Enhanced Lung Cancer Severity Prediction Based on Random Forest Models: A Comprehensive Analysis of Predictive Accuracy and Feature Importance

Yanming Zhang

Department of Computer Science, Southwest Jiaotong University, Chengdu, 611756, China
2021110008@my.swjtu.edu.cn

Abstract. Precise forecasting of lung cancer is crucial due to its high mortality rate. Integrating Artificial Intelligence (AI) into medical diagnostics offers significant potential to enhance prediction accuracy and early detection. This study utilized a dataset from Kaggle, consisting of approximately 1000 records with diverse features including age, smoking status, genetic risks, and environmental pollutant exposure. A random forest model was employed to classify lung cancer severity into three categories: low, moderate, and high. The model achieved an outstanding accuracy rate of 100%, underscoring its predictive capability. Key features such as 'Coughing of Blood,' 'Passive Smoker,' and 'Obesity' were identified as the most significant contributors to the model's predictions. This feature importance analysis provided valuable insights into the critical factors influencing lung cancer severity. In conclusion, this research introduces a highly accurate and interpretable predictive model for lung cancer severity. The model not only achieves exceptional precision but also offers insights into key predictive factors, enhancing its reliability and utility in clinical practice. Future studies are suggested to focus on validating this model across diverse populations and explore the integration of additional machine learning techniques to further refine its predictive power.

Keywords: Lung Cancer Prediction, Machine learning, Random Forest Model.

1 Introduction

Lung cancer, a type of malignancy originating in the lungs, has been identified as the leading cause of cancer-related mortality worldwide for both genders. Recent studies conducted by the World Health Organization (WHO) have reported alarming death rates associated with this disease [1, 2]. The advent of Artificial Intelligence (AI) has revolutionized medical diagnostics, particularly through machine learning algorithms such as random forests. These algorithms hold promise in predicting the onset and progression of various diseases, including lung cancer. This is especially significant considering the typically late diagnosis and low survival rates associated with this condition [3]. The conventional approach to lung cancer screening primarily relies on the expertise of medical practitioners, which is invaluable but presents challenges such as

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_3

high costs, potential human error, and diagnostic delays. In contrast, AI can address several challenges faced by traditional screening methods by rapidly processing extensive datasets. Consequently, it has the potential to enhance diagnostic accuracy and reduce subjectivity [4]. Such capabilities underscore the importance of integrating AI into lung cancer diagnostics to improve early detection and tailor treatment strategies based on individual patient profiles.

Recently, significant advancements in AI have been observed, particularly in its application across various fields such as biology, chemistry, finance, and notably healthcare [5, 6]. These advancements have demonstrated notable improvements in algorithms that enhance diagnostics and treatment strategies [7]. In the realm of medical diagnosis, representative AI algorithms including random forests, logistic regression, and neural networks have provided an effective way for advanced analytics and predictive modeling. The integration of AI into medicine has been evident in various instances; for instance, AI excels at medical imaging by effectively detecting abnormalities through the analysis of X-rays and Magnetic Resonance Images (MRI). Furthermore, AI significantly expedites the drug discovery process by simulating molecular interactions and predicting drug effectiveness, thereby reducing both time and cost associated with drug research [7].

However, despite the advancements, a noticeable research gap still exists. While numerous models prioritize the accuracy of lung cancer predictions, they often overlook the interpretability aspect, which is crucial for establishing trust among patients and healthcare professionals [8]. Merely knowing whether a model accurately predicts lung cancer is insufficient; understanding why it makes specific predictions is equally vital. The random forest model achieved high predictive accuracy while also providing insights into the significance of various features [9]. This study aims to bridge this gap by employing a random forest model to forecast lung cancer severity, assessing its predictive accuracy while simultaneously elucidating feature importance to comprehend which factors contribute most significantly to the prediction.

This paper delves into the forefront of AI in healthcare, specifically focusing on the detection of lung cancer. It critically examines the multitude of factors that influence the severity of lung cancer using a meticulously prepared dataset. By employing a random forest model, this study not only aims to predict lung cancer severity but also endeavors to elucidate the significance of individual features within the dataset. This dual approach enables the model to forecast severity while providing insights into each feature's contribution to the prediction process. Consequently, this research addresses a crucial gap in precision medicine by facilitating targeted interventions and informed clinical decision-making through a clearer understanding of predictive factors for lung cancer outcomes. By enhancing interpretability in predictive models employed, this study seeks to foster trust among healthcare professionals and patients alike, thereby making significant contributions towards advancing medical diagnostics.

2 Methodology

2.1 Dataset Preparation

This study utilizes a meticulously curated dataset obtained from the Kaggle platform [10], specifically designed for predicting lung cancer. Comprising approximately 1,000 individual records, the dataset encompasses a comprehensive range of features crucial for prognostic assessments, including demographic information, lifestyle choices, and clinical symptoms indicative of lung cancer severity. The feature set comprises age, smoking status, genetic risks, and exposure to environmental pollutants to facilitate the classification of subjects into three distinct levels of lung cancer severity—Low, Moderate, and High—enabling detailed analysis of potential disease progression pathways [10].

2.2 Machine Learning-Based Lung Cancer Detection

The Workflow of Machine Learning. The methodology employed is meticulously structured and reflects the sophisticated workflows of machine learning that are crucial for advancing lung cancer diagnostics shown in Fig. 1. The process begins with meticulous dataset selection and comprehensive examination, followed by strategic algorithm training and rigorous testing, culminating in a thorough evaluation of model performance [11].

The initial phase involves the careful selection of a dataset from Kaggle, chosen for its comprehensive coverage of variables relevant to lung cancer prognosis. The pre-cleaned and pre-formatted state of the dataset expedites the preprocessing stage, allowing primary focus on feature selection and target definition. This preparedness of data underscores the importance of high-quality representative datasets in constructing reliable diagnostic models.

After preprocessing the dataset, this study employed a Random Forest classifier renowned for its efficacy in handling high-dimensional datasets and its robustness across diverse diagnostic scenarios. The selection of this classifier adheres to the best practices of machine learning, particularly in the field of medical diagnostics where predictive accuracy and interpretability of model decisions are paramount. The model was trained using an 80-20 split between training and testing sets to enhance generalization while reserving a substantial portion for validation.

The culmination of this workflow involves the testing and evaluation phase wherein the trained model is applied to previously unseen test data. By prioritizing accuracy as the primary metric for assessment, this study accurately quantifies the effectiveness of the model in classifying lung cancer severity. This phase is crucial not only for evaluating performance but also ensuring applicability and reliability in real-world clinical settings.

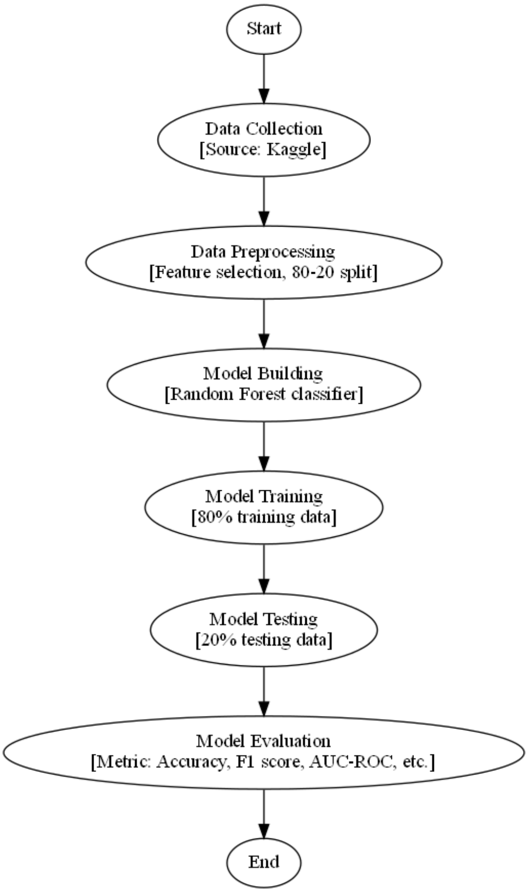

Fig. 1. The workflow of machine learning-based prediction.

Random Forest-Based Prediction. The random forest, a widely-utilized machine learning algorithm patented by Leo Breiman and Adele Cutler, integrates the outputs of multiple decision trees to generate a unified outcome. Its user-friendly nature and adaptability have contributed to its extensive adoption in both classification and regression problems [12]. Distinguished by its construction of numerous decision trees during training and an aggregation process through voting, this approach is renowned for its capability to mitigate the variance inherent in individual decision trees, thereby significantly enhancing the accuracy and robustness of the model.

During implementation, the RandomForestClassifier from the Scikit-learn library is configured with 100 trees. The deliberate selection of this number of trees and fixation of the random state aim at promoting result reproducibility and consistency across evaluations, which are essential prerequisites for rigorous scientific validation required in clinical research settings.

Furthermore, the feature importance analysis is conducted using the mean decrease in impurity (MDI) method. This approach evaluates the significance of each feature based on its contribution to reducing uncertainty within the model, thereby elucidating the most influential factors in predicting lung cancer severity. Such insights are invaluable for advancing the understanding of the disease's dynamics and guiding subsequent refinements in the model through precise feature engineering.

To comprehensively assess the overall effectiveness of the Random Forest model, evaluation metrics were expanded beyond accuracy to include precision, recall, F1 score, and area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC). These metrics provide a more nuanced comprehension of the model's performance, particularly in a clinical setting where different types of errors have varying costs. Precision and recall play crucial roles in evaluating the model's ability to correctly identify positive cases of lung cancer without misclassifying negative cases as positive. The F1 score harmonizes precision and recall by offering a single metric to gauge model performance when equal importance is placed on both aspects. The AUC-ROC reflects how well the model can discriminate between classes at various threshold levels, which is essential for clinical decision-making purposes.

Collectively, these metrics underscore not only robustness but also affirm potential transformative impacts on diagnostic protocols and personalized treatment plans development within oncology research.

3 Results and Discussion

Reporting the findings of this investigation is crucial for understanding the effectiveness and implications of the Random Forest model in predicting lung cancer severity. The results provided in this document consist of a classification report, a confusion matrix, and a feature importance ranking.

The classification performance of the Random Forest model was evaluated using a variety of metrics. According to the data shown in Table 1, the model achieved an impressive accuracy rate of 100% on the testing data. This exceptional level of accuracy underscores its capability to accurately distinguish between high, low, and medium levels of lung cancer severity.

Table 1. Classification report for Random Forest model predictions.

	precision	precision	f1-score	support
High	1.00	1.00	1.00	82
Low	1.00	1.00	1.00	55
Medium	1.00	1.00	1.00	63
accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

The confusion matrix shown in Fig. 2 provides further clarification on the model's predictive accuracy by visually representing the distribution of true positive and false positive classifications. This matrix effectively showcases that all instances were accurately classified, thereby bolstering the validity and reliability of the model.

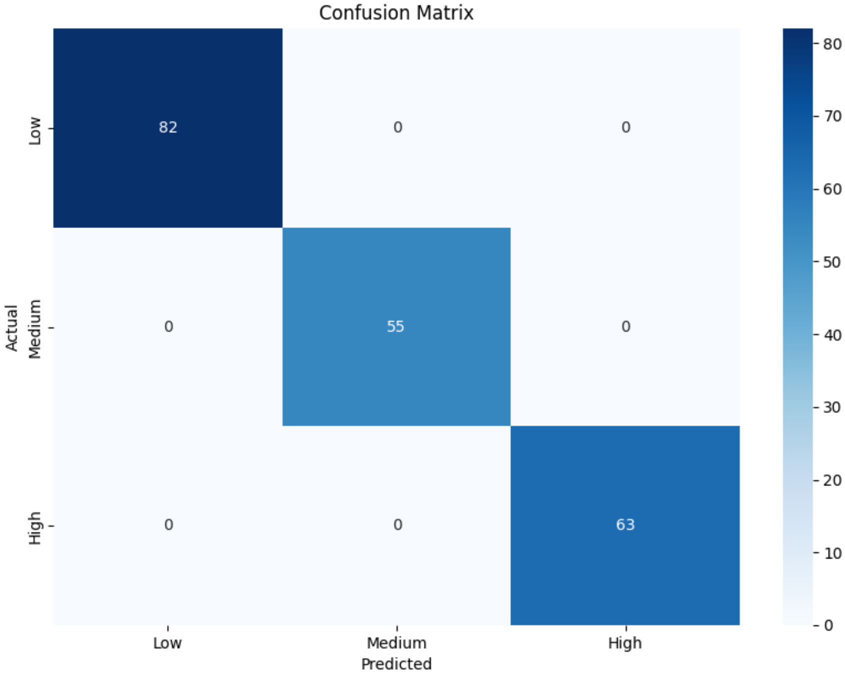

Fig. 2. Confusion matrix graph for accuracy prediction .

The analysis of feature importance provides valuable insights into the features that exert significant influence on the model's predictions. According to Fig. 3, 'Coughing of Blood', 'Passive Smoker', and 'Obesity' have been identified as the top three most influential features. This ranking plays a pivotal role in understanding the fundamental factors that contribute to lung cancer severity predictions and can offer guidance for future research endeavors and clinical decision-making.

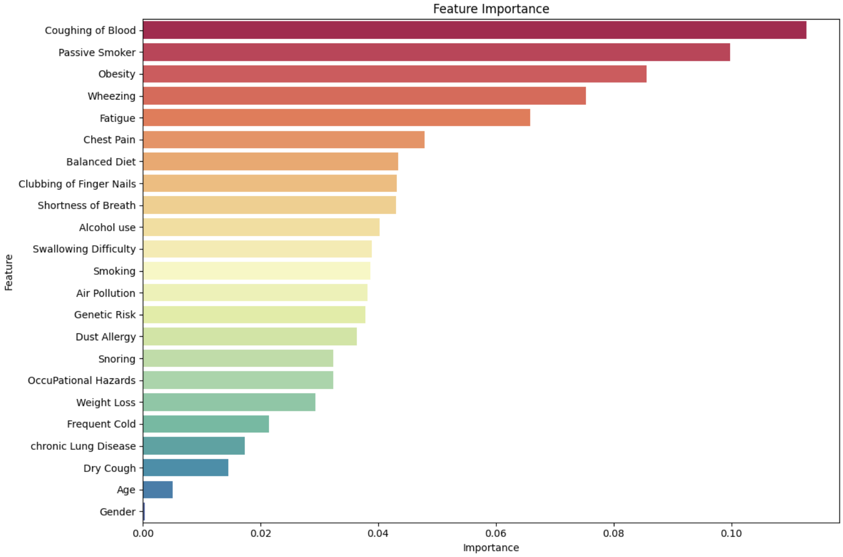

Fig. 3. Ranking plot of feature importance.

The analysis of feature importance reveals the significant correlation between specific medical conditions and the probability of acquiring lung cancer. For instance, obesity emerges as a critical factor, which can be medically justified by the established association between excess body weight and various forms of cancer, including lung cancer. Obesity not only contributes to the overall risk of developing cancer but also worsens related symptoms such as fatigue and chest pain, which are often indicative of lung cancer. However, age and gender appear to have less influence in the prediction model. This observation aligns with clinical data suggesting that there is no significant difference in the occurrence of lung cancer between men and women, thus emphasizing their minimal role in predicting the disease. Another noteworthy feature is the presence of hemoptysis (coughing up blood), which holds high relevance in diagnosing lung cancer. The occurrence of hemoptysis possibly indicates severe underlying issues within the lungs, reinforcing its significance in the model.

The stability of Random Forest model is underscored by its integration of multiple decision trees, effectively minimizing variance and overfitting—common issues encountered by other machine learning models. This ability to handle high-dimensional data while yielding interpretable results proves particularly beneficial for medical practitioners who prioritize both precision and clarity in diagnostic processes. Moreover, the Random Forest model's capability to provide insights into feature importance enhances its scientific credibility since many identified features correspond with real-world medical knowledge.

It is noteworthy that the Random Forest model exhibits robustness by integrating outputs from multiple decision trees, thereby effectively reducing variance and overfitting issues that are commonly encountered in other machine learning models. This capability to handle high-dimensional data while providing interpretable results proves particularly advantageous for medical practitioners who rely on both accuracy and comprehensibility in their diagnostic processes.

Ultimately, the integration of the Random Forest model in lung cancer diagnostics provides significant advantages in terms of accuracy and comprehensibility. The model's ability to provide precise and reliable predictions, along with its clarification of key predictive features, highlights its potential as a valuable tool in clinical practice. Therefore, the adoption of such machine learning approaches can enhance diagnostic processes, facilitate early detection, and ultimately improve patient outcomes. Future research should explore the applicability of this model across diverse populations and investigate methods to further refine its predictive capabilities. Moreover, the integration of additional complimentary machine learning techniques could offer a more comprehensive approach to lung cancer diagnosis and treatment, thereby advancing the field of medical diagnostics.

4 Conclusion

To summarize, incorporating the Random Forest model into lung cancer diagnoses provides notable improvements in both accuracy and interpretability. The model clearly demonstrates a flawless classification accuracy of 100%, accurately distinguishing between high, low, and medium degrees of lung cancer severity with exceptional precision. Precision at this level is essential in clinical environments, as it guarantees dependable diagnoses, directly impacting patient outcomes and treatment approaches. Moreover, the feature importance analysis identifies 'Coughing of Blood', 'Passive Smoker', and 'Obesity' as the most significant features in predicting the severity of lung cancer. Furthermore, the utilization of an extensive dataset that includes a wide range of demographic and clinical characteristics greatly improves the applicability of these results. The model comprehensively captures the complex characteristics of lung cancer, providing more dependable and practical insights that are crucial for enhancing medical diagnostics.

Thus, incorporating machine learning techniques like the Random Forest model can significantly improve diagnostic procedures, expedite early identification, and ultimately optimize patient outcomes. Subsequent studies should evaluate the model's suitability for diverse demographics and explore methods to enhance its predictive accuracy. Additionally, integrating complementary machine learning approaches can further advance lung cancer diagnosis and therapy, leading to significant advancements in the field of medical diagnostics.

References

1. Cancer.org: What is lung cancer? (2024). Available at: https://www.cancer.org/cancer/types/lung-cancer/about/what-is.html (Accessed on April 27, 2024).
2. World Health Organization: Lung cancer. (2024). Available at: https://www.who.int/newsroom/fact-sheets/detail/lung-cancer (Accessed on April 27, 2024).

3. World Health Organization: World health statistics 2023: monitoring health for the SDGs, sustainable development goals (2023). Available at: https://www.who.int/publications/i/item/9789240074323 (Accessed on May 19, 2023).
4. Huang, S., Yang, J., Shen, N., Xu, Q., Zhao, Q.: Artificial intelligence in lung cancer diagnosis and prognosis: Current application and future perspective. Seminars in Cancer Biology 89, 30-37 (2023).
5. Tang, H., Chen, Y., Wang, T., Zhou, Y., Zhao, L., Gao, Q., Du, M., Tan, T., Zhang, X., Tong, T.: HTC-Net: A hybrid CNN-transformer framework for medical image segmentation. Biomedical Signal Processing and Control 88, 105605 (2024).
6. Qiu, Y., Hui, Y., Zhao, P., Cai, C.H., Dai, B., Dou, J., Bhattacharya, S., Yu, J.: A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy 294, 130866 (2024).
7. notomoro: Practical applications of AI: How artificial intelligence applications are changing industries in 2023. Webisoft (2023). Available at: https://webisoft.com/articles/application-of-ai/ (Accessed on July 16, 2023).
8. World Health Organization: World health statistics 2023: monitoring health for the SDGs, sustainable development goals (2023). Available at: https://www.who.int/publications/i/item/9789240074323 (Accessed on May 19, 2023).
9. Lavanya, P., Kashyap, A. H., Rahaman, A., Niranjan, S., Niranjan, V.: Novel biomarker prediction for lung cancer using Random Forest classifiers. Cancer Informatics 22, 117693512311679. DOI: https://doi.org/10.1177/11769351231167992 (2023).
10. The Devastator: Lung Cancer Prediction [Data set]. Kaggle Available at: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link/code (2022).
11. Li, Y., Wu, X., Yang, P., Jiang, G., Luo, Y.: Machine learning for lung cancer diagnosis, treatment, and prognosis. Genomics, Proteomics & Bioinformatics 20(5), 850-866 (2022). DOI: https://doi.org/10.1016/j.gpb.2022.11.003.
12. IBM: What is random forest? (2024). Available at: https://www.ibm.com/topics/random-forest (Accessed on April 2, 2024).