Lung Cancer Prediction Based on K-nearest Neighbor and Other Algorithms

Yimo Ren

Jinan Foreign Language School, Jinan, 250000, China

Abstract. Lung cancer is still the most affected type of cancer in the world. The purpose of this study is to achieve a certain accuracy of lung cancer prediction based on a variety of computer algorithms, to effectively reduce the prevalence of cancer in the future. The computer algorithms mainly used in this paper include Random forest, K-nearest neighbours, and Logistic regression. By collecting lung cancer patients and clinical data sets, basic prediction is realized through programming code, and data visualization is finally realized to complete prediction. Finally, it is found that the prediction of lung cancer using a single variable is not accurate, and there are many factors leading to lung cancer. It is necessary to import as many data sets as possible to increase the reliability of prediction. The study found that smoking had the greatest impact on the risk of developing lung cancer. After the study in this paper, it is recommended that all people carry out a healthy life schedule, which can effectively prevent lung cancer. At the same time, the study found that the prediction of lung cancer by computer algorithm is achievable, and more algorithms can be combined to achieve higher precision prediction in the future.

1 Introduction

Lung cancer, which is a malignant tumor that starts in the bronchi's epithelium or glands, is quite common and deadly all over the world. The treatment of lung cancer involves surgery, chemotherapy, radiotherapy, targeted therapy, and immunotherapy, which requires a large number of medical resources, while huge medical expenses will bring a heavy economic burden. With the global rise in the prevalence of lung cancer, the pressure on medical equipment is also rising year by yearAccurate prediction of lung cancer can effectively reduce the mortality caused by lung cancer, and improve the early detection rate, effective prevention, and individualized treatment.

The most widely used method for screening for lung cancer at the moment is chest CT. Its high resolution offers a distinct benefit in the early detection of lung cancer and can more clearly display the connections between nearby organs and blood vessels. CT scans do have certain drawbacks, though, namely the possibility of health problems from high radiation

Corresponding author: renhao@cha-tm.com

dosages and the inability to detect small nodules in time. These elements have encouraged the use of artificial intelligence technology for lung cancer early detection. The ability of AI technology to increase job productivity, stability, and diagnostic accuracy is astounding. AI-assisted lung cancer diagnosis is currently widely used in clinical research and practice. For lung cancer adenocarcinoma, the microenvironmental feature has been shown to be an independent predictive factor. Wang developed an automatic cell categorization algorithm called ConvPath that uses the spatial distribution features of different cell types to generate the microenvironmental features of malignancies. Additionally, their research demonstrated that the survival of patients may be predicted by analyzing the spatial arrangement of 48 HE-stained pictures of lung cancer. The late survival rates of the high-risk group were significantly lower than those of the group with a low risk. This research shows that AI diagnostic systems can provide useful quantitative data regarding a patient's prognosis for lung cancer.

The main purpose of this paper is to effectively improve the accuracy of lung cancer prediction through different algorithms and transformation programs. This paper will be from the data source introduction and source, algorithm and model concept introduction and application, as well as the output and result analysis of several aspects to elaborate.

2 Data and Methods

2.1 Data source

The above picture is the data set used in this article, showing the relationship between masculine and feminine and gender, etc. When examining data per gender, this observation demonstrates that Yellow Thumb, Sneezing Persistent Disease, Chest discomfort, and Allergy are crucial symptoms [1].

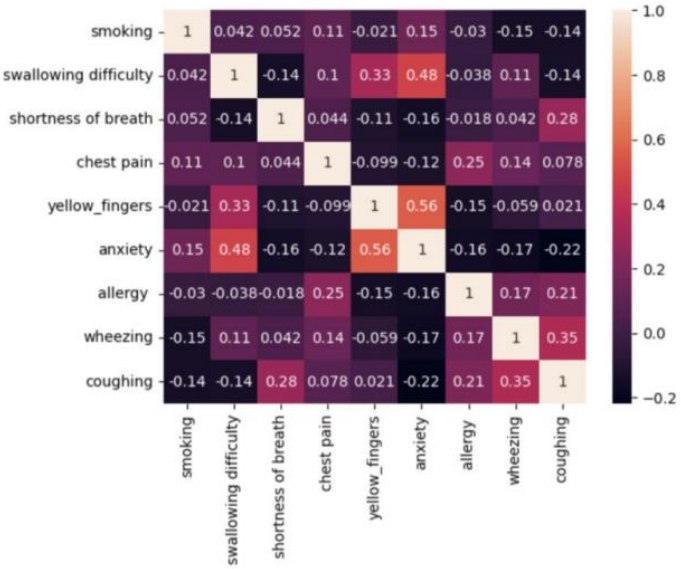

Fig. 1. Heat map with the probability of developing lung cancer (Photo/Picture credit: Original)

Visualize the data by building a data set and running it in Python, Fig. 1 shows the different symptoms associated with the probability of developing lung cancer. Additionally, it displays the likelihood of lung cancer development under certain circumstances. Since the anticipated course of the disease and the chance of life can guide treatment choices, prognosis

prediction is a crucial component of clinical oncology. Prognosis and patient survival may be predicted by DL when applied to genomic, transcriptomic, and other data types. The most common approach to forecast survival is the Cox proportional risk regression framework (Cox-PH), a model of multivariate linear regression that searches for associations between predictor variables and survival time. Because of its linear nature, which may overlook intricate and possibly nonlinear relationships between parts, Cox-PH is difficult to apply to genomic and transcriptomic data [2].

2.2 Method

The algorithms used include Random forest, K-Nearest neighbor (KNN), and Extra tree. After repeated experiments and calculations, it is found that the KNN has the highest accuracy.

Feature selection, regression, classification, and anomaly detection all make extensive use of the well-liked and potent machine learning method known as random forest. Random forest algorithm is an inheritance learning method that constructs multiple decision trees for classification or regression prediction. However, its disadvantages include poor model interpretability and the possibility of overfitting problems in some cases [3].

KNN algorithm is a basic and easily understood classification and regression method. It works by measuring the distance between different eigenvalues. It is suitable for classification and regression problems, but it costs a lot to calculate and store large data sets. The generalization ability of unbalanced data sets is poor. In short, the KNN algorithm is widely used in various fields, such as image recognition, recommendation systems, pattern recognition, etc., because of its simplicity and effectiveness. However, practical applications may require some adjustments to the algorithm to adapt to the needs of specific problems [4].

A popular classification approach in computer learning and data, logistics regression is particularly useful for binary classification issues. Its purpose is to predict the probability of an event, that is, to predict the probability that an instance belongs to a certain category. This algorithm model is simple and easy to understand, has high computational efficiency, can handle linearly separable data, and is suitable for probabilistic prediction. However, for nonlinear data, the performance may be poor and susceptible to outliers. The case of multicollinearity between features is not well handled. Logistic regression is the preferred algorithm for classification problems in many fields, including financial risk assessment, medical diagnosis, customer churn prediction, etc. Although it is a linear model, it can still provide good predictive performance in many practical problems through probability estimation and regularization techniques [5].

2.3 Evaluation indicators

After preliminary statistical analysis, the study applied a variety of machine learning methods to the lung cancer clinical dataset. According to the correlation analysis of the qualities, the data gathering has been streamlined for the lung cancer prediction model. To evaluate the efficacy of the algorithms for learning the confusion matrix, nevertheless, the curve of receiver operating characteristics (ROC) and the area under the ROC curve (AUC) have been taken into account; also, a comprehensive classification analysis has been made available with each approach. The evaluation is divided into three sections: (1) Confusion Matrix, (2) ROC/AUC, and (3) Classification Report. The confusion matrix helps determine precision and recall, which in turn affects the F1-score. AUC guarantees the model's dependability, while the classification report provides the models' overall statistics. The confusion matrix for all of the machine learning models that have been described is shown in Fig. 1, which also compares all of the machine learning techniques that have been addressed [1].

3 Result

By writing code in Python and importing a large number of data sets, it can be found that the basic prediction of lung cancer can be realized through artificial intelligence and big data calculation, which also increases the possibility of using AI to predict the success rate of the possibility of people suffering from lung cancer in the future. It can be seen from the results that as long as enough data is mastered, different ages can be calculated. The likelihood of lung cancer in people who like it [6].

The classification report provides the model's general statistics, and the area under the curve guarantees the model's dependability by using the confusion matrix to compute accuracy and recall rates. The confusion matrix of all the machine learning models that have been discussed is shown in Fig. 2, which also compares all of the methods for machine learning that have been mentioned. The AUC diagram, which compares all of the machine learning techniques presented on the confusion matrix, is displayed in Fig. 3 for each of the models that have been reviewed. Fig. 2 shows that the KNN has the highest accuracy (92.86%), followed by Bernoulli Naive Bayes and Gaussian Naive Bayes, which rank second (91.07%) [7].

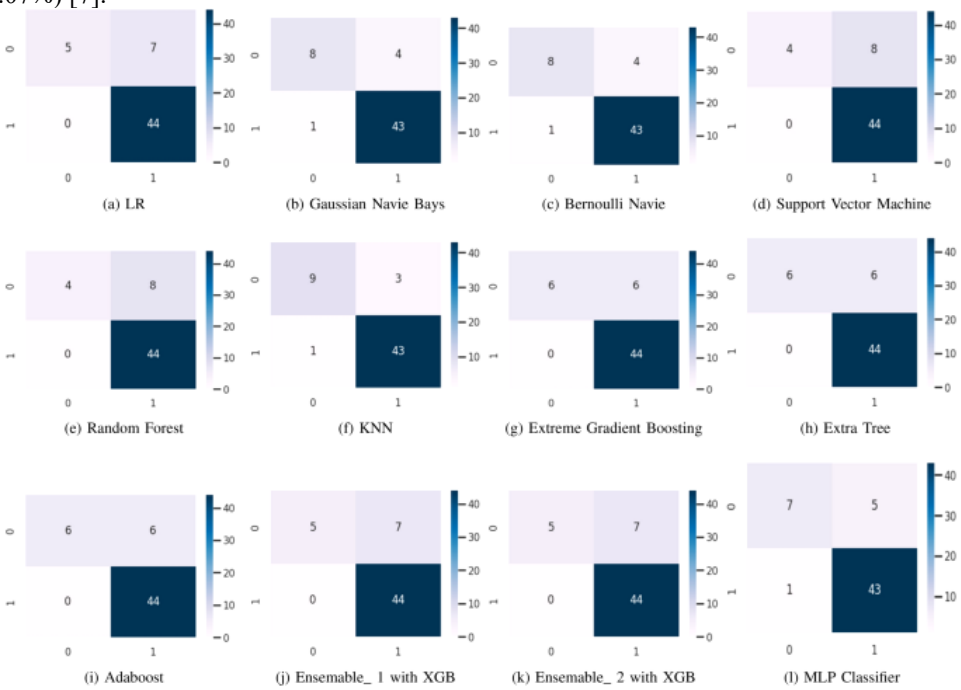

Fig. 2. A comparative study of learning algorithm through confusion matrix (Photo/Picture credit : Original)

Fig. 2 shows the confusion matrix for all machine learning algorithms that have been described, comparing all of the machine learning methods that have been mentioned on the confused matrix. The confusion matrix is useful for calculating precision and recall.

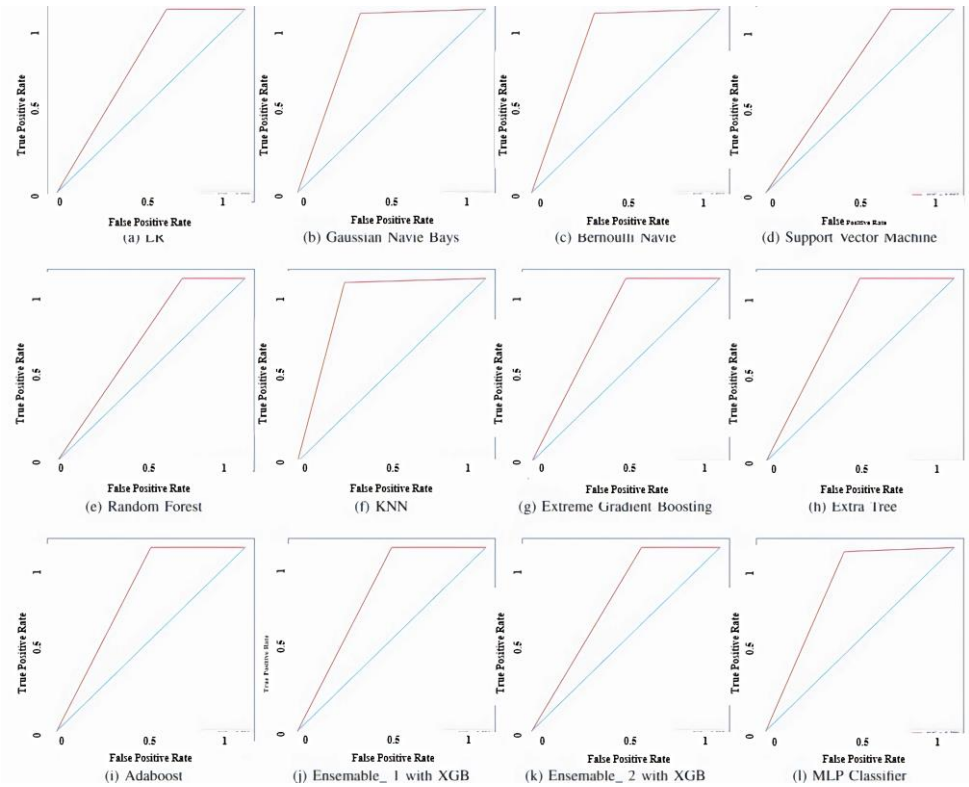

Fig. 3. A comparative study of learning algorithm through ROC/AUC over lung cancer dataset (Photo/Picture credit : Original)

As shown in Fig. 3, the X-axis represents the probability of being false and the Y-axis represents the probability of being true, and the closer the two are, the more accurate the prediction is.

According to the comparison study, KNN has the highest accuracy (92.86%), followed by Bernoulli Naive Bayes and Gaussian Naive Bayes (91.07%) in Fig. 3. Thus, we can ultimately say that the KNN and Bernoulli Naive Bayes models perform better on the smaller datasets that have binary features. They work better in datasets where features and attributes are extremely independent. Other models could not do better for the dataset since they rely on correlation and the data set's training/testing separation.

Lung cancer forecasting can be helpful if the technique is effective once symptoms are identified and also correlates with the patient's lifestyle and low-risk cancer status. Furthermore, depending on the patient's cancer risk level, the specialist may suggest the best course of action. However, when forecasting lung cancer in a patient, accuracy is crucial. After processing the 310 instances of raw data to identify positive results by gender, each attribute's individual positive cases were compared by gender. According to a preliminary examination of the data, yellow thumb, and allergies are the most common symptoms in a correlation study over alcohol consumption patterns. With a precision of 92.86% and 91.07%, respectively, the KNN and Bernoulli Naive Bayes models—which perform similarly well as Gaussian Naive Bayes—were determined to be appropriate in this study's thorough examination of twelve possible machine learning methods.

For larger datasets, certain possible algorithms, such as Multilayer Perceptron (MLP) and Ensemble 1 with XGB and ADA, might be further examined. In addition to statistical analysis, information from the collection was used in the course of the research [8].

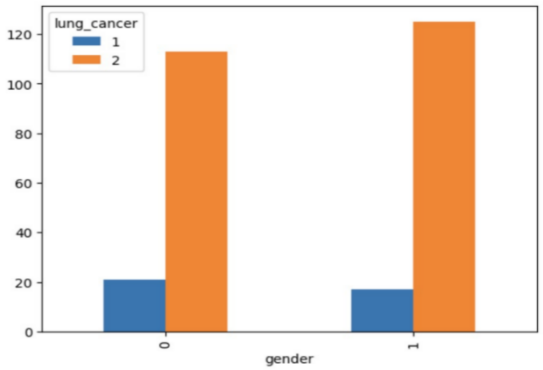

Fig. 4. The probability of lung cancer by age and sex (Photo/Picture credit: Original)

These are two charts created by Python visualization, respectively, showing the possibility of lung cancer under different genders and different ages (Fig. 4). Based on clinical data and cases, they are highly reliable. It can be seen that the possibility of lung cancer has a large gap between genders, which is also related to the fact that the proportion of men smoking is much higher than that of women. Different genders have different maximum and minimum ages for lung cancer, which has a lot to do with daily life and so on.

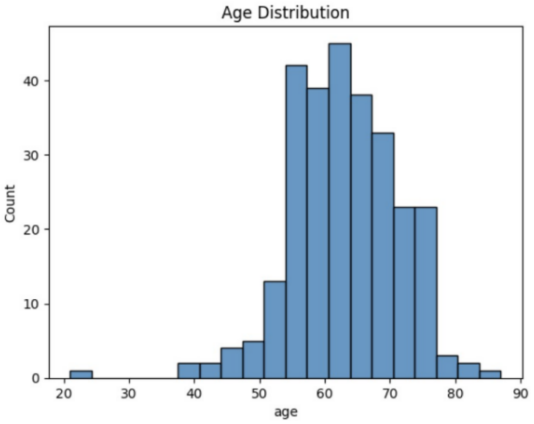

Fig. 5. The probability of lung cancer in different age groups (Photo/Picture credit: Original)

Fig. 5 is another visual experiment by Python, which shows the possibility of lung cancer in different ages. It can be seen that the peak of lung cancer is between 60 and 70, which also reflects the feasibility of predicting lung cancer through data [9,10].

4 Conclusion

In this paper, a variety of computer algorithms were used to predict lung cancer, and code writing and data visualization were carried out. This article mainly compares and analyzes 12 machine learning algorithms including Logistic regression, Gaussian Naive Bayes, and Bernoulli Naive Bayes. Through the results, it can be found that KNN is the most accurate prediction, which also proves the feasibility of using a computer algorithm to predict lung cancer. At the same time, research has found that the cause of lung cancer is related to a variety of factors, including gender, age, and so on, and will cause different symptoms, Yellow Finger, Coughing, Chronic Disease, Chest Pain, and Allergy are critical symptoms.

Therefore, the impact of various data should be considered before the prediction results become reality. A single variable can skew the results.

In addition to the machine learning model method used in this paper for lung cancer prediction, other methods can be used, such as time series analysis, prediction of lung cancer prevalence through autoregressive summing moving average model and neural network autoregressive model, or the use of imaging omics, which can quantify medical imaging data. A large number of image features can be extracted and combined with machine learning or deep learning models for lung cancer prediction and prognosis analysis. By combining different methods, the accuracy of lung cancer prediction can be effectively improved, which is also the future research direction. At the same time, the prediction of lung cancer is of great significance. The survival rate of early-stage lung cancer is much higher than that of late-stage lung cancer. Through the predictive model, early treatment can be carried out, significantly improving the survival rate and quality of life of patients, and can also help doctors to develop personalized prevention and treatment plans for each patient, improve the effectiveness of treatment, and reduce unnecessary treatment, save medical resources, reduce medical costs, promote scientific research in related fields, and so on.

Reference

1. S. P. Maurya, et al., Performance of Machine Learning Algorithms for Lung Cancer Prediction: A Comparative Approach, Scientific Reports, 14, 1, Nature Portfolio, (2024).
2. K. A. Tran, et al., Deep Learning in Cancer Diagnosis, Prognosis and Treatment Selection, Genome Medicine, 13, 1, (2021).
3. L. Breiman, Random Forests, Machine Learning, 45, 1,5-22, (2001).
4. S. Uddin, et al., Comparative Performance Analysis of K-Nearest Neighbour (KNN) Algorithm and Its Different Variants for Disease Prediction, Scientific Reports, 12, 1,(2022).
5. Z. Bobbitt, Introduction to Logistic Regression, Statology,27, (2020), www.statology.org/logistic-regression/.
6. P. G. Mikhael, et al., Sybil: A Validated Deep Learning Model to Predict Future Lung Cancer Risk from a Single Low-Dose Chest Computed Tomography, Journal of Clinical Oncology, 41, 12, (2023).
7. N. D. Rao, et al., Household Contributions to and Impacts from Air Pollution in India, Nature Sustainability, 4, 10, (2021).
8. E. Nemlander, et al., Lung Cancer Prediction Using Machine Learning on Data from a Symptom E-Questionnaire for Never Smokers, Former Smokers, and Current Smokers, PloS One, 17, 10, (2022).
9. R. Mahmoud, Lung Cancer Prediction Using ML Regression Models, Kaggle.com, Kaggle, (2024). www.kaggle.com/code/rawanmahmoud19/lung-cancer-prediction-using-ml-regression-models/notebook#Gender. Accessed 17 Oct. (2024).
10. J. Brownlee, How to Develop an Extra Trees Ensemble with Python, Machine Learning Mastery, 21 Apr. (2020), machinelearningmastery.com/extra-trees-ensemble-with-python/.