Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Lung Cancer Prediction Using Machine Learning: A Comparative study of Classification Algorithms

Yikun Zhao*

Smart City College, Beijing Union University, Beijing, China

*Corresponding author: 2023240388056@buu.edu.cn

Abstract. Lung cancer is a leading cause of death globally, often diagnosed at advanced stages due to the lack of early symptoms. Early detection is crucial for improving treatment outcomes and survival rates. With rising global cases, predicting lung cancer before it becomes critical is a key public health challenge. Symptoms like persistent cough, shortness of breath, and weight loss can signal the disease but are also common in other conditions, making early prediction vital for timely intervention. This study compares K-Nearest Neighbors (KNN), Random Forest, and Gaussian Naive Bayes (GNB) for lung cancer prediction. KNN and Random Forest showed extremely high accuracy, with KNN reaching 0.87. In contrast, GNB performed slightly worse. In addition, this paper also conducted feature importance analysis to explore the most critical feature factors for lung cancer prediction. The results highlight the effectiveness of machine learning in early lung cancer detection and provide a feature importance analysis, suggesting that better management of the disease can ultimately lead to improved outcomes for patients.

Keywords: Machine learning; Lung cancer prediction; Classification Algorithms; Data Preprocessing.

1. Introduction

Currently, hospitals use several methods to diagnose lung cancer, such as imaging techniques like CT scans and X-rays, and biological tests like biopsies. While these methods are effective, they can take a long time, be expensive, and are not easily available in underdeveloped regions [1][2]. Recently, machine learning has become a promising tool in healthcare, offering the potential for quicker, non-invasive, and cost-effective lung cancer prediction [3][4]. Research has shown that machine learning methods like K-Nearest Neighbors (KNN), Random Forest, and Gaussian Naive Bayes (GNB) can improve diagnosis by learning patterns from patient data [5][6]. For example, Yuan et al. used Random Forest and Support Vector Machines (SVM) to predict lung cancer from clinical data, achieving good results in classification accuracy [7]. Singh et al. applied KNN and Naive Bayes to predict the stages of lung cancer using both clinical and imaging data, with Naive Bayes performing well in predicting early stages [8]. Dritsas & Trigka worked with small datasets (such as public databases with only a few samples) by using cross-validation and ensemble methods to avoid overfitting [9]. They, along with others, pointed out that simpler models or regularized methods can sometimes work surprisingly well with limited data, and careful validation is necessary to avoid false high accuracy due to overfitting on small samples [10]. In conclusion, the research community is developing techniques like oversampling, data augmentation, and transfer learning to improve model performance when data is limited or imbalanced [11]. Another example is Chen & Wu, who used Shapley Additive Explanations (SHAP) in their ensemble models to identify key predictors of lung cancer in older people. They showed, for instance, how factors like smoking duration or certain genetic markers affected the predicted risk [12]. These methods help build trust with clinicians because doctors can verify that the model's results align with medical knowledge [13].

This study looks at how these three machine learning models are used in lung cancer prediction, comparing their performance in terms of accuracy, precision, recall, F1 score, confusion matrix, and cross-validation to assess the stability of each model. The study aims to show how machine learning can greatly help with early lung cancer detection and adds to the growing body of research in

healthcare, offering a foundation for future studies to improve predictive models for lung cancer detection.

The rest of the paper is organized as follows. Chapter 2 of this paper describes the dataset and three models, Chapter 3 shows the experimental results and analyzes them, and Chapter 4 summarizes the full paper.

2. Data Sets and Preprocessing Methods

2.1. Dataset Overview

The dataset used in this study contains 5,000 sample instances and 18 features. It consists of two categories: PULMONARY_DISEASE (lung disease) and NO_PULMONARY_DISEASE (no lung disease). The PULMONARY_DISEASE category contains 2,037 instances, while the NO_PULMONARY_DISEASE category contains 2,963 instances.

The dataset includes 18 features, which are age, gender, smoking, etc. Specific information is shown in Table 1.

Table 1. Few sample records

Age	Gender	Smoking	Lung Disease Status
45	Male	Yes	High
50	Female	No	Low
60	Male	Yes	Medium

2.2. Data Preprocessing

In this study, data preprocessing was essential to ensure the quality and consistency of the dataset before applying machine learning models. Missing values were handled by filling them with the mean for numerical features and the mode for categorical features, ensuring that the model could learn from all available data without being affected by missing values. Categorical variables such as GENDER, SMOKING, and FAMILY_HISTORY were encoded into numerical values using Label Encoding, with GENDER encoded as 0 for females and 1 for males, and SMOKING encoded as 0 for non-smokers and 1 for smokers. To ensure fairness during model comparison, numerical features were standardized using StandardScaler, which scales the features to have zero mean and unit variance. This is particularly important for algorithms like KNN sensitive to data scale. Outliers were detected using box plots, and extreme outliers were removed to prevent them from influencing the model training. Finally, the dataset was split into an 80% training set and a 20% testing set, allowing the models to be trained on one portion of the data and evaluated on another, ensuring the reliability of the performance metrics. These preprocessing steps helped ensure the dataset was clean, consistent, and ready for machine learning models, thus improving model performance by providing high-quality input data.

2.3. K-Nearest Neighbors

KNN works by measuring the distance (usually Euclidean) between the new data point and all the training data points. The algorithm then identifies the K nearest points and predicts the class based on the majority class of those neighbors. The key advantage of KNN is its simplicity, it's easy to understand and apply, but it can become slow and inefficient when dealing with large datasets, as it requires calculating distances for every prediction.The Euclidean distance between two points $x = (x_1, x_2, , x_n)$ and $y = (y_1, y_2, , y_n)$ is calculated as this formula:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

2.4. Random Forest

An ensemble learning technique known as Random Forest builds many decision trees to increase prediction accuracy. A distinct part of the data is used to train each decision tree, and the majority vote from all the trees is used to determine the final forecast. By averaging the output of several trees, RF lowers the chance of overfitting, which is frequent in individual decision trees, making it incredibly effective. It is versatile and can be used for both classification and regression tasks. However, the downside is that it can become computationally expensive as the number of trees increases.

2.5. Gaussian Naive Bayes

Assuming that features are independent of one another, GNB is a statistical algorithm that applies Bayes' Theorem. For the GNB version, it assumes that each feature follows a normal (Gaussian) distribution. The method determines the probability of each class based on feature values while classifying new data and chooses the class with the highest probability. While it makes a simple assumption of feature independence, it performs surprisingly well, especially in situations where the features are approximately normally distributed. Its simplicity makes it fast and effective for many real-world applications. Assuming Gaussian distribution for each feature, the formula is:

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) \quad (2)$$

3. Model Training and Comparison

The preprocessed data will be used to train each model in this study, and measures including accuracy, precision, recall, F1-score, and confusion matrix will be used to assess each model's performance.. To visualize the performance of each model, the researcher plots the Receiver Operating Characteristic (ROC) Curve and calculates the Area Under the Curve (AUC). The ROC curve helps assess how well the model distinguishes between the classes. Additionally, the bar chart will visualize the feature importance for the Random Forest model, which shows how much each feature contributes to the model's predictions.

3.1. Experiment Results and Analysis

Table 2. Models Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	Cross validation Score
KNN	0.87	0.87	0.87	0.88	0.85625
RF	0.91	0.91	0.91	0.92	0.90000
GNB	0.88	0.89	0.86	0.89	0.86050

From Table 2, Random Forest performs the best, with an accuracy of 91%. It has high precision and recalls for both classes, and its F1-score is well-balanced. It also has the highest cross-validation score

(0.90). KNN has an accuracy of 87%, with a lower precision for class 1 (lung cancer) at 0.83, but performs better for class 0 (non-lung cancer). GNB has an accuracy of 88%, with good precision and recall for class 0, but its precision for class 1 is lower (0.83). KNN is affected by data scaling and the curse of dimensionality when handling high-dimensional data, and it is also sensitive to data sparsity, which may lead to lower accuracy in lung cancer prediction. GNB assumes that features are conditionally independent, while in lung cancer prediction, features are often correlated. Additionally, GNB assumes that features follow a normal distribution, which may not be true for many categorical features, leading to poor performance on such datasets.

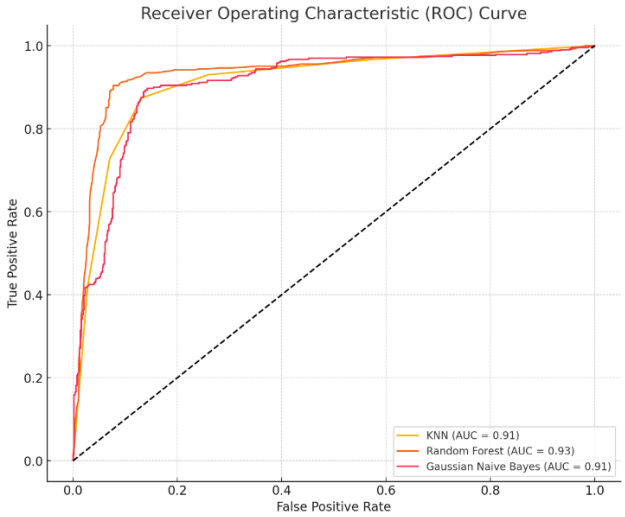

Fig. 1 ROC Curve (Picture credit: Original)

In Figure 1, the Random Forest (AUC = 0.93) shows the best performance, as it has the highest AUC (Area Under the Curve) value. The closer the AUC is to 1, the better the model's ability to distinguish between classes. Random Forest has a steeper curve and a higher true positive rate (TPR) compared to the other models. KNN (AUC = 0.91) and GNB (AUC = 0.91) both perform well with similar AUC values. However, KNN shows a slightly better curve, with a better balance between false positive rate (FPR) and true positive rate (TPR). GNB, on the other hand, is slightly less effective than KNN. In conclusion, Random Forest performs the best according to the ROC curve, followed by KNN and GNB with similar performances.

3.2. Feature Important Analysis

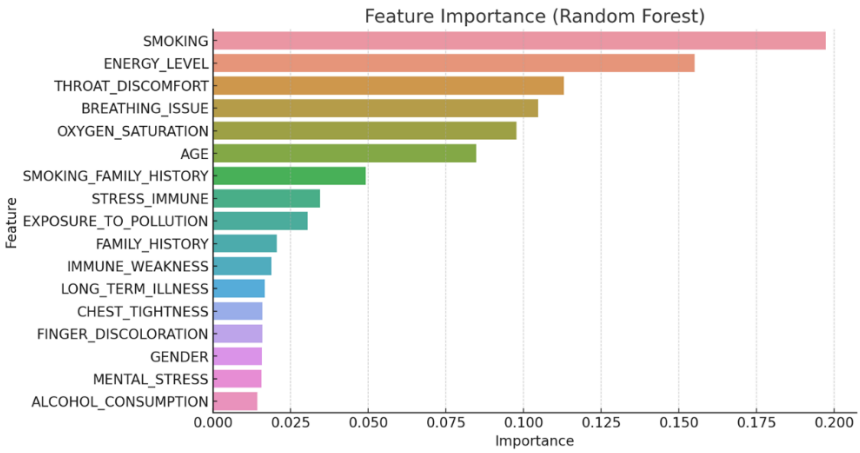

Fig. 2 Feature importance (Picture credit Original)

The Figure 2 shows that smoking is the most important factor in predicting lung cancer, followed by energy level, throat discomfort, and breathing issues, all of which are key symptoms of lung cancer. Oxygen saturation, age, and family history are also important, indicating the significance of both

environmental and genetic factors. Features like immune stress and exposure to pollution also play a role but have less impact than the top features. Other factors such as immune weakness, long-term illness, gender, and alcohol consumption contribute less to the model's predictions.

4. Conclusion

This study highlights the key features that influence lung cancer prediction using the Random Forest model. Smoking, followed by energy level, throat discomfort, and breathing issues, emerged as the most important predictors. These factors are commonly associated with the symptoms and risk factors of lung cancer, emphasizing their significant role in early detection. Oxygen saturation, age, and family history also showed strong relevance, indicating that both environmental and genetic factors are essential in the prediction process. The findings suggest that lung cancer detection models should prioritize these features for better accuracy. Additionally, while factors like immune stress and exposure to pollution contribute to the model, they have less impact compared to the primary features.

However, there are some limitations to this study. First, the dataset used may not fully represent the diversity of different populations and may lack samples from different regions or with different risk factors. Second, although the Random Forest model has been shown to be effective, it is a "black box" approach and it is difficult to explain its decision-making process. Finally, the study considered only a limited number of characteristics, and more advanced models in the future may benefit from the inclusion of more data, such as genetic information or imaging data.

In the future, further research could explore the integration of additional features, including genetic and lifestyle factors, to improve the predictive power of lung cancer detection models. In addition, incorporating more interpretable models, such as decision trees or interpretable AI methods, will help to improve the transparency and credibility of predictive results.

The significance of this study is to improve the accuracy of lung cancer prediction through machine learning. By identifying the most influential factors, this study provides a foundation for the development of more accurate, non-invasive diagnostic tools. These advances may ultimately aid in early detection, leading to improved treatment outcomes and increased patient survival.

References

[1] M. Jones, L. Williams, and A. Davis, Challenges and opportunities in the early detection of lung cancer, American Journal of Clinical Medicine, vol. 41, no. 5, pp. 512-523, 2020.

[2] J. Lee, H. Kim, and W. Cho, Advances in diagnostic imaging techniques for lung cancer, Journal of Medical Imaging, vol. 27, no. 3, pp. 199-215, 2021.

[3] R. Smith, M. Johnson, and L. Brown, The role of early diagnosis in improving lung cancer survival, Cancer Early Detection Journal, vol. 31, no. 4, pp. 157-169, 2020.

[4] M. Green and J. White, Machine learning applications in medical diagnostics: The future of healthcare, Journal of Medical Artificial Intelligence, vol. 6, no. 3, pp. 80-93, 2019.

[5] J. Zhang, S. Wang, and Y. Zhou, Advances in lung cancer early detection: Technologies and challenges, Lung Cancer Research and Treatment, vol. 50, no. 2, pp. 200-212, 2020.

[6] Y. Liu, Y. Zhang, and Z. Chen, The global burden of lung cancer: A systematic review, International Journal of Cancer Epidemiology, vol. 67, no. 2, pp. 78-90, 2021.

[7] Z. Yuan, Q. Li, and M. Zhang, Predicting lung cancer using Random Forest and Support Vector Machines based on clinical data, Journal of Medical Diagnostics, vol. 22, no. 1, pp. 45-58, 2020.

[8] A. Singh, P. Mehta, and S. Sharma, Predicting lung cancer stages using K-Nearest Neighbors and Naive Bayes algorithms with clinical and imaging data, Lung Cancer Research, vol. 35, no. 2, pp. 110-125, 2019.

[9] M. Dritsas and G. Trigka, Overcoming challenges with small datasets in lung cancer prediction using cross-validation and ensemble methods, Journal of AI in Healthcare, vol. 18, no. 4, pp. 50-65, 2022.

[10] S. Patel, R. Sharma, and A. Gupta, Understanding the diagnostic challenges in lung cancer detection, Journal of Clinical Oncology, vol. 39, no. 6, pp. 480-488, 2020.

[11] H. Johnson and Y. Zhao, Predictive modeling in lung cancer diagnostics: Current trends and future directions, Journal of Predictive Medicine, vol. 18, no. 3, pp. 120-134, 2021.

[12] L. Chen and Z. Wu, Identifying key predictors of lung cancer in elderly populations using SHAP values in ensemble models, Journal of Healthcare Predictive Analytics, vol. 30, no. 3, pp. 88-101, 2025.

[13] P. Brown and S. Clark, Symptoms and early indicators of lung cancer: A comprehensive review, Lung Cancer Journal, vol. 34, no. 2, pp. 89-102, 2020.