Lung Cancer Prediction based on KNN, Logistic Regression, and Random Forest Algorithm

Yunzhe Liao *

Department of Schools of software & Internet of things engineering, Jiang Xi University of Finance and Ecnomics, Nanchang, China

* Corresponding author: 2202203465@stu.jxufe.edu.cn

Abstract. Cancer has become the biggest killer of human health in the 21st century, among them, lung cancer is the "number one killer" of cancer ", and has become the disease with the highest incidence rate of cancer. However, many factors cause lung cancer. Therefore, to effectively control cancer and reduce the incidence rate of lung cancer, it is very important to predict the incidence of lung cancer caused by different factors. By calling the dataset in Kaggle, we can visualize and preprocess data from 25 factors that may cause lung cancer incidence, and use the machine of K-nearness model, random forest model and logical regression model Learn the model for prediction and obtain the model with the highest accuracy. It was found that the random forest algorithm performed the best among the three models used, so the random forest algorithm was used for model tuning, adjusting parameters to improve accuracy. Finally, the model accuracy was increased from 0.96 to 0.98, achieving the expected effect. This study is beneficial for a clearer and more accurate understanding of the impact of different factors on the incidence of lung cancer, as well as for more efficient prediction of lung cancer in the medical field.

Keywords: Lung cancer, Prediction model, Evaluation, Machine learning.

1. Introduction

With the rapid growth of the global economy in the 21st century, human living standards have been improved, and the average life expectancy of humans has unprecedented room for improvement [1]. However modern life has posed many new threats to human health, such as cancer. Many carcinogenic factors in daily life threaten human health, especially in modern industrial production such as mining, smelting, and nuclear power applications, which can form carcinogenic pollution. Cancer has become the biggest killer of human health in the 21st century [2], Among them, lung cancer is the "first killer" of cancer [3]. According to media reports, the first case of lung cancer has been more than 100 years. In the past 100 years, lung cancer has changed from an "unattractive" disease to a disease with the highest incidence rate. cause lung cancer. Therefore, many factors and precursor manifestations that can cause the onset of lung cancer. However, these factors have a wide variety of types and varying degrees of impact. Therefore, to effectively control cancer and reduce the incidence rate of lung cancer, it is very important to predict the incidence of lung cancer caused by different factors,

Early lung cancer prediction models were mostly analyzed using few factors or traditional communication methods. Compared with previous genome-wide association studies based on lung cancer risk prediction models in the Chinese population, three risk prediction models based on smoking information, genetic scores, and the combined use of smoking and genetic information were used to predict the risk of lung cancer in the Chinese population [4], as well as the previous use of the "Delphi statistical method", Statistical prediction was made on the development trend of cancer with the top 10 incidence rate in China. The problems to be solved were sent to each expert separately using communication for advice, and then the opinions of all experts were collected and summarized, and the comprehensive opinions were sorted out. Subsequently, feedback will be provided to experts separately, and opinions will be solicited again. Each expert will modify their original opinions based on comprehensive opinions, and then summarize them. [5], which may have the drawbacks of being time-consuming and the process is complex.

This article will also focus on predicting the incidence of lung cancer based on different factors. By calling the dataset in Kaggle, 25 possible factors that may cause lung cancer will be classified from the perspectives of hundreds of respondents, such as their physical condition and lifestyle. These factors will be approximated into corresponding Arabic numerals for visualization and data preprocessing, and machine learning models will be used for prediction. This article will use the K-nearest neighbor model, random forest model, and logistic regression model, and provide corresponding AUC and ROC curves. This article will focus on various behavioral patterns and physical symptoms of residents, and research more influencing factors. Through artificial intelligence machine learning, the model will have better generalization ability in more complex and large amounts of data compared to before. This article will use the three models provided, accuracy, macro average, weighted average, and F1 value as evaluation indicators, and use 4-fold cross-validation to evaluate various models, compare different model structures, explore which model has the highest accuracy, and provide the model with the highest accuracy.

2. Method and Data

2.1. Datasets

The dataset cited in this article regarding the incidence factors of lung cancer is from Kaggle's HTTPs://www.kaggle.com/datasets/mysar Ahmadbhat/lung cancer/data website. Quoting this dataset can improve the effectiveness of lung cancer prediction, help people understand their cancer risk at a low cost, and also help people make appropriate decisions based on their cancer risk status. This includes objective factors such as gender and age, as well as environmental factors such as air pollution and dust allergies, as well as daily factors such as smoking, and secondhand smoke, dietary habits, and disease factors such as chest pain, shortness of breath, dry cough, and coughing up blood. The relationship between various factors and the degree of causing lung cancer will be specifically presented in Section 3.1 Data Visualization, which only provides a brief introduction to the variables.

2.2. Model Introduction

K-nearest neighbor (k-NN) is a classic classification and regression algorithm. The basic idea is to classify or regress predictions based on feature similarity. K-NN classifies the test sample as the most frequently occurring category among neighboring samples or predicts its continuous value by comparing the feature distance between the test sample and the training sample. The main advantages of the k-NN algorithm are its simplicity and ease of implementation, no need for training process and parameter estimation, no need to select the best parameters for model tuning, and good interpretability. In addition, the k-NN algorithm is suitable for multi-class problems, has strong adaptability, no assumptions about data distribution, has high fault tolerance and is not sensitive to noise and outliers. Therefore, it is more effective in dealing with the classification problem of lung cancer prediction. Use the K-nearest neighbor algorithm to optimize the k-value and visualize and evaluate the prediction results. First, create an empty list to store the cross-validation scores corresponding to different k values. Create a K-nearest neighbor classifier through a loop from k=1 to k=19, calculate the average score of k-fold cross validation, and add the score to the previously created list. Create a K-nearest neighbor classifier, set k to 1, and use this model to fit the training data. By plotting the relationship between k-values and cross validation scores, it is possible to help select the optimal k-value, thereby improving the model performance of the K-nearest neighbor algorithm. Confusion matrices and classification reports can be used to evaluate the predictive performance of models.

Random Forest is a powerful machine learning model used to solve classification and regression problems. The Random Forest (RF) algorithm is a machine learning algorithm based on decision trees, which has good classification and variable screening performance [6]. Each decision tree in a random forest is constructed based on a randomly selected feature subset, and bootstrap is performed on the training samples to produce trees with uniqueness and diversity. This randomness helps to reduce overfitting problems and improve the model's generalization ability. By integrating the prediction

results of each decision tree on the training set, a comprehensive prediction is ultimately obtained. Random search cross validation was used to optimize the hyperparameters of the random forest classifier, and a parameter dictionary was defined, focusing on the number of decision trees. Then a random search cross validation object was created, using a random forest classifier and defining parameters. Perform a random search cross validation process on the incoming training data to find the optimal parameter combination. And use the trained model to predict the test set and obtain the prediction results. Random search cross validation can help us randomly select parameter combinations for model training and evaluation within a given parameter range, in order to find the best parameter combination and improve model performance.

A logistic regression model is a statistical learning method used to solve binary classification problems, which predicts the probability of a sample belonging to a certain category by mapping the output results of the linear regression model to the [0,1] interval. The logistic regression model is a linear classification model, and its results can be interpreted as the degree to which different features affect the prediction results. The positive or negative impact of features on prediction results can be determined by the positive or negative coefficients. High computational efficiency: The logistic regression model has a small computational load, and fast training speed, and performs well on datasets of this size. This is because the objective function of the logistic regression model is a convex function, and optimization algorithms such as gradient descent can be used to quickly find the optimal solution. Logistic regression models can handle multi-classification problems in a one-to-many manner. For K categories of problems, K logistic regression models can be trained, each model predicts the probability of a certain category and then selects the category with the highest probability as the final prediction result.

2.3. Evaluation Indicators

Accuracy is the percentage of correctly predicted samples to the total amount. Specifically, it is to classify and predict the samples in the test set, and then calculate the ratio between the number of correctly predicted samples and the total number of samples. This indicator is simple and clear, but one drawback is that it cannot effectively evaluate the performance of the model when the sample size of the data is imbalanced.

Macro average refers to the arithmetic mean of each statistical indicator value for all categories, including accuracy, recall, and F1 value. For each category, calculate the precision, recall, and F1 value for that category, and then take the average of these values as the result of the macro average. This indicator can avoid the problem of imbalanced data samples because when calculating the average, the statistical indicator values of each category have the same weight.

Weighted average is the process of weighting the average value of each category based on its importance. Specifically, for each category, a weight value is assigned based on its importance, and then the accuracy, recall, and F1 values of the category are calculated. The values of these three indicators are then multiplied by the corresponding weight values, and finally, the weighted average is obtained by summing up. This indicator can also avoid the problem of imbalanced data samples because when calculating the average, the statistical indicator values of each category have different weights.

Precision is an evaluation indicator for predicting results. The percentage of positive samples predicted by the model. This indicator is relatively cautious and has a high classification threshold, which can avoid predicting too many negative samples as positive samples.

Recall rate is an evaluation indicator for the original sample. The percentage of actual positive samples predicted to be positive. This indicator can measure whether the model can predict all positive samples.

F1 Score is a comprehensive evaluation indicator, which is the harmonic average of accuracy and recall. This indicator can comprehensively evaluate the overall performance of the model, avoiding potential conflicts between accuracy and recall. When the F1 score is high, it indicates that the model performs well in terms of accuracy and recall.

3. Result

3.1. Data visualization and Data visualization analysis

Data visualization is an important component of any project in data science because it can let us gain a deeper understanding of the attributes that display different features most intuitively. This article will present data visualization from two perspectives: the correlation between various factors and the correlation between various factors and the incidence of lung cancer.

Fig. 1 shows the correlation between various factors, with anxiety level as a parameter of 1. Comparing the correlations of various categories, the correlation between yellow finger level and anxiety level is the highest, reaching 0.57, followed by anxiety level and chewing difficulty level reaching 0.49, followed by the correlation between alcohol consumption and gender, reaching 0.45. In addition, there is a low correlation between the degree of swallowing difficulty and age, allergy level, and smoking frequency.

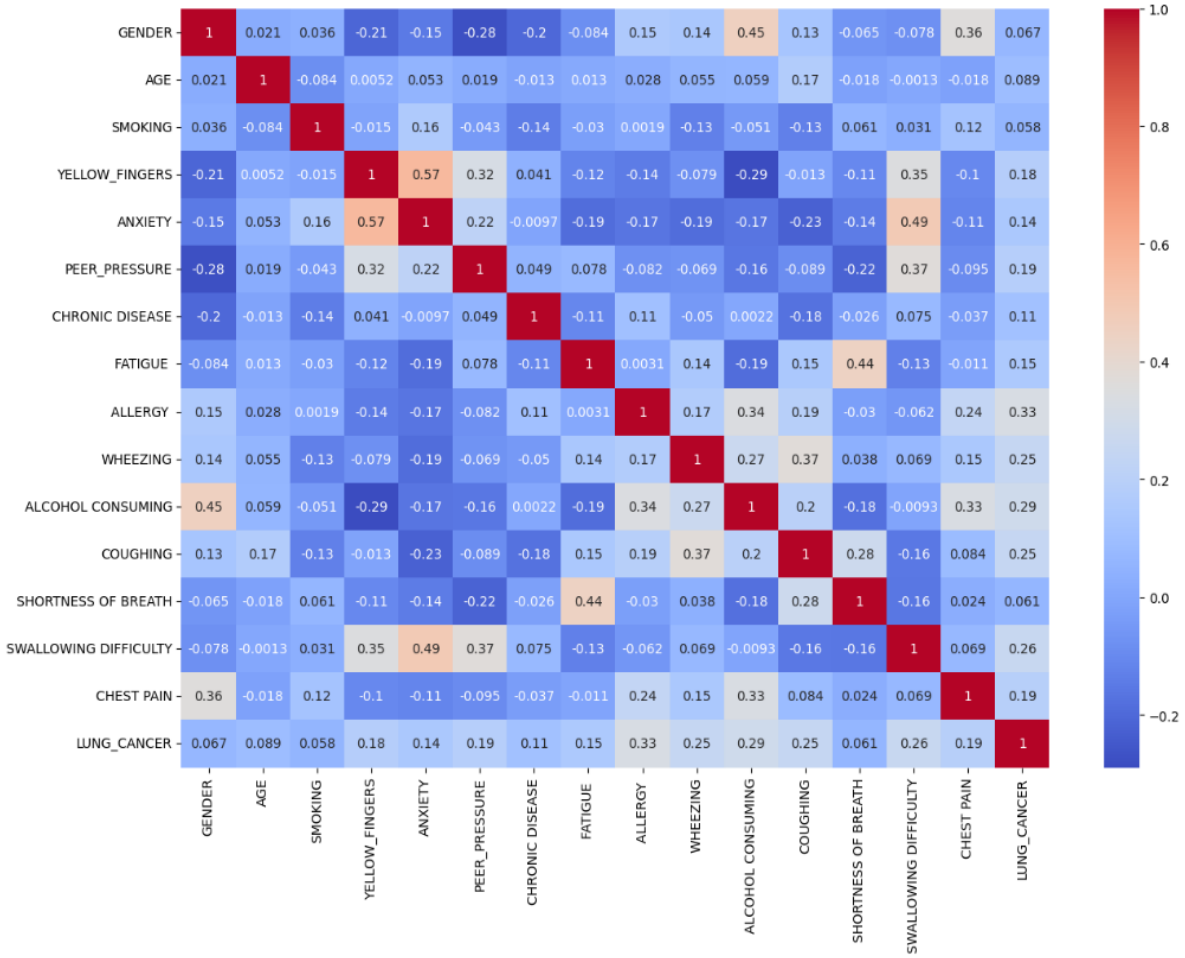


Figure 1. Heatmap of the correlation between various factors

The heatmap of the correlation between various factors in Fig. 1 and Fig. 2 shows the correlation between each factor and the incidence of lung cancer. It can be seen that allergy and lung cancer have the highest correlation, reaching around 0.33, followed by alcohol consumption. This indicates that allergic factors and alcohol consumption may be important factors in the onset of lung cancer. The correlation between smoking frequency and the incidence of lung cancer is relatively low. Previous scholars have conducted population-based case-control studies in areas with high tumor incidence. Compared to non-smokers, smokers have an increased risk of developing lung cancer, and the risk of lung cancer from smoking to the lungs is 4.44 times higher than non-smokers. Therefore, smoking is the main risk factor for the disease. And it increases with the increase of smoking years, smoking volume, and smoking depth [7]. However, this conclusion is contradictory to the results of this study,

possibly due to bias in the data used, that is, the correlation between smoking and lung cancer may not be fully covered in the data. If the proportion of smokers to non-smokers in the data sample is imbalanced, or other related factors are not fully considered, the prediction results of the model may be affected, or the dataset is relatively small, the model may not be able to fully learn the relationship between smoking and lung cancer, or the model used may not be sensitive enough to the relationship between smoking and lung cancer, or the model may be too simple to capture complex associations.

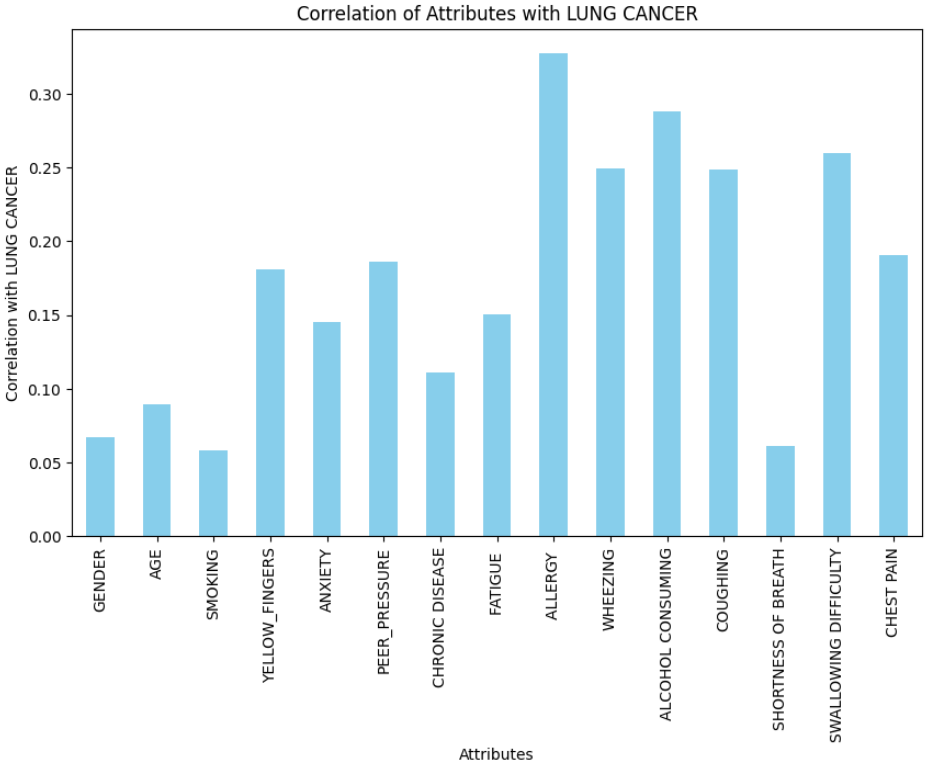


Figure 2. The correlation between various factors and the incidence of lung cancer

3.2. Machine learning model result & Visualization

Fig. 3 and Fig.4 show the confusion matrix and classification evaluation reports of accuracy, recall, F1 value, etc. using KNN, logistic regression, and random forest algorithm, respectively. When the training set sample is 162, the KNN algorithm predicts that the number of healthy individuals is 80, and the number of unhealthy individuals is 5; The predicted number of healthy individuals for lung cancer is 1, and the number of unhealthy individuals is 76. The logistic regression algorithm predicts that the number of healthy individuals is 76, while the number of unhealthy individuals is 11; The predicted number of healthy individuals for lung cancer is 5, and the number of unhealthy individuals is 70. The random forest algorithm predicts that the number of healthy individuals is 81, and the number of unhealthy individuals is 6; The predicted number of healthy individuals for lung cancer is 0, and the number of unhealthy individuals is 75. The accuracy, macro average, and weighted average of the KNN algorithm under the recall rate all reach 0.96. The accuracy, macro average, and weighted average corresponding to the F1 value are also 0.96. The accuracy, recall, macro average, and weighted average of the logistic regression algorithm all reach 0.90. The accuracy, macro average, and weighted average corresponding to F1 value are also 0.90. The accuracy of the random forest algorithm corresponds to a macro average and weighted average of 0.97. The macro average and weighted average under recall rate are 0.96. The accuracy, weighted average, and macro average corresponding to F1 value are all 0.96. By comparing the accuracy, the models of KNN and random forest algorithm were first selected, both achieving 0.96. The current goal is to select one of these models for tuning parameters to obtain the final model, which has been introduced in section 2.3 and provides the numerical basis. By comparing the macro mean and weighted mean, the KNN and

random forest are 0.96, 0.96, and 0.97, 0.97, respectively (Table 1). Therefore, the random forest algorithm will be chosen.

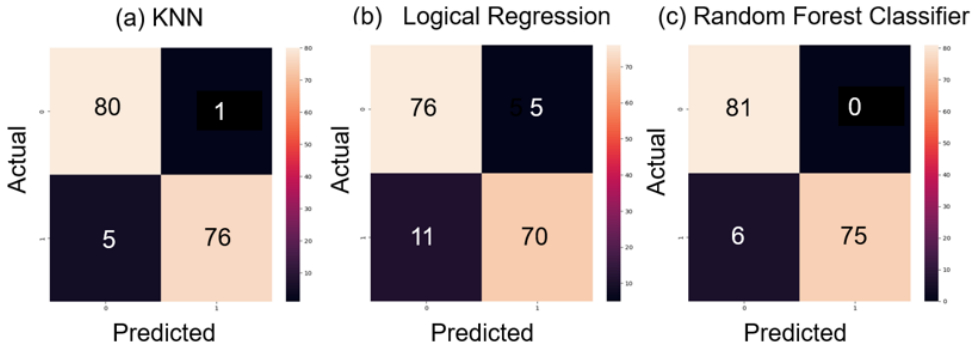


Figure 3. Confusion Matrix of Three Models

Table 1. Evaluation indicators for three models

		Precision	Recall	F1-score
KNN	Accuracy			0.96
	Macro avg	0.96	0.96	0.96
	Weighted avg	0.96	0.96	0.96
Logical Classifier	Accuracy			0.90
	Macro avg	0.90	0.90	0.90
	Weighted avg	0.90	0.90	0.90
Random Forest Classifier	Accuracy			0.96
	Macro avg	0.97	0.96	0.96
	Weighted avg	0.97	0.96	0.96

3.3. Machine learning results after tuning utilizing the data visualization analysis

In the analysis report of Fig. 3, the optimal parameter is 0.75, which will be used for model tuning in this section, constructing a random forest model using 75 decision trees. Then, use the training dataset to train the model, and next, use the trained model to predict the test dataset.

As with previous model training, to evaluate the performance of the model, the confusion matrix function was called to calculate the difference between the predicted results and the actual labels. The predicted number of healthy individuals is 81, and the number of unhealthy individuals is 4; The predicted number of healthy individuals for lung cancer is 0, and the number of unhealthy individuals is 77. Then, using the heatmap function in the seaborn package, the confusion matrix is visualized in the form of a heatmap, and a report on classification metrics such as accuracy, recall, and F1 score is printed. Fig. 4 and Table 2 shows the visualized results, and the accuracy, macro average, and weighted average of the random forest algorithm under quadruple cross-validation all reach 0.98. The accuracy, macro average, and weighted average corresponding to the F1 value are also 0.98, achieving the expected effect.

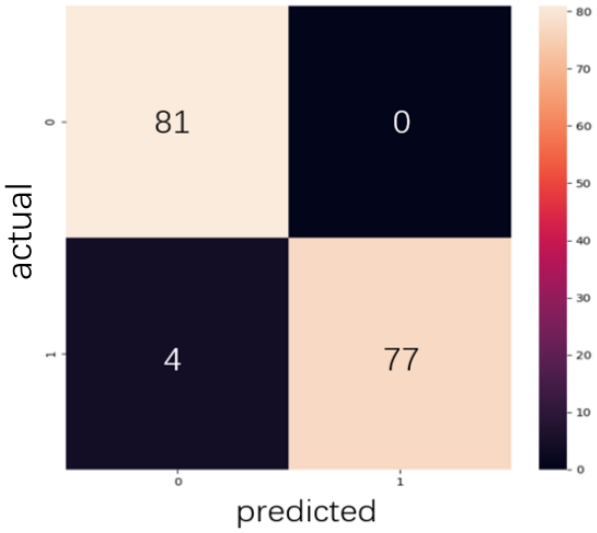


Figure 4. The Confusion Matrix of Random Forest classifier under 75 Decision Trees

Table 2. Evaluation Report on Random Forest Classifier under 75 Decision Trees

	Precision	Recall	F1-score
Accuracy			0.98
Macro avg	0.98	0.98	0.98
Weighted avg	0.98	0.98	0.98

4. Conclusion

This article drew a heatmap describing the correlation between various factors based on a dataset of lung cancer incidence factors, and obtained the yellow finger, with the highest correlation between severity and anxiety level. Next are the levels of anxiety and difficulty chewing, followed by the correlation between alcoholic beverages, dietary habits, and gender. A chart was also drawn to visualize the correlation between various factors and the incidence of lung cancer. Allergies and lung cancer have the highest correlation, followed by alcohol consumption, swallowing difficulties, wheezing, and coughing. At the same time, by comparing KNN, logistic, and regression random forest algorithms on the dataset of lung cancer incidence factors, the prediction of lung cancer incidence is carried out. Use this dataset to compare accuracy, macro average, and weighted average to obtain the optimal model, and adjust the parameters of the model to optimize it and achieve higher accuracy. The random forest algorithm performed the best among the three models used, so it was used for model tuning, adjusting parameters to improve accuracy. Finally, the accuracy of the model was improved from 0.96 to 0.98, achieving the expected effect. In addition to providing only three models in this study, more model algorithms such as SVM, neural network models, naive Bayesian models, decision tree models, etc. can be chosen to compare the optimal prediction models. This study can provide readers with a clearer understanding of the different pathogenic factors of lung cancer. It can predict whether different factors will cause lung cancer, to effectively control the incidence rate of lung cancer. It has the advantages of a simple process, saving manpower and material resources.

References

[1] Fang Yiliang. With the continuous growth of the global economy, the significant improvement of living standards, the fundamental improvement of medical conditions, and the steady extension of peacetime, the average life expectancy of humanity has achieved unprecedented upward space. However, even so, the century old people, who can be regarded as "auspicious", still belong to the rare category, which can be said to be [J] [2023-12-04].

[2] Zhao Chenghui. Carcinogenic factors in daily life [J]. Medicine and Health, 2007, 15 (2): 1. DOI: CNKI: SUN: TAKE.0.2007 - 02 - 016.

[3] Liang Zhongshi .au@sio2 Preparation of nanocomposites and their application in photothermal therapy for liver cancer [D]. East China Normal University [2023-12-04] DOI: 10.76666/d. y1904196.

[4] Zhu Meng, Cheng Yang, Dai Juncheng, et al. A predictive model for lung cancer risk in the Chinese population based on genome-wide association studies [J]. Chinese Journal of Epidemiology, 2015, 36 (10): 6. DOI: 10.3760/cma.j.issn.0254 - 6450 - 2011.10002.

[5] Zhao Boyang. Prediction of Top Ten Common Cancer Incidence Trends [J]. Wishing You Good Health, 1995 (3): 2. DOI: CNKI: SUN: ZNJK. 0.1995 - 03 - 041.

[6] Cheng Xuexin. Research on Particle Swarm Optimization Weighted Random Forest Algorithm [D]. Zhengzhou University [2023-12-04]. DOI: CNKI: CDMD: 21017. 146473.

[7] Gu Xiaoping, Wang Yincun, Zhi Hengkui, et al. The impact and interaction of smoking and drinking on the incidence of lung cancer [J]. Jiangsu Preventive Medicine, 2015, 26 (5): 3. DOI: 10.13668/j.issn.1006 - 9070. 2015. 05. 013.