Lung Cancer Prediction based on Machine Learning

Yuqi Zou *

Ningbo Foreign Language School, Ningbo, 315000, China

* Corresponding Author Email: EasonZ2008@outlook.com

Abstract. Lung cancer, a malignancy with high incidence and mortality rates, underscores the critical importance of early diagnosis and treatment. Traditional prediction methods possess limitations, whereas advancements in machine learning technologies offer novel avenues for lung cancer prediction. This paper utilizes data from public databases of lung cancer patients, employing various machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and Neural Networks (NN) to diagnose and predict lung cancer. The results indicate that the Random Forest model performs optimally, particularly achieving a high AUC value. Notably, smoking status, age, and yellowing of fingers are identified as crucial features for lung cancer prediction. This study provides new insights and methods for early diagnosis and treatment of lung cancer, possessing significant clinical implications and application value.

Keywords: Lung cancer; prediction; machine learning.

1. Introduction

Lung cancer, ranking among the top malignancies globally in terms of both incidence and mortality, poses a severe threat to human health [1, 2]. Due to its inconspicuous early symptoms, patients often find themselves in the advanced stages upon diagnosis, missing the optimal treatment window and often succumbing to the disease. Consequently, early diagnosis and treatment are paramount. Traditional lung cancer prediction methods, while offering some assistance in early detection, are limited. For instance, X-ray examinations have low sensitivity for early-stage lung cancer, leading to missed diagnoses and delayed treatments. Recently, the rapid development of machine learning technology has broadened its application in the medical field, especially in cancer probability assessment, disease diagnosis, and drug development. Machine learning algorithms can process vast amounts of medical data, uncover valuable information and patterns, and provide doctors with more accurate and personalized diagnostic and treatment suggestions [3, 4].

In lung cancer prediction, machine learning technologies have been extensively researched and applied in clinical practice. These techniques construct prediction models by analyzing clinical data, imaging information, and genetic profiles of patients to evaluate lung cancer risk [5]. For example, Sun et al. developed multiple lung cancer diagnostic models based on CT images of lung cancer patients, including SVM, and found that the SVM model achieved the best prediction performance with an AUC value of 0.94. Additionally, Yin et al. utilized Random Forest to predict survival outcomes in non-small cell lung cancer patients, demonstrating high prediction accuracy with an AUC value close to 0.8, indicative of stable prediction performance [6]. These studies not only validate the effectiveness of machine learning in lung cancer prediction but also lay a foundation for further research and applications [7, 8].

This paper aims to diagnose and predict lung cancer using machine learning algorithms, enhancing early diagnosis rates and prognostic prediction accuracy. Firstly, we collected clinical data, imaging information, and genetic profiles of lung cancer patients from public databases. Subsequently, we trained machine learning models with these data and evaluated their performance. Through optimizing model parameters and selecting algorithms, we aspired to construct efficient and accurate lung cancer prediction models. Furthermore, we explored the clinical application value of this model to provide more precise and personalized treatment plans for lung cancer patients. This study endeavours to offer novel ideas and methods for early detection and prognostic diagnosis of lung cancer, ultimately improving patients' survival rates and prognosis.

2. Data and Methodology

2.1. Data Sources and Preprocessing

This study aims to elevate the precision of lung cancer diagnosis and prognosis prediction through the application of machine learning algorithms. To achieve this goal, we have sourced clinical data, imaging materials, and genetic information of lung cancer patients from public databases. Specifically, the data utilized in this research was obtained from Kaggle, encompassing comprehensive records of 150 patients, including details such as gender, age, smoking habits, anxiety levels, and chronic disease status, among others.

Table 1. Information on the number of lung cancer patients (incomplete statistics)

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE
M	69	1	2	2	1
M	74	2	1	1	1
F	59	1	1	1	2
M	63	2	2	2	1
F	63	1	2	1	1
F	75	1	2	1	1
M	52	2	1	1	1
F	51	2	2	2	2
F	68	2	1	2	1
M	53	2	2	2	2

Below is Table 1, presenting a snapshot of the patient demographics and various attributes. Please note that only ten entries are showcased for brevity [9, 10].

Upon obtaining the raw data, we embarked on a series of preprocessing steps to ensure data quality and the validity of our analysis. Initially, we inspected the data for missing values. Given the relatively small dataset size and high feature completeness, observations containing missing values were promptly removed. Subsequently, we scrutinized and rectified any abnormal or erroneous records to maintain the integrity of the data, ensuring that all values fell within reasonable ranges (e.g., ensuring ages were non-negative and smoking status was accurately categorized). We then converted categorical variables into numerical formats to facilitate machine learning algorithm processing. For instance, gender was encoded as male=1 and female=0, while smoking habits were transformed into non-smoker=0, smoker=1, and heavy smoker=2.

Lastly, numerical features underwent standardization or normalization procedures to mitigate the influence of different scales during model training.

2.2. Models

To construct an efficient lung cancer prediction model, we selected and evaluated multiple machine learning algorithms, primarily including Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting Decision Trees (GBDT), and Neural Networks (NN). These algorithms excel in handling classification tasks and possess distinct characteristics:

SVM is a robust and versatile machine learning algorithm, it excels in tackling classification problems, particularly those involving high-dimensional data or nonlinear relationships. However, its computational cost may escalate notably with an increasing sample size, particularly when dealing with large-scale datasets.

RF is an ensemble learning method that constructs multiple decision trees and combines their predictions to enhance overall prediction accuracy and robustness. Nevertheless, it may falter in capturing complex interactions within certain datasets, potentially compromising prediction performance. Additionally, its model interpretability can be limited, particularly with a large number of decision trees, making it challenging to intuitively comprehend the prediction process.

GBDT is an ensemble learning algorithm utilizing decision trees as base learners. It iteratively constructs multiple weak learners to strengthen the overall model's predictive capabilities, apt for managing intricate data relationships.

3. Model Training and Evaluation

After preprocessing, model training and evaluation proceeded in four key steps:

Firstly, the preprocessed dataset is randomly divided into a training set, validation set, and test set. Typically, we allocate 70% as the training set, 15% as the validation set, and 15% as the test set. Then, the selected machine learning models are trained using the training set data, and model parameters are adjusted through cross-validation to optimize performance. Secondly, multiple evaluation metrics such as accuracy, precision, recall, F1 score, and AUC value are used on the validation set and test set to assess model performance. It is particularly important to note that in lung cancer prediction, the costs of misdiagnosis (i.e., diagnosing non-lung cancer patients as lung cancer patients) and missed diagnosis (i.e., diagnosing lung cancer patients as non-lung cancer patients) may differ, so it is necessary to select appropriate evaluation metrics based on the actual situation. Finally, the optimal model is selected based on its performance on the validation set, and the final performance verification is conducted on the test set.

4. Result Analysis

Table 2. Evaluation data of each model.

Models	Accuracy	Precision	Recall	F1 Score	AUC Value
SVM	0.95	0.9	0.97	0.92	0.96
RF	0.95	0.92	0.98	0.94	0.97
GBDT	0.90	0.86	0.92	0.88	0.93

From Table 2, it is evident that RF excels in all evaluated metrics, particularly achieving the highest AUC value of 0.97, demonstrating its superiority in lung cancer prediction tasks. SVM also performs well, albeit slightly inferior to RF and GBDT, and while impressive, falls short in recall and F1 score.

Furthermore, we computed the feature importance in the RF model. By comparing the importance scores of various features, we identified smoking status (SMOKING), age (AGE), and the presence of yellow fingers (YELLOW_FINGERS) as the most critical predictors of lung cancer. This finding aligns closely with clinical medical knowledge, reinforcing the notion that smoking is a primary risk factor for lung cancer, often manifesting as yellow fingers among long-term smokers.

To delve deeper into the influence of individual features on lung cancer prediction, we conducted an in-depth analysis of the key features:

Emerged as the most influential feature, emphasizing the strong link between smoking and lung cancer risk. Our analysis reveals that the longer the smoking duration, the higher the risk of developing lung cancer. Consequently, regular lung cancer screening is recommended for chronic smokers to facilitate early detection and treatment.

Serves as another crucial predictor. As individuals age, their organ functions gradually decline, and immunity weakens, making older adults more susceptible to various illnesses. In lung cancer prediction, older patients typically face higher risks. Therefore, enhanced health education and routine health check-ups are crucial for this demographic.

Commonly observed in long-term smokers, this feature significantly elevates lung cancer risk compared to individuals with normal fingers. This finding underscores the association between smoking and lung cancer, offering clinicians a straightforward preliminary screening method.

5. Conclusion

This study investigates the utilization of machine learning algorithms for lung cancer diagnosis and prediction. By collecting clinical data, imaging information, and genetic profiles from public databases, we trained various machine-learning models and evaluated their performance. Our results highlight the outstanding performance of the Random Forest model in lung cancer prediction, especially with its high AUC value. Further analysis underscores smoking status, age, and yellow fingers as the most pivotal predictors. This research not only validates the efficacy of machine learning in lung cancer prediction but also paves the way for novel approaches to early lung cancer detection and treatment. Looking ahead, we aim to further refine model performance and explore advanced machine learning techniques to enhance the accuracy and reliability of lung cancer predictions.

References

[1] World Health Organization. =Cancer Fact Sheet: Lung Cancer. Geneva: WHO, 2023.

[2] Zhang, Renfeng, Zhang Yan, Wen Fengbiao, Wu Kai, & Zhao Song.Analysis of Pathological Types and Clinical Epidemiological Characteristics of 6,058 Lung Cancer Patients. Chinese Journal of Lung Cancer, 2016.

[3] Kaur, I., Doja, M. N., & Ahmad, T. Data mining and machine learning in cancer survival research: An overview and future recommendations. Journal of Biomedical Informatics, 2020.

[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, (2023, February).

[5] Chen, Shihui, Liu Weixiang, Qin Jing, Chen Liangliang, Bin Guo, Zhou Yuxiang, Wang Tianfu, & Huang Bingsheng. Research Progress in Computer-Aided Diagnosis of Cancer Based on Deep Learning and Medical Images. Journal of Biomedical Engineering, 2027, 28(3).

[6] Yin, Bincan Research Team. Application of Random Forest Algorithm in Predicting Postoperative Survival of Patients with Non-Small Cell Lung Cancer. Chinese Journal of Cancer, 2017.

[7] Srikanth, R., Tamil Priya, D., Jagadeesan, S., Patil, S. P., Ingale, A. K., Vivekanandan, M., & Ramalingam, V. (2024, March 13). A comprehensive review on cancer prediction using machine learning techniques. International Journal of Intelligent Systems and Applications in Engineering.

[8] Li, M., et al. Deep learning for lung cancer detection: A review. Artificial Intelligence Review (2024).

[9] American Cancer Society. Key Statistics for Lung Cancer. Atlanta, GA: ACS (2024, January 29).

[10] National Cancer Institute. SEER Cancer Statistics Review, 1975-2015. Bethesda, MD: NCI (2018, September 10)