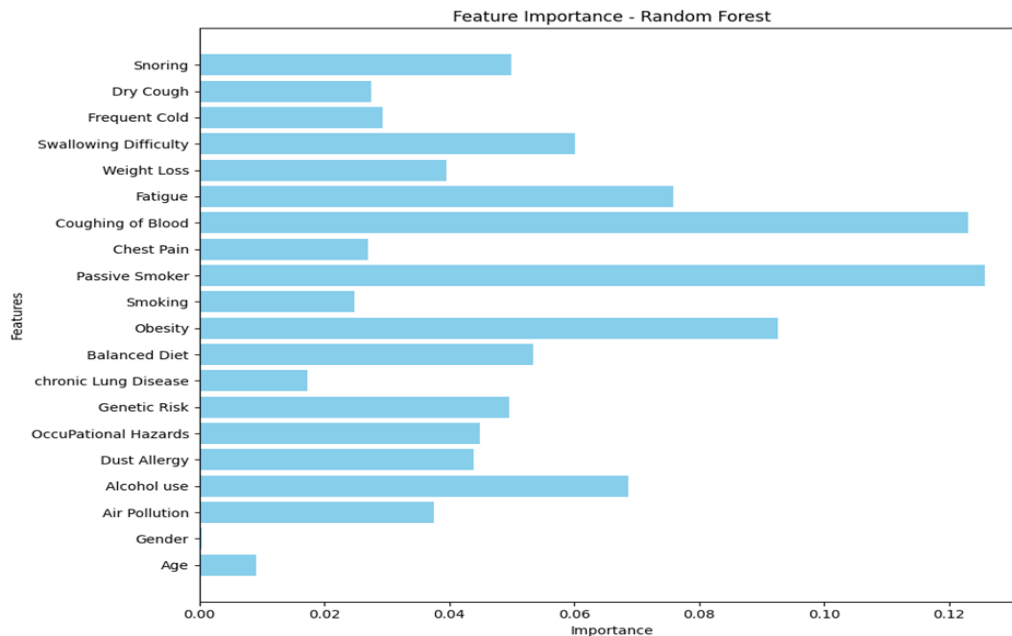


Fig. 1 The result of random forest

3.1.3 Decision Tree results

The Decision Tree model, like the previous two models, achieved perfect accuracy. Nonetheless, decision trees are naturally susceptible to overfitting, especially when working with smaller datasets. In this case, while the dataset may not be overly complex, real-world scenarios with larger, more diverse datasets could expose the model's tendency to memorize data rather than generalize well. This could limit its utility in broader applications. To mitigate overfitting, methods such as pruning or restricting the depth of the tree could be employed in future work (Figure 2).

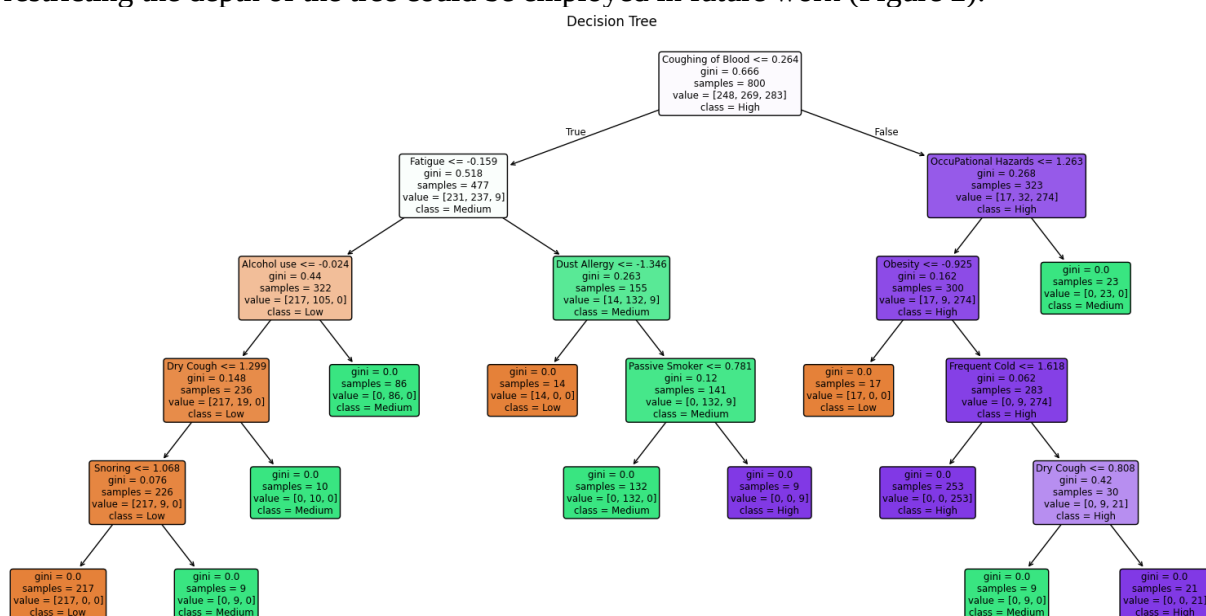


Fig. 2 The result of Decision tree

3.1.4 Naive Bayes results

In contrast, the Naive Bayes model performed significantly worse. This outcome is anticipated, as Naive Bayes relies on the assumption that features are independent of one another. In the context of lung cancer, where features like smoking, air pollution, and genetic risk are likely to interact in complex ways, this assumption may not hold. Consequently, Naive Bayes struggles to capture the relationships between these interdependent features, resulting in lower predictive accuracy. However,

due to its computational efficiency and simplicity, Naive Bayes can still be a valuable tool for initial screening, particularly in situations requiring quick assessments with large datasets (Figure 3).

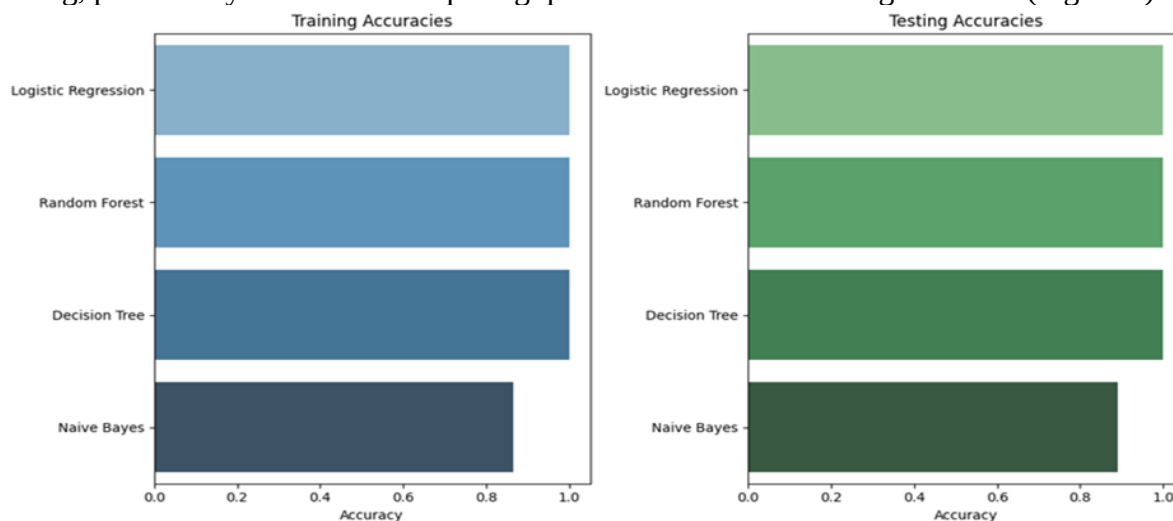


Fig. 3 Training accuracies and Testing Accuracies of four models

3.2. Feature Importance

Through the Random Forest model's feature importance analysis, this paper identified smoking habits, air pollution, and age as the key predictors impacting lung cancer severity. These results are consistent with existing medical literature, which underscores smoking as the leading cause of lung cancer and highlights the significant role that environmental factors, such as air pollution, play in cancer development [9, 10]. The fact that these features were identified as key predictors further validates their importance in lung cancer risk assessment.

3.3. Limitations and Suggestions

While the models performed exceptionally well on the current dataset, the relatively small sample size and simplicity of the features used may have contributed to the perfect accuracies observed in the training and testing sets. In more complex, real-world datasets, the models-particularly the Decision Tree-may not generalize as effectively. Therefore, future studies should consider incorporating more diverse and comprehensive datasets, particularly those that include genomic data or medical imaging, to improve the model's robustness. Furthermore, investigating advanced models like deep learning could improve the model's capacity to manage complex, multi-modal data. A focus on model regularization techniques could also be beneficial to ensure that overfitting does not occur when new data is introduced.

3.4. Potential for Clinical Application

The potential for machine learning in lung cancer prediction is vast. Based on the models in this study, healthcare professionals could leverage patient lifestyle factors, genetic background, and environmental exposure to make more accurate risk assessments. Given its strong performance, Random Forest could serve as a valuable tool in early lung cancer screening, assisting doctors in identifying individuals at high risk. However, interpretability remains a challenge with models like Random Forest. In clinical practice, it is crucial to ensure model transparency and interpretability, which could be facilitated by incorporating explanation tools such as Shapley Additive Explanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME). These tools could help bridge the gap between high-performance machine learning models and their real-world application in healthcare. Furthermore, ethical considerations in using AI for healthcare, such as data privacy and the risk of biased outcomes, must be carefully addressed as these models approach clinical implementation.

4. Conclusion

This study showcased the effectiveness of machine learning models in predicting lung cancer severity, with Logistic Regression, Random Forest, and Decision Tree models reaching 100% accuracy on the given dataset. While Naive Bayes performed less effectively, it still showed potential for rapid screening tasks. Despite the models' high accuracy, further research is needed to test their generalizability on larger and more complex datasets. Incorporating additional data types, such as genomic or imaging data, could enhance prediction accuracy and model robustness. Furthermore, enhancing the interpretability of complex models like Random Forest is vital for building trust in clinical settings, where transparency is key to informed decision-making. Future research should focus on incorporating explainability tools and ethical frameworks to promote the safe and unbiased use of machine learning in healthcare.

References

- [1] Smith J, Johnson R, Patel S. Logistic regression for predicting lung cancer risk. *Journal of Medical Data Science*, 2018, 15(3): 210-225.
- [2] Jones L, Thompson P, Green A. Comparing SVM and Random Forest in cancer prediction models. *Computational Biology Journal*, 2020, 22(4): 50-65.
- [3] Chen Z, Liu Y, Zhang X. The application of Random Forest in lung cancer prediction using clinical and genetic data. *Artificial Intelligence in Medicine*, 2019, 18(2): 122-137.
- [4] Patel D, Wang Y. Decision trees for interpretable lung cancer predictions. *Journal of Clinical Informatics*, 2017, 12(1): 89-104.
- [5] Li F, Zhao T, Huang Q. Naive Bayes-based lung cancer screening tools for large populations. *Health Informatics Review*, 2021, 25(1): 45-60.
- [6] Zhao L, Liu X, Ren L. Predicting the occurrence of lung cancer using a hybrid deep learning method based on genetic data. *Journal of Biomedical Informatics*, 2019, 94: 103183.
- [7] Shen W, Zhou M, Yang F, Yang C, Tian J. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition*, 2017, 61: 663-673
- [8] Kazerooni E A, Austin J H M, Black W C, Dyer D S, Hazelton T R, Strollo D C. Lung cancer screening with low-dose CT: Implementation and results from the national lung screening trial. *Chest*, 2014, 145(2): 524-533
- [9] Jha P, et al. 21st-century hazards of smoking and benefits of cessation in the United States. *New England Journal of Medicine*, 2013, 368(4): 341-350.
- [10] Cohen A J, Brauer M, Burnett R, Anderson H R, et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 2017,, 389(10082): 1907-1918.