International Journal of Healthcare Information Systems and Informatics
Volume 19 • Issue 1

Classification Prediction of Lung Cancer Based on Machine Learning Method

Dantong Li, Weifang Hospital of Traditional Chinese Medicine, China
Guixin Li, Weifang Hospital of Traditional Chinese Medicine, China*
Shuang Li, Weifang Hospital of Traditional Chinese Medicine, China
Ashley Bang, The St. Nicholas School, Vietnam

ABSTRACT

The K-nearest neighbor interpolation method was used to fill in missing data of five indicators of coronary heart disease, diabetes, total cholesterol, triglycerides, and albumin;, and the SMOTE algorithm was used to balance the number of variable indicators. The Relief-F algorithm was used to remove 18 variable indicators and retain 42 variable indicators. LASSO and ridge regression algorithms were used to remove eight variable indicators and retain 52 variable indicators; The prediction accuracy, recall, and AUC values of the linear kernel support vector machine model filtered using Relief-F and LASSO features are high, and the prediction results are optimal; The test result of random forest screened by Relief-F and LASSO features is better than that of the support vector machine model. It is concluded that the random forest model screened by Relief-F features is better as a prediction of lung cancer typing. The research results provide theoretical data support for predicting lung cancer classification using machine learning methods.

KEYWORDS

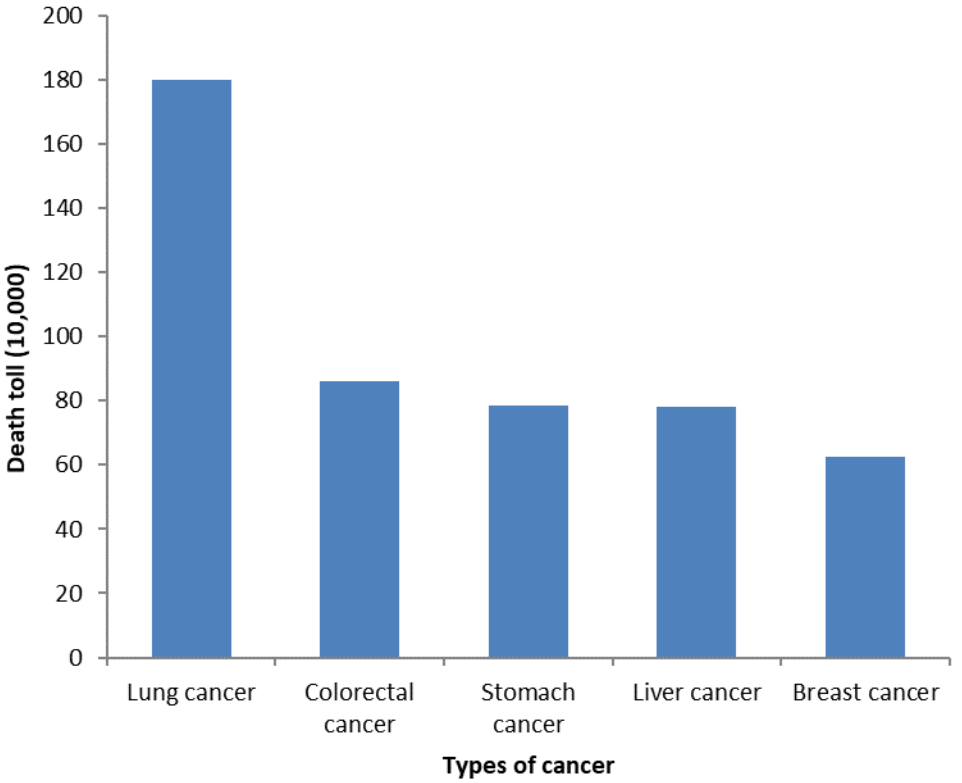
Lung Cancer Typing, Machine Learning, Random Forest, Support Vector Machine

INTRODUCTION

With rapid economic and social development, human lifestyle and eating habits have changed significantly. In the face of long-term unhealthy living conditions, ionizing radiation, poor environment, and other adverse factors, the incidence rate of cancer in China is increasing year by year, and the types of cancer are also increasing. According to a survey from the International Agency for Research on Cancer, the number of cancer deaths worldwide is growing exponentially. The number of deaths due to different cancers in 2019 was 1.8 million for lung cancer, 870,000 for colorectal cancer, 780,000 for gastric cancer, 780,000 for liver cancer, and 630,000 for breast cancer (Wang & Yuan, 2019) (see Figure 1). Among them, lung cancer has the highest incidence rate and is particularly prominent among men. Lung cancer, as the most common fatal disease worldwide, is influenced by multiple factors. Smoking has been identified as the main risk factor for lung cancer, and smokers are more than 10 times more likely to develop lung cancer than nonsmokers. Harmful substances

DOI: 10.4018/IJHISI.333631 *Corresponding Author

Figure 1. Statistics of cancer death cases



such as PM2.5, sulfur dioxide, and carbon monoxide in the air are also increasing the risk of lung cancer. At the same time, professions such as mining, welding, and painting are increasing the risk of lung cancer due to long-term exposure to harmful substances. According to the latest cancer burden data, there are over 4 million confirmed lung cancer patients worldwide each year, with nearly half of them dying from cancer.

Lung cancer poses a huge threat to human survival and health. The confirmed cases of lung cancer are mainly adenocarcinoma, small cell lung cancer, and squamous cell carcinoma of the lung. The treatment methods for different types of lung cancer vary greatly (Abdullah et al., 2021). At the same time, it is necessary to pay attention to the patient's psychological state and prescribe appropriate drugs before treatment. Tumor markers and imaging diagnosis of lung cancer are widely used in clinical practice, but some markers, such as carcinoembryonic antigen, are not specific enough to cause errors in clinical diagnosis. Imaging diagnosis (such as chest X-ray, CT, magnetic resonance imaging, etc.) has certain value for diagnosis; however, small pulmonary nodules or lymph node metastases may be missed due to poor imaging. The main treatment methods for lung cancer include surgery, radiotherapy, chemotherapy, and targeted therapy. For early-stage lung cancer patients, surgical treatment can be used and is currently the most effective treatment method. Radiotherapy, which kills cancer cells to alleviate symptoms, is mainly aimed at patients whose cancer cannot be surgically removed or who have residual cancer cells after surgery. Chemotherapy mainly targets patients with advanced lung cancer, killing cancer cells through intravenous injection or oral medication. Targeted therapy is the targeted killing of lung cancer cells by identifying their molecular targets. With the development of lung cancer screening technology, most lung cancer is easily detected in the early stage. At the

same time, with the rapid growth of medical data information, a large amount of medical diagnostic information has been digitized. Establishing lung cancer prediction models to assist diagnosis and treatment has important research significance.

Today, the incidence and mortality rate of lung cancer have rapidly increased, and this has become the cancer with the highest mortality rate in the world. By analyzing lung cancer medical data through machine learning, a complete lung cancer prediction model is established to provide a basis for assisting in lung cancer prevention, diagnosis, and treatment measures. This paper selects the clinical diagnosis, treatment, and experimental data of lung cancer patients in the database of the US National Center for Biotechnology Information (NCBI) and uses the K nearest neighbor interpolation and synthetic minority over-sampling technique (SMOTE) to complete missing values and solve the problem of data imbalance. The Relief-F filtering method and least absolute shrinkage and selection operator (LASSO) embedding method are used to extract the characteristics of patient indicators, and the prediction model is constructed through support vector machines and random forest machine learning methods. Then, the prediction effect is compared between the recall rate and area under curve (AUC) indicators through the accuracy rate.

LITERATURE REVIEW

Based on the establishment of a large database, machine learning algorithms can effectively infer patterns from massive amounts of data and automatically deduce scientific mathematical models (Liu et al., 2021). At present, machine learning is mainly applied in the prediction of lung cancer in image recognition, data mining, gene analysis, and clinical decision support; it helps doctors make more accurate diagnoses and more scientific treatment plans and improve the early diagnosis rate and treatment effect of lung cancer (Sadhwani et al., 2021).

Some researchers have conducted deep learning-based lung cancer screening research, proposing the use of 3D-CNN deep convolutional neural network deep learning algorithm to automatically recognize and locate lung nodules in low-dose chest CT scan images, classify and segment them, and achieve end-to-end lung cancer screening with an accuracy rate of up to 93.5% (Chaturvedi et al., 2021). Some researchers have studied deep learning lung cancer risk prediction models and proposed deep learning algorithms using deep neural networks (DNN) to mine and analyze patient medical history, symptoms, and physiological indicators data (Yawen et al., 2018). They automatically learn and extract features related to lung cancer to predict the occurrence and development trend of lung cancer with an accuracy of 90%, providing decision support for doctors' diagnosis and treatment. Some researchers have used central focus convolutional neural networks for lung nodule segmentation and established an automatic localization and segmentation model for lung nodules using CF-CNN central focus convolutional neural networks (Chauhan et al., 2017). The prediction accuracy has reached 91%, thus achieving early diagnosis and treatment of lung cancer. Some researchers have established deep learning lung cancer risk prediction models using 2D convolutional neural networks (CNN) to automatically learn and extract feature indicators related to lung cancer from low-dose CT images (Ubaldi et al., 2021). The prediction accuracy of lung cancer risk in patients is 84%, which can provide decision-making for doctors' diagnosis. Some researchers have applied deep learning to the treatment of lung cancer, using the deep learning algorithm of CNN-LSTM deep convolutional recurrent neural network to automatically learn and extract medical image data from patients, predicting lung cancer treatment response and survival rate with a prediction accuracy of 81% (Liu et al., 2021). Other researchers have studied the random forest algorithm to establish a machine learning prediction model for lung cancer (Jena et al., 2019). By selecting a certain number of CT image data of lung cancer patients, 132 medical index features are extracted, and the random forest algorithm is used to establish a machine learning model to predict whether there is gene mutation in lung cancer patients. The prediction accuracy rate has reached 76%. By applying machine learning to lung cancer

prediction, better support is provided for early diagnosis and treatment of lung cancer. However, these algorithms still need more practical validation and optimization to be applied in clinical practice.

Some researchers use machine learning algorithms to predict malignant and benign pleural effusion (Jung et al., 2017). They compare pleural detection techniques using adaptive neurofuzzy inference algorithms, support vector machines, and sampling lifting trees. By setting model variables as detection indicators for serum and pleural fluid, they compare and validate the prediction results of each model on the data of 260 clinical patients. The results show that the prediction effect of the support vector machine prediction model is the best, and the prediction effect of sample lifting tree model is not ideal. Researchers have proposed the optimal deep neural network and linear discriminant algorithm to analyze lung CT images after computed tomography (Zhu et al., 2021). By extracting lung CT image features and reducing the dimensionality of image features, the lung cancer labels are divided into two types: benign and malignant. Based on the improved gravity search algorithm, the deep neural network is optimized and the processed image information is brought into the optimized model to compare the prediction effects before and after optimization, The optimized model significantly improves prediction.

EXPERIMENTAL METHODS

Research Subjects

The NCBI database is an authoritative biotechnology information center that collects a large amount of biomedical data, including genes, proteins, and sequences. The data obtained from the NCBI database has high reliability and accuracy. This article collected data for 1,000 patients—500 males and 500 females—from the NCBI database. Patients' lungs mainly collect protein and cell data, and data analysis software is used to determine 812 benign cases and 188 malignant cases. A total of 80% of these patients are used as the training set, and the rest are used as the test set. This article mainly studies the age, gender, etiology, symptoms, treatment plan, hemoglobin, platelets, blood pressure, blood lipids, and 25 other indicators of lung cancer patients. Conventional blood tests are used and, based on the test report, the patient's blood cell status can be clearly understood, making scientific symptoms and etiological judgments and thus providing a basis for subsequent diagnosis.

In the NCBI database, due to differences in diagnostic methods and medical records, there are significant differences in case entry information and certain missing indicators for patients (Wang et al., 2022). At the same time, the patient information indicates the diagnosis of lung cancer but not which subtype it belongs to. By removing the patient data samples with missing information, the final available patient data includes 750 cases.

A total of 60 data characteristic indicators are selected in this paper. Table 1 shows the basic characteristics of discrete categorical variable indicators, mainly including coronary heart disease, diabetes, alcohol, hypertension, hepatitis, tuberculosis, and smoking, which are basically related to patients' living habits and existing chronic diseases. The basic characteristics of continuous variable indicators mainly include 53 indicators; Table 2 provides the most commonly used 8 indicators, all of which are the most common detection indicators for patients, such as blood routine, serum, and vital signs. It can be seen from Table 2 that each variable indicator has certain data loss, but there is no loss of age and respiration. Table 1 and Table 2 are sorted according to the loss rate of variable indicators. Among the discrete categorical variable indicators, coronary heart disease and diabetes have the most loss, while among the continuous numerical variable indicators, total cholesterol and triglyceride have the most loss, accounting for 41.3% and 32.1% respectively.

When there is a serious imbalance in the data information of the selected variable indicators, it will greatly affect the prediction results and the final result will point to one or several specific response variable indicators. This article uses K-nearest neighbor interpolation to fill in missing variable indicator data. During interpolation, all indicators are uniformly interpolated, while

Table 1. Basic characteristics of discrete categorical variable indicators

Feature	Amount of missing data	Missing percentage (%)	Number of categories
Coronary heart disease	32	17.2	2
Diabetes	34	12.8	3
Drink	18	9.4	4
Hypertension	14	6.4	2
Hepatitis	9	4.3	2
Pulmonary tuberculosis	6	3.2	2
Smoke	2	2.1	3

Table 2. Basic characteristics of continuous variable indicators

Feature	Amount of missing data	Missing percentage (%)
Total cholesterol	70	41.3
Triglyceride	59	32.1
Albumin	36	10.2
Red blood cells (whole blood)	13	3.6
White blood cells (whole blood)	9	0.4
Total bilirubin	3	0.2
Age	0	0
Breathing	0	0

taking into account the characteristics of various variable indicators to make the interpolated results as close as possible to the original data (Wang et al., 2022). Figure 2 shows the data missing after interpolation. It can be seen that coronary heart disease, diabetes, total cholesterol, triglycerides and albumin are still missing significantly, of which total cholesterol is missing more, and variable index data is still unbalanced. When the data of variable indicators is imbalanced, it has a significant impact on the construction and prediction results of the prediction model. For relatively few variable indicators, it is not sensitive to the impact of the prediction model and is prone to classification errors. The SMOTE algorithm is a method used to solve sample imbalance problems. It increases the number of minority class samples in the training set by synthesizing new minority class samples. In this article, due to the serious imbalance in the data of variable indicators, it is necessary to use the SMOTE algorithm for data balancing. By increasing the number of minority class samples, the model's recognition ability for minority class samples can be improved, thereby improving the performance of the prediction model. Figure 3 shows the completeness of the balanced data; all data has reached completeness, thus ensuring the true effect of model construction.

Figures 2 and 3 show the integrity of the data after processing; most of the variable indicators have been effectively interpolated and balanced, thus ensuring the reliability and stability of the data. By using processed data, the constructed prediction model can more accurately assess the risk of lung cancer, providing strong support for clinical diagnosis and treatment.

Figure 2. Data integrity after interpolation processing

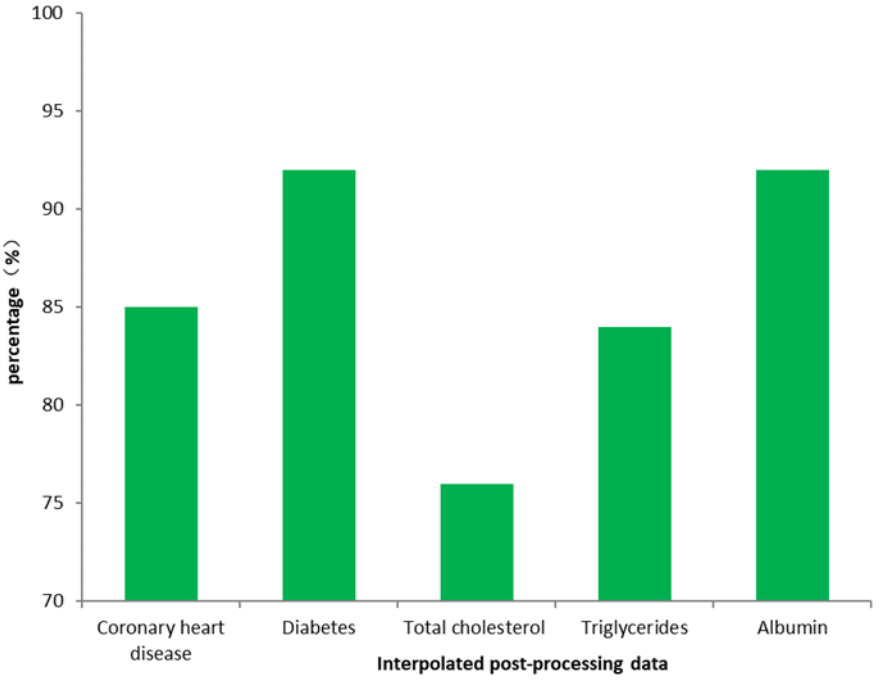
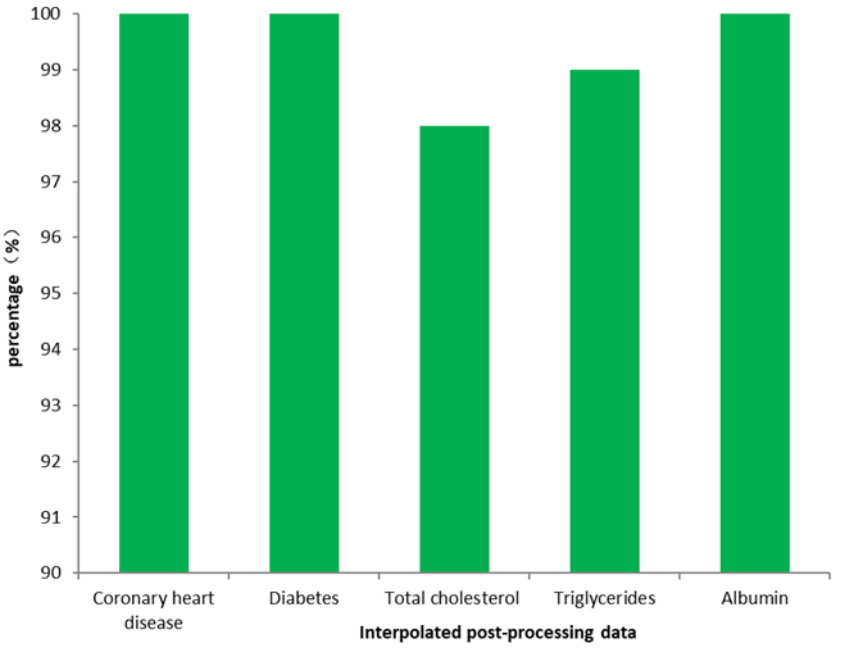


Figure 3. Balanced processing data integrity



Research Methods

After obtaining variable indicators for lung cancer patients, a lung cancer prediction model is constructed using machine learning algorithms. After processing the collected data with missing, outlier, and discrete data, machine learning algorithms (decision tree, logical regression, support vector machine) are used to build a lung cancer prediction model and cross validation and other technologies are used to evaluate the model to ensure the accuracy and reliability of prediction.

When establishing predictive models, the classification indicators of clinical data have high-dimensional features and different indicators may have interdependence and correlation (Zhao et al., 2017). This high-dimensional feature can lead to low computational efficiency, overfitting, and other issues. Therefore, it is necessary to screen multiple variable feature indicators to obtain effective variable features. The goal of feature selection is to select a minimum subset of features, so that the features contained in the subset can provide the optimal discriminative information for the target variable, thereby achieving the goal of improving prediction accuracy.

When screening variable features, the variance selection method is first used to calculate the intra group square difference of variable indicators and determine whether a single variable indicator has an impact on the classification indicators (Lu et al., 2018). Before calculating variance, in order to improve computational efficiency, the data should be normalized first and then the intra group variance of variable indicators should be calculated. The closed value should be set to 1. The calculation results show that all variable indicators have variances greater than 1, demonstrating the important value of variable indicator classification.

In this paper, the interactive information method is used to visualize the variable indicators and analyze the accuracy of the disease category in the determined variable indicator values (Pietrowska & Widlak, 2012). The Relief-F algorithm mainly removes invalid variable indicators by continuously updating the weights of the interval between variable indicators within the sample and the interval between variable indicators between classes. This article sets the sampling ratio to 100% for all samples but, overall, due to the small sample size and low computational efficiency requirements, the removal effect of all samples is better. The feature weight threshold is set to 0.006 and, through continuous sampling and iterative processing, variable indicators with a mean weight lower than 0.006 are removed. A total of 18 variable indicators are removed using the Relief-F algorithm, while 42 variable indicators are retained (Fan, 2020). Table 3 provides the names of variable indicators for filtered feature filtering.

Embedded feature selection is also the most commonly used variable indicator selection method. By combining variable indicator selection with classifier machine learning algorithms, variable indicator selection is carried out. This article uses a combination of LASSO and ridge regression algorithm to extract features by reducing the discreteness of variable indicator feature vectors. LASSO

Table 3. Filtered feature filtering partial variable indicator names

Serial number	Feature	Serial number	Feature
1	Smoke	9	Potassium
2	Total albumin	10	Calcium
3	Glucose	11	Hypertension
4	Carcinoembryonic antigen	12	Total cholesterol
5	Magnesium	13	Sodium
6	Drinking alcohol	14	Hepatitis
7	Total bilirubin	15	Breathing
8	Urea nitrogen	16	Lymphocyte count

is a linear model, which can realize feature selection through L1 norm of constraint parameters. In this article, LASSO extracts feature by reducing the discreteness of variable metric feature vectors, and gradually removes unimportant features during this process to obtain the final feature set. In the end, the LASSO method eliminated a total of 8 variable indicators and retained 52 variable indicators. Table 4 provides the names of variables and indicators for LASSO feature screening.

EXPERIMENTAL RESULTS

Support Vector Machine Lung Cancer Prediction Model

After obtaining the data features of lung cancer variable indicators, the support vector machine algorithm is used for data feature classification processing. After feature processing of the data, support vector machines are selected to build a lung cancer prediction model for the feature data (Pietrowska & Widlak, 2012). Support vector machine is a commonly used classification algorithm that has good performance in processing high-dimensional and nonlinear data. During machine learning, the dataset can be divided into a training set and a testing set in a 6:4 ratio for training and validating the model, respectively. In addition to accuracy, it is also necessary to consider indicators such as the recall rate and AUC value of the model. The recall rate reflects the model's ability to recognize positive samples, while the AUC value can evaluate the overall classification ability of the model.

In this paper, after selecting the data characteristics of variable indicators through the Relief-F filtering method, the filtered data characteristics are introduced into three support vector machine models—namely, linear kernel, linear separability, and polynomial kernel—and the prediction accuracy, recall, and AUC values are obtained (Table 5). Table 5 indicates that the linear kernel support vector machine model has a prediction accuracy of 0.827, a recall rate of 0.825, and an AUC value of 0.93, all of which are the maximum values of the three support vector machine models.

Table 4. LASSO feature screening partial variable indicator names

Serial number	Feature	Serial number	Feature
1	Basophil	9	Drink
2	White blood cells (whole blood)	10	Hematocrit
3	Hemoglobin	11	Glutamate dehydrogenase
4	Coronary heart disease	12	Serum sialic acid
5	Diabetes	13	Glutamate transaminase
6	Hypertension	14	Hepatitis
7	Phosphorus	15	Breathing
8	Breathing	16	Urea nitrogen

Table 5. Relief-F feature screening support vector machine test results

Indicator	Linear Kernel Support Vector Machine	Linear Separable Support Vector Machine	Polynomial Kernel Support Vector Machine
Accuracy	0.827	0.812	0.810
Recall	0.825	0.811	0.813
AUC value	0.93	0.92	0.92

Therefore, it can be considered that the performance of the linear kernel support vector machine is superior to the other two models.

Table 6 presents the test results of LASSO feature screening support vector machine. From Table 6, it can be seen that using LASSO feature screening support vector machine testing, the linear kernel support vector machine model has a prediction accuracy of 0.867, a recall rate of 0.853, and an AUC value of 0.94, which are superior to the other two support vector machine models. Therefore, it can be considered that the linear kernel support vector machine performs better than the other two models.

When using the Relief-F filtering method and the LASSO embedded method for feature selection of variable indicator data, the linear kernel support vector machine model yields the best prediction results (Zhou et al., 2019). This is mainly because the filtering and embedded selection of data features use the linear correlation method, so choosing the linear kernel support vector machine model is more optimal.

Random Forest Lung Cancer Prediction Model

This paper uses random forest algorithm to predict lung cancer and mainly analyzes the prediction accuracy, recall rate, and AUC value (Cloutier et al., 2021). The grid search algorithm is used to calculate the lung cancer prediction model of random forest. The grid search algorithm will perform cross validation on different parameter combinations to find the optimal parameter combination. The core parameters of random forest are selected. The decision tree of random forest is set to 65, the maximum depth of each decision tree is set to 4, the maximum number of features of each tree is set to 15, the minimum node sample is set to 10, the minimum leaf node sample is set to 4, and the remaining values are the default values of the system.

Table 7 shows the results of random forest test after screening of different characteristics. It can be seen from Table 7 that the random forest test produces excellent results. The prediction accuracy and recall rate in the random forest test results screened by Relief-F and LASSO features exceed 0.87, and the AUC value exceeds 0.97. At the same time, compared with the support vector machine model, the prediction accuracy, recall rate, and AUC value of the random forest model are significantly higher and the prediction results of the model are better (De Vos et al., 2015). Therefore, it is better to select the random forest model screened by Relief-F characteristics as the prediction of lung cancer classification.

Table 6. LASSO feature screening support vector machine test results

Indicator	Linear Kernel Support Vector Machine	Linear Separable Support Vector Machine	Polynomial Kernel Support Vector Machine
Accuracy	0.867	0.927	0.816
Recall	0.853	0.925	0.814
AUC value	0.94	0.93	0.92

Table 7. Random forest test results after screening of different characteristics

Indicator	Relief-F	LASSO
Accuracy	0.893	0.876
recall	0.882	0.872
AUC value	0.97	0.98

CONCLUSION

This article uses machine learning technology to analyze lung cancer medical data and establish a lung cancer prediction model, which is of great significance for the prevention, diagnosis, and treatment of lung cancer. The prediction model is constructed by the support vector machine and random forest machine learning methods to compare the prediction accuracy, recall, and AUC indicators of the two models. The main research achievements include:

- The linear kernel support vector machine model using Relief-F feature filtering has a prediction accuracy of 0.827, a recall rate of 0.825, and an AUC value of 0.93. The linear kernel support vector machine model using LASSO feature screening has a prediction accuracy of 0.867, a recall rate of 0.853, and an AUC value of 0.94. Research has found that using linear kernel support vector machine models yields the best prediction results.
- The prediction accuracy and recall rate of random forest test results screened by Relief-F and LASSO features exceeded 0.87, and the AUC value exceeded 0.97, both exceeding the support vector machine model. It is verified that the random forest model screened by Relief-F features has a better prediction effect on lung cancer classification.

This article uses machine learning technology to establish a lung cancer prediction model. Future research can explore how to apply this model to the diagnosis, screening, and treatment of lung cancer and combine it with traditional clinical diagnostic methods to provide doctors with more accurate diagnostic references.

ACKNOWLEDGMENT

The authors would like to extend sincere thanks to those who have contributed to this research.

DATA AVAILABILITY

The figures and tables used to support the findings of this study are included in the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENT

This work was not supported by any funding.

REFERENCES

Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2021). Lung cancer prediction and classification based on correlation selection method using machine learning techniques. *Qubahan Academic Journal*, *1*(2), 141–149. doi:10.48161/qaj.v1n2a58

Banerjee, N., & Das, S. (2020). *Prediction Lung cancer–In machine learning perspective*. *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA),* Gunpar, India. doi:10.1109/ICCSEA49143.2020.9132913

Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021). Prediction and classification of lung cancer using machine learning techniques. *IOP Conference Series. Materials Science and Engineering*, *1099*(1), 012059. doi:10.1088/1757-899X/1099/1/012059

Chauhan, D., & Jaiswal, V. (2017). An efficient data mining classification approach for detecting lung cancer disease. *2016 International Conference on Communication & Electronics Systems*, Coimbatore, India. doi:10.1109/CESYS.2016.7889872

Cloutier, M., Grégoire, Y., Choucha, K., Amja, A., & Lewin, A. (2021). Prediction of donation return rate in young donors using machine-learning models. *ISBT Science Series*, *16*(1), 119–126. doi:10.1111/voxs.12618

De Vos, B. D., De Jong, P. A., Wolterink, J. M., Vliegenthart, R., Wielingen, G. V., Viergever, M. A., & Išgum, I. (2015, March). Automatic machine learning based prediction of cardiovascular events in lung cancer screening data. In Medical Imaging 2015: Computer-Aided Diagnosis, 9414, 85-90. doi:10.1117/12.2082242

Fan, C., Xu, F., Qi, X., Li, C., & Yao, L. (2020). Classification of Alzheimer's disease based on brain MRI and machine learning. *Neural Computing & Applications*, *32*(4), 1927–1936. doi:10.1007/s00521-019-04495-0

Hosseinzadeh, F., Ebrahimi, M., Goliaei, B., & Shamabadi, N. (2012). Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. *PLoS One*, *7*(7), e40017. doi:10.1371/journal.pone.0040017 PMID:22829872

Jena, S. R., George, T., & Ponraj, N. (2019). Texture analysis based feature extraction and classification of lung cancer. *2019 IEEE International Conference on Electrical, Computer and Communication Technologies.* IEEE. doi:10.1109/ICECCT.2019.8869369

Jung, H., Oh, A., Apte, R., & Al-Lozi, J. (2017). Towards prediction of radiation pneumonitis arising from lung cancer patients using machine learning approaches. *Journal of Radiation Oncology Informatics*, *1*(1), 30–43. doi:10.5166/jroi-1-1-5

Liu, A., Xiao, Y., Wu, M., Tan, Y., He, Y., Deng, Y., & Tang, L. (2022, June). Diagnosis and classification prediction model of pituitary tumor based on machine learning. *Neural Computing & Applications*, *34*(12), 9257–9272. doi:10.1007/s00521-021-06277-z

Liu, P., Jin, K., Jiao, Y., He, M., & Fei, S. (2021). *Prediction of second primary lung cancer patient's survivability based on improved eigenvector centrality-based feature selection*. IEEE. doi:10.1109/ACCESS.2021.3063944

Lu, C. F., Hsu, F. T., Hsieh, K. L. C., Kao, Y. C. J., Cheng, S. J., Hsu, J. B. K., Tsai, P.-H., Chen, R.-J., Huang, C.-C., Yen, Y., & Chen, C. Y. (2018). Machine learning–based radiomics for molecular subtyping of gliomas. *Clinical Cancer Research*, *24*(18), 4429–4436. doi:10.1158/1078-0432.CCR-17-3445 PMID:29789422

Pietrowska, M., & Widłak, P. (2012). MALDI-MS-based profiling of serum proteome: Detection of changes related to progression of cancer and response to cancer treatment. *International Journal of Proteomics*, *2012*, 1–10. doi:10.1155/2012/926427 PMID:22900176

Priya, T. S., & Meyyappan, T. (2021). Disease prediction by machine learning over big data lung cancer. *International Journal of Scientific Research in Computer Science Engineering and Information Technology.*, *7*(1), 16–24. doi:10.32628/CSEIT206669

Sadhwani, A., Chang, H. W., Behrooz, A., Brown, T., Auvigne-Flament, I., Patel, H., Findlater, R., Velez, V., Tan, F., Tekiela, K., Wulczyn, E., Yi, E. S., Mermel, C. H., Hanks, D., Chen, P. C., Kulig, K., Batenchuk, C., Steiner, D. F., & Cimermancic, P. (2021). Comparative analysis of machine learning approaches to classify tumor mutation burden in lung adenocarcinoma using histopathology images. *Scientific Reports*, *11*(1), 16605. doi:10.1038/s41598-021-95747-4 PMID:34400666

Shahweli, Z. N., & Ban, N. D. (2018). In silico model for lung cancer prediction based on tp53 mutations using neural network. *Al-Nahrain Journal of Science*, *1*(1), 196–201. doi:10.22401/ANJS.00.1.26

Ubaldi, L., Valenti, V., Borgese, R. F., Collura, G., Fantacci, M. E., Ferrera, G., Iacoviello, G., Abbate, B. F., Laruina, F., Tripoli, A., Retico, A., & Marrale, M. (2021). Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. *Physica Medica*, *90*, 13–22. doi:10.1016/j.ejmp.2021.08.015 PMID:34521016

Wang, H. N., Zheng, L. X., Pan, S. W., Yan, T., & Su, Q. L. (2022). Image recognition of pediatric pneumonia based on fusion of texture features and depth features. *Computational and Mathematical Methods in Medicine*, *2022*, 1–10. doi:10.1155/2022/1973508 PMID:36060651

Wang, L., & Yuan, Y. (2019). A prediction strategy for academic records based on classification algorithm in online learning environment. *IEEE International Conference on Advanced Learning Technologies.* IEEE. doi:10.1109/ICALT.2019.00007

Yawen, X., & Jun, W. (2018). A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine*, *153*, 1–9. doi:10.1016/j.cmpb.2017.09.005 PMID:29157442

Zhao, S., Yu, J., & Wang, L. (2017). Machine learning based prediction rain metastasis of patients with iiia-n2 lung adenocarcinoma by a three-mirna nature. *Translational Oncology*, *11*(1), 157–167. doi:10.1016/j.tranon.2017.12.002 PMID:29288987

Zhou, J., Luo, L. Y., Dou, Q., Chen, H., Chen, C., Li, G. J., Jiang, Z.-F., & Heng, P. A. (2019). Weakly supervised 3D deep learning for breast cancer classification and localization of the lesions in MR images. *Journal of Magnetic Resonance Imaging*, *50*(4), 1144–1151. doi:10.1002/jmri.26721 PMID:30924997

Zhu, R., Dai, L., Liu, J., & Guo, Y. (2021). Diagnostic classification Lung cancer using deep transfer learning technology and multi-omics data. *Chinese Journal of Electronics*, *30*(5), 843–852. doi:10.1049/cje.2021.06.006