Journal of Intelligent Medicine and Healthcare
DOI: 10.32604/jimh.2022.032405
Article

Tech Science Press

Prognosis Analysis of Lung Cancer Patients

Yicheng Xie[1,*] and Jinyue Xia[2]

[1]School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[2]International Business Machines Corporation (IBM), NY, 10504, USA
*Corresponding Author: Yicheng Xie. Email: xieyicheng0516@163.com
Received: 25 April 2022; Accepted: 27 May 2022

Abstract: Lung cancer is now the most common type of cancer worldwide, with high levels of morbidity and mortality. The cost of treatment and emotional stress put a high burden on families and society. This paper aims to collect relevant information and provide predictive analysis for the prognosis of patients with lung cancer. Using the public data of SEER database and the method of machine learning, a model is constructed to predict the five-year survival of patients with lung cancer. The re-coding method is used for data processing, the eigenvalues are re-coded to adapt to the construction of the model, and the data are balanced by a variety of sampling methods to improve the applicability of the model. The construction method of the model is based on logistic regression, fully connected neural network, random forest and XGBOOST, evaluate the performance and select the optimal model. The results show that among the four constructed models, XGBOOST is selected as the optimal model with faster training speed, higher accuracy and the highest AUC value, it has advantages in memory occupation and time-consuming. The tumor stage and whether surgery or not and treatment difficulty have certain decisive factors for the prognosis of patients.

Keywords: Lung cancer; prognosis; machine learning

1 Introduction

According to the relevant data compiled by the International Agency for Research on Cancer, there were about 18.1 million new cancer cases and 9.6 million cancer deaths in 2018 [1], while in 2020, there were more than 19 million newly diagnosed patients and 9.6 million cancer deaths in the world, including about 3 million deaths in China. The existing statistics have shown that with the development of society, both cancer incidence and mortality rates are gradually rising in recent years, while lung cancer is at a high level among the incidence and mortality rates of various diseases in China, and the incidence and mortality rates of lung cancer have double high attributes [2]. Cancer is a malignant neoplastic disease, and its treatment is generally based on radiotherapy and chemotherapy, supplemented by other treatment methods, and the treatment time is long. Since the lung is a major organ, when the lung has cancerous lesions, it brings a series of problems, including but not limited

to the occurrence of respiratory system damage, metastasis of other organs, etc. Meanwhile, cancer treatment is expensive, which is a huge burden to the society and family.

Lung cancer is a malignant tumor originating from the bronchial mucosa or glands of the lung, with common symptoms such as coughing and blood in sputum. Generally speaking, the all-age standardized five-year survival rate of lung cancer patients is about 16.1%, and the general diagnosis is advanced lung cancer, which can only be treated by surgery and adjuvant radiotherapy and chemotherapy, making the family and social burden of lung cancer treatment heavy. Meta analysis data on the influencing factors of lung cancer in Chinese population showed that exposure to toxic substances, history of tuberculosis, history of respiratory diseases, smoking, psychiatric factors, family tumors history, passive smoking, exposure to lampblack, and alcohol drinking are all risk factors for lung cancer in Chinese population [3,4]; vegetable intake is a protective factor for lung cancer in Chinese population.

Based on the publicly available (SEER) medical database (without ethical review due to the absence of private information), certain inclusion and exclusion criteria are used to limit the statistical scope of data. The data are first subjected to a univariate analysis to initially screen for relevant parameters, including but not limited to demographic characteristics, disease-related characteristics, and final results, etc. Among them, categorization and algebraic proxies are performed mainly for differentiation class and survival months to accommodate adaptation and adjustment of the system learning parameters. Significant associated factors are screened out. And certain machine learning approaches are used to model them.

(1) According to the existing medical data and the collected patient data (desensitized) in SEER database, appropriate independent and dependent variables are selected according to certain processing methods, and the selection of data model parameters and final objectives are analyzed; finally determine the parameters and results of the selected model construction.
(2) According to the selected final target type, select the appropriate algorithm for model prediction and output the relevant evaluation parameters; after the construction of multiple models, the optimal results are selected and analyzed.

2 Machine Learning Theory

Machine learning (ML) is simply to let machines "learn" like humans, the machine itself does not have the ability to learn, that is, the so-called learning is a series of operations given to machines by people using modern computer technology, the process and the result of this giving is machine learning [5,6]. Machine learning has been developed from the 1960s to the present, and has emerged in various aspects of society, such as the underlying technology of text recognition, various recommendation algorithms, and beauty technology are all the results of machine learning [7,8]. Machine learning is divided into approaches based on different aspects according to different focus. Here, classified by learning modality, it can be divided into three categories: supervised learning, unsupervised learning, and reinforcement learning [11,12]. Different learning approaches represent different labeling categories in the original data and the result orientation after learning.

Machine learning is to find correlations in many data samples and present the results in a self-learning way. Data samples have multiple different attributes to describe, and each attribute has different values and categories (quantification and classification) [9,10]. When performing machine learning, it is necessary to understand the problem orientation and the data in the data sample in advance and to choose different processing methods to process the sample data [14–16]. Generally speaking, there are data preprocessing and quality improvement for data samples [17,18]; data

preprocessing is result-oriented and adjusts data according to different machine learning approaches to make data more suitable for machines; quality improvement is to avoid data where some samples of large data do not meet the cluster characteristics, such as outliers, blank data, etc. The quality improvement of data promotes the learning ability and model construction of machine learning, and avoids the occurrence of GIGO (garbage in, garbage out) to a certain extent [19–21].

After data preprocessing, according to the needs of different project development, different algorithms are selected to construct different types of results for the data (the model results output after training) [22]. The data is divided into training set and test set. Ideally, the model should be able to fit the data in the training set well, however, in practice there is no perfect model and there are always some errors. Errors are generally divided into empirical errors and generalization errors, which are errors generated on different data sets, and the essence is the generalization ability of machine learning model, which determines the prediction accuracy of machine learning model [23–25].

Once the classification model is generated, it is necessary to evaluate and analyze the performance of the classification model. Generally, the accuracy rate and recall rate cannot have both, and the F1 index is used for evaluation (the larger the F1 index, the better) [13].

3 Data Access and Preprocessing

3.1 Data Access

The data of this research is obtained by accessing the SEER database, and the specific options are Incidence-SEER Research Data, 18Registries, Nov2020Sub (2000–2018). As the leading factor of lung cancer is acquired and subject to multiple factors. The data from 2010 to 2015 were selected, and the selected control factors are age (over 25 years old), gender (male, female), tumor stage (TMN stage), tumor differentiation grade, influence of adjacent organs, number of lymph nodes sent for examination, number of positive lymph nodes in tissues sent for examination, tumor size, return visit status, survival month, and other attributes [26–28].

Where tumor stage used Derived AJCC Stage Group, 7th ed (2010–2015) and tumor differentiation grade was classified in 2017.

The overall inclusion criteria of the data are that the primary site (Site recode ICD-O-3) is the lung, age is over 25 years old, and there is no vacancy in the data items.

All data types, with a total of 31589 pieces of data, and the data of each characteristic are shown in Tab. 1.

Table 1: List of data

Characteristics	Data type	Number of categories	Data range
age	classification	13	
gender	classification	2	
tumor stage (TMN stage)	classification	12	
tumor grade	classification	4	
influence of adjacent organs	quantification		100–999

(Continued)

Table 1: Continued

Characteristics	Data type	Number of categories	Data range
degree of transfer	quantification		0–99
tumor size	quantification		0–999
number of lymph nodes sent for examination	quantification		0–99
number of positive lymph nodes in tissues sent for examination	quantification		0–99
survival month	quantification		0–95
surgical status	classification	2	
survival state	classification	2	

3.2 Data Conversion

Query the relevant medical data, for tumor TMN stage is according to the scope of the primary tumor, the degree of infiltration, whether the regional lymph node metastases, whether the regional lymph node metastases, whether there are blood and other metastases, and thus dividing the tumor into four major stages I, II, III, and IV. At the same time, abcd can also be marked in the lower right corner to distinguish the different states in each major stage [29].

In order to better perform machine learning (the text content requires natural language processing, which brings uncertain factors to the model training), the data are transformed to a certain extent, and the tumor differentiation level is newly coded. The results are shown in Tab. 2 below.

Table 2: Tumor differentiation grade recoding table

Original code	New code
Well differentiated; Grade I	0
Moderately differentiated; Grade II	1
Poorly differentiated; Grade III	2
Undifferentiated; anaplastic; Grade	3

The type conversion of tumor TMN stage data is also similar.

For the transition from the survival month to the final state, take the five-year survival as the boundary, when the survival month is greater than 60, it is converted to 1, otherwise it is converted to 0.

4 Model Construction and Optimization

4.1 Prognostic Model of Lung Cancer Based on Logistic Regression

The processed sample data are trained, and in order to ensure the repeatability of the experimental results, the value of random_state is fixed at 33 and the sample share is 0.25. After fixing random_state, the model constructed is consistently the same each time, the generated dataset is the same, and the splitting results are the same each time. The optimization algorithm selects one of the following

four types: newton-cg, liblinear, lbfgs, liblinear_l1, while iteratively finding the optimal regularization coefficients. The first three all use L2 regularization, and the fourth one uses L1 regularization.

The corresponding graphs of the four optimization algorithms are shown in Fig. 1.

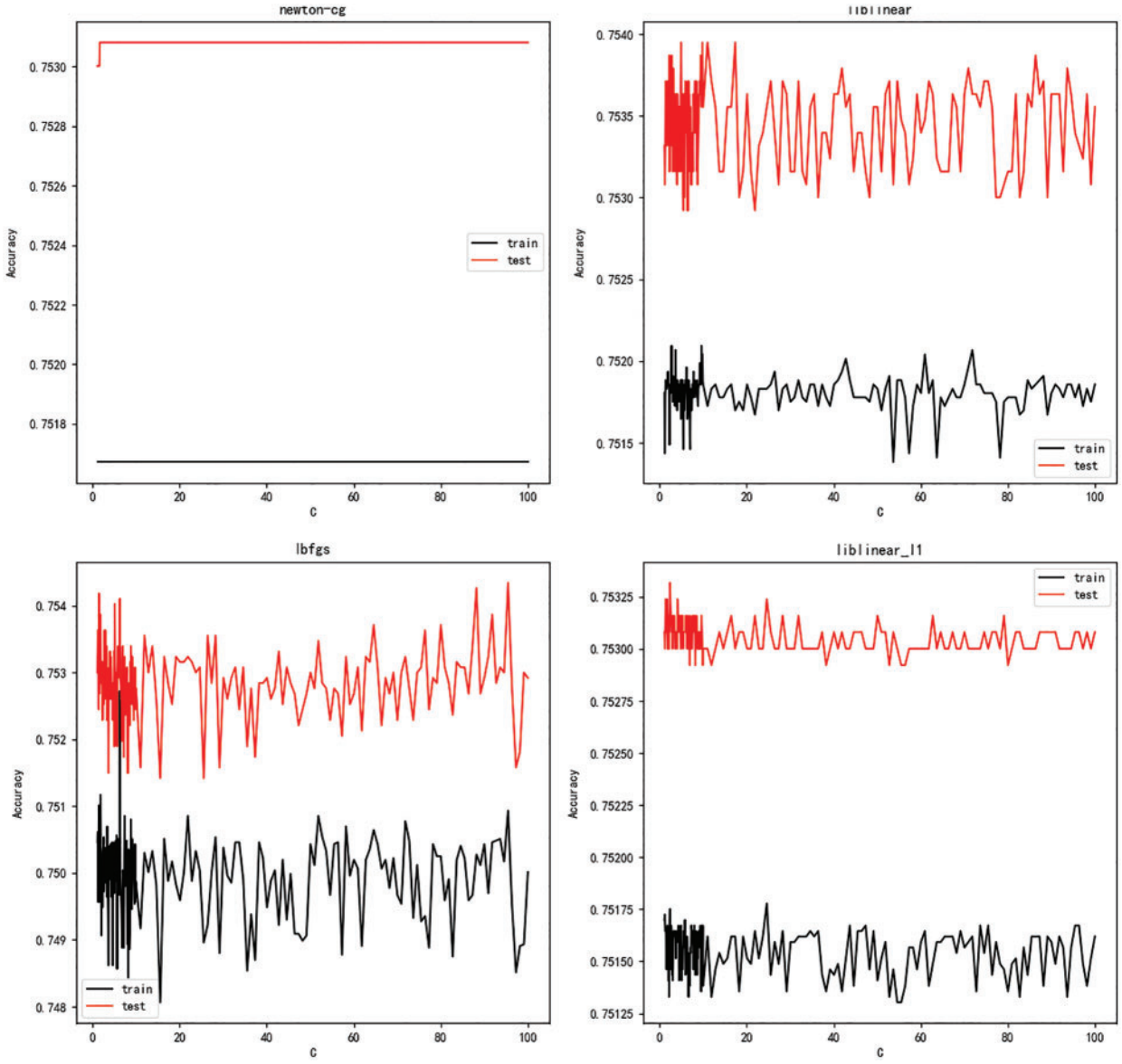

Figure 1: Graph of each optimization algorithm

After training, the optimal optimization algorithm is L-bgfs, the corresponding regularization coefficient is 95.459, the accuracy of training set is 0.751, the accuracy of training set is 0.751. The model evaluation results are shown in Tab. 3 below.

Table 3: Evaluation effect of logistic regression

Category	Accuracy rate	Recall rate	F1	Sample size
Dead	0.79	0.69	0.74	6282
Alive	0.73	0.82	0.77	6380
Accurary			0.75	12662
macro avg	0.76	0.75	0.75	12662
weighted avg	0.76	0.75	0.75	12662

4.2 Prognostic Model of Lung Cancer Based on Fully Connected Neural Network

In order to ensure the repeatability of this prognostic model, the random seed number is 42 and the sample proportion is 0.25. The original data are balanced. Two hidden layers and one classification layer are used to complete the construction of neural network, where the classification layer is binary classification. Use torch from_ numpy converts arrays into tensors, and the converted training set is integrated with TensorDataset to fit the algorithm model. The data fed to the trainer are scrambled to improve the generalization ability of the model.

The optimizer uses the Adam optimizer to update the variables, and the step factor (learning rate) is chosen to be 0.001. The smaller step factor allows for better convergence performance, with the difference that the efficiency is slower during initial training. The loss function adopts cross entropy loss function. The number of training rounds is specified as 100.

According to the obtained optimal parameters, the final accuracy of the obtained convergence model is 0.756 on the test set. The effect evaluation table is shown in Tab. 4 below.

Table 4: Evaluation of fully connected neural networks

Category	Accuracy rate	Recall rate	F1	Sample size
Dead	0.81	0.66	0.73	6282
Alive	0.72	0.85	0.78	6380
Accurary			0.76	12662
macro avg	0.77	0.76	0.75	12662
weighted avg	0.76	0.76	0.75	12662

The ROC curve graph of this algorithm is shown in Fig. 2, the calculated AUC (area) = 0.756. The AUC value is lower than 0.85, the prediction effect of the model is general and similar to the logistic regression results.

4.3 Prognostic Model of Lung Cancer Based on Random Forest

In order to ensure the repeatability of this prognostic model, the number of random seeds is fixed at 13 and the sample share is 0.25. The parameters are adjusted in order to obtain better random forest parameters. The initial n_estimators are randomized and limited between 10 and 60. This parameter determines the number of trees in the random forest, and the more the number, the longer the memory overhead and training time. The maximum depth of the decision tree is limited to the interval of

1–10. The random forest is constructed using RandomForestClassifier function to find the optimal parameters.

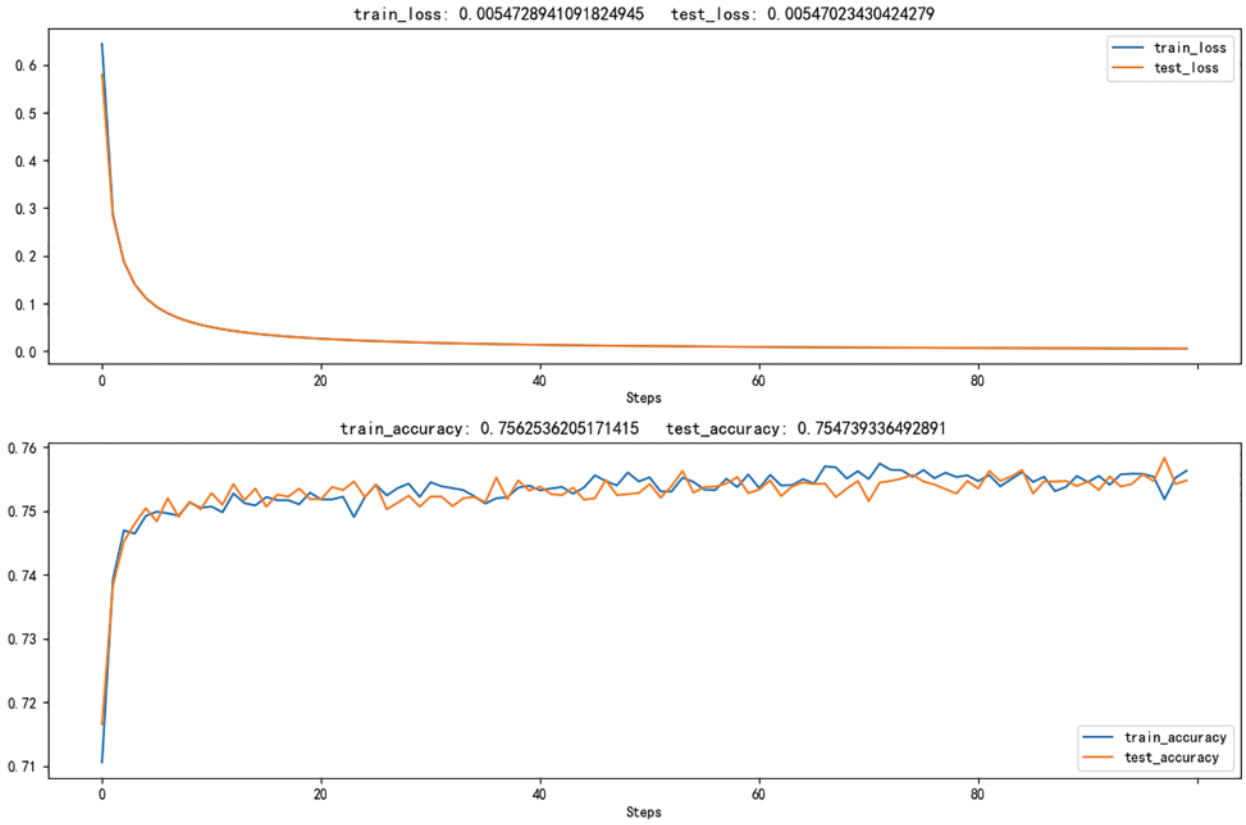

Figure 2: 100 rounds of training for fully connected neural networks

The best number of clusters is 37 trees with a maximum depth of 9. Its accuracy is 0.840 on the training set and 0.829 on the test set. The random forest model is evaluated and its specific results are shown in Tab. 5 below.

Table 5: Evaluation of random forest effect

Category	Accuracy rate	Recall rate	F1	Sample size
Dead	0.87	0.77	0.82	6275
Alive	0.80	0.88	0.84	6385
Accurary			0.83	12660
macro avg	0.83	0.83	0.83	12660
weighted avg	0.83	0.83	0.83	12660

The ROC curve graph of this algorithm is shown in Fig. 3, and the calculated AUC (area) = 0.828. The prediction of the model is better. In this experiment, random forest model is better than the fully connected neural network and logistic regression.

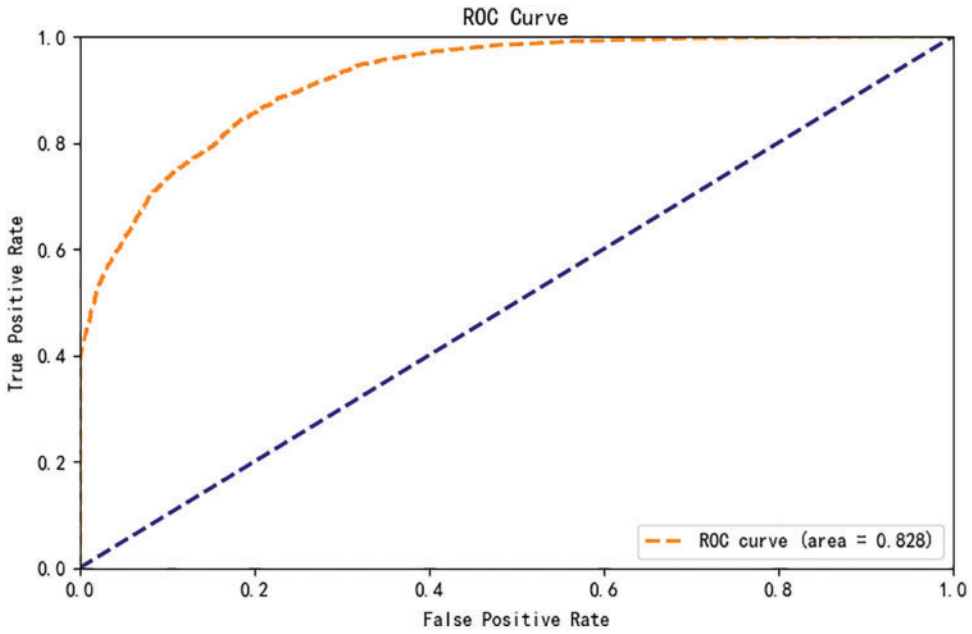

Figure 3: ROC curve of the optimal random forest model

4.4 Prognostic Model of Lung Cancer Based on XGBOOST

The full name of XGBOOST is eXtreme Gradient Boosting, which was proposed by Dr. Tianqi Chen of the University of Washington and used in the Higgs subsignal recognition competition of Kaggle, where it has excellent efficiency and high prediction accuracy [8]. Its essence is an integrated algorithm, which optimizes the loss function and makes a quadratic Taylor expansion. A regularization is added to the function in addition to the objective function so as to obtain the optimal value when solving it.

In the process of generating trees, the algorithm is generated one by one, and we expect that the overall prediction effect will be improved when each tree is generated. For each subtree added to it, when there are too many leaf nodes, it will potentially cause an overfitting risk to the algorithm, and we need to control this risk by introducing a penalty term to form a constraint on the objective function. Therefore, in practice, each extension needs to traverse all the schemes listed. For a particular expansion, we need to determine the amount of loss before and after the segmentation, traverse all the results, and select the expanded model with the largest change. When looking for the optimal segmentation point, different from the random forest strategy, a few segmentation point candidates are selected according to the percentile results, and the calculation method is used to obtain the optimal value between the segmentation points, so as to realize a similar greedy strategy. The memory consumption and time consumption are optimized compared with the original greedy strategy, which is not too far from the result. At the same time, the feature columns are sorted and placed in memory in blocks, which can be reused in subsequent iterations, thus enabling parallel processing.

The best number of clusters is 52 trees with a maximum depth of 7. Its accuracy is 0.904 on the training set and 0.872 on the test set. The XGBOOST model is evaluated and its specific results are shown in Tab. 6 below.

Table 6: Evaluation of XGBOOST effect

Category	Accuracy rate	Recall rate	F1	Sample size
Dead	0.84	0.91	0.88	6275
Alive	0.91	0.83	0.87	6385
Accurary			0.87	12660
macro avg	0.87	0.87	0.87	12660
weighted avg	0.87	0.87	0.87	12660

The ROC curve graph of this algorithm is shown in Fig. 4, and the calculated AUC (area) = 0.892. The AUC value is higher than 0.85, and the model has a good prediction effect. It is better than other ways of constructing the model in this experiment.

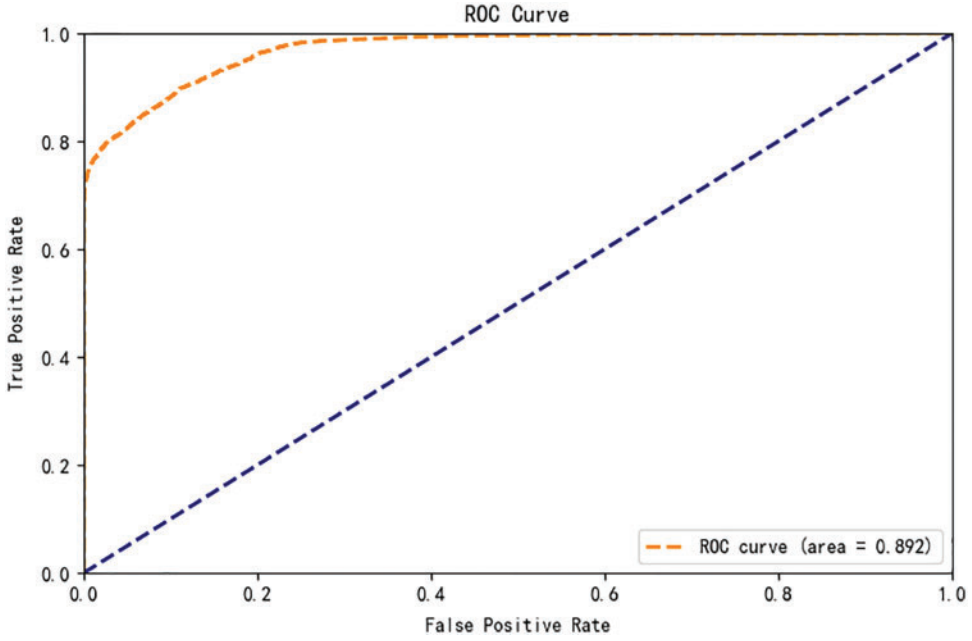

Figure 4: ROC curve of XGBOOST model

5 Model Selection

5.1 Optimal Model Construction Method

The optimal model is selected based on the overall consideration of training time, loading and running time, memory occupation, etc. Priority is given to XGBOOST and random forest, which have the advantages of relatively low memory occupation and short training time. At the same time, according to the graph of time and accuracy in the previous paper, XGBOOST has the advantages of short loading time and high accuracy, and XGBOOST is finally selected as the optimal model construction method. As shown in Fig. 5, four model construction methods are compared.

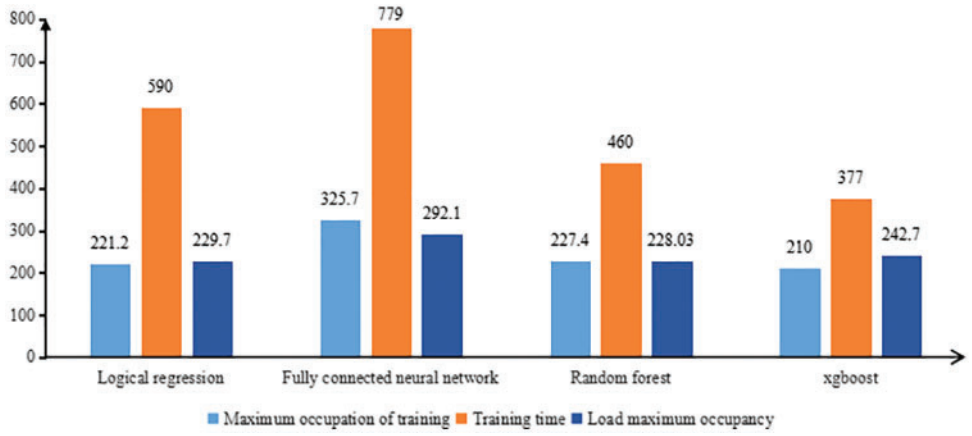

Figure 5: Comparison of four model construction methods

5.2 Optimal Model Selection

The optimal decision tree is obtained by outputting the decision tree in XGBOOST during training, and the retrospective results showed that tumor size appeared most frequently in all trees throughout the five-year survival of lung cancer patients, followed by age and number of lymph nodes and tumor grade. Fig. 6 shows the frequency of various features in xgboost.

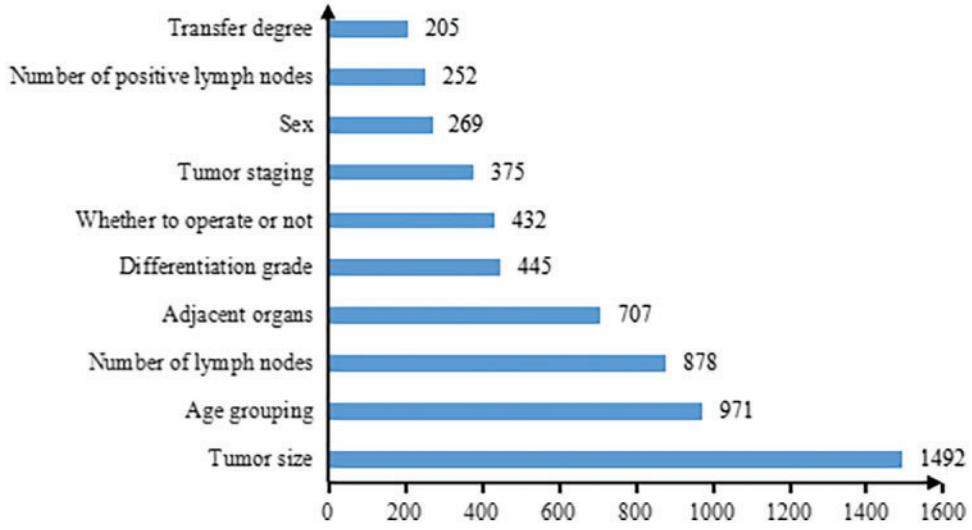

Figure 6: Frequency of features in xgboost

Looking at the final generated decision tree (see the attached figure for the decision tree), the tumor stage has a decisive effect, the more advanced its stage, the more likely its five-year survival, which is also the same as the conclusion of medical studies. At the same time, the next level is whether surgery or not or not, which shows that surgery should be performed as soon as possible after diagnosis to improve the possibility of survival. Subsequently, the two characteristics of tumor size and metastasis are related to the difficulty of treatment of lung cancer. The smaller the tumor and lower the degree of metastasis, the better the treatment effect and the higher the possibility of survival.

In summary, XGBOOST is selected as the means to build the model, while the prognosis of lung cancer should be adhered to early detection and treatment, and early surgical intervention to enable patients to have a greater possibility of survival. To extend, we can improve the frequency of physical examination and increase physical examination items according to the level of personal commitment in daily life to better deal with diseases.

6 Conclusion

The data in this paper are selected from the medical data of patients diagnosed with lung cancer from 2010 to 2015 in SEER database, combined with machine learning to analyze the medical data of patients, aiming to provide some reference significance for subsequent patients diagnosed with this type of cancer to help clinicians with decision aids.

The model is constructed based on logistic regression, fully connected neural network, random forest and XGBOOST to understand the applicable scope and performance of the model. Among them, both random forest and XGBOOST are algorithms combining integrated learning and decision tree, therefore, in binary classification problem, integrated learning combined with decision tree has certain advantages over other methods.

The memory consumption and time consumption based on logistic regression and fully connected neural network are higher than those based on the other two methods. In terms of performance, for binary classification problems, random forest and XGBOOST decision tree based learning methods have more advantages.

The best model in this paper is constructed based on XGBOOST method, which can improve the accuracy and AUC of model prediction compared with other methods. At the same time, the timely detection and treatment of lung cancer can make patients survive better, and early surgical treatment is a good means to prolong patients' lives. Extending to other diseases can increase the frequency of physical examination to reduce the risk of disease and improve individual survival.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] F. Bray, J. Ferlay and I. Soerjomataram, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] R. Zheng, S. Zhang and H. Zeng, "Cancer incidence and mortality in China, 2016," *Journal of the National Cancer Center*, vol. 2, no. 1, pp. 1–9, 2022.

[3] L. Q. Zeng, S. J. Xia and R. H. Peng, "Meta-analysis of factors influencing lung cancer among Chinese people, 2006–2016," *South China Journal of Preventive Medicine*, vol. 44, no. 5, pp. 431–435, 2018.

[4] S. Zhang, "Effect of building national demonstration areas for comprehensive prevention and control of non-communicable diseases: A case study of a Mega-city," M.S. dissertation, Chinese Center for Disease Control and Prevention, China, 2019.

[5] C. Yin, "Application of machine learning in cancer diagnosis," M.S. dissertation, University of Electronic Science and Technology of China, China, 2020.

[6] D. V. Cicchetti, "Neural networks and diagnosis in the clinical laboratory: State of the art," *Clinical Chemistry*, vol. 38, no. 1, pp. 9–10, 1992.

[7] K. Kourou, T. P. Exarchos and K. P. Exarchos, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.
[8] T. Chen and C. Guestrin, "XGBOOST: A scalable tree boosting system," in *Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, pp. 785–794, 2016.
[9] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, New York: Springer, vol. 4, no. 4, pp. 738, 2006.
[10] I. H. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques with Java implementations," *ACM Sigmod Record*, vol. 31, no. 1, pp. 76–77, 2002.
[11] J. N. Wang, Z. Xu and J. Lin, "Clinical characteristics and prognosis analysis of patients with liver metastasis from breast cancer: A retrospective study based on SEER database," *Chinese Journal of Breast Disease (Electronic Edition)*, vol. 12, no. 4, pp. 202–208, 2018.
[12] H. Li, *Statistical Learning Method*, Beijing, China: Tsinghua University Press, pp. 96–102, 2012.
[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[14] N. E. Sondak and V. K. Sondak, "Neural networks and artificial intelligence," *ACM*, vol. 21, no. 1, pp. 241–245, 1989.
[15] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.
[16] W. Wang, "On the control of continuous stirred tank reactor," M.S. dissertation, Beijing Jiaotong University, China, 2013.
[17] J. M. Zhao and Y. M. Wang, "The study of learning rate based on BP neural network," *Microcomputer Applications*, vol. 34, no. 8, pp. 89–92, 2018.
[18] Z. S. Cui and J. R. Wu, "Application of data classification based on random forest," *Journal of Shanxi Datong University (Natural Science)*, vol. 35, no. 5, pp. 31–33, 2019.
[19] S. H. Song, W. J. Liu and H. T. Shi, "Research on road icing prediction based on random forest," *Technology and Information*, vol. 28, pp. 131–132, 134, 2020.
[20] Y. F. Yan, "Research and application of regression model method based on decision forest," Ph.D. dissertation, Zhejiang University, China, 2019.
[21] L. T. Zheng, "Research on annual air conditioning load forecasting of shopping malls in pearl river delta region based on machine learning method," Ph.D. dissertation, South China University of Technology, China, 2019.
[22] S. B. Yang, "A software enterprise engineer's job performance based on machine learning approach evaluation application research," M.S. dissertation, Qingdao University of Science and Technology, China, 2020.
[23] X. Zhao, "Research on forest aboveground biomass estimation based on airborne LiDAR data," M.S. dissertation, Xi'an University of Science and Technology, China, 2020.
[24] Y. S. Wang and S. T. Xia, "A survey of random forests algorithms," *Information and Communications Technologies*, vol. 12, no. 1, pp. 49–55, 2018.
[25] E. M. O. Silveira, S. H. G. Silva and F. W. Acerbi-Junior, "Object-based random forest modelling of aboveground forest biomass outperforms a pixel-based approach in a heterogeneous and mountain tropical environment," *International Journal of Applied Earth Observation and Geoinformation*, vol. 78, pp. 175–188, 2019.
[26] L. Ge, W. Ge and Y. G. Wu, *The AJCC Tumor Staging Manual*, Beijing, China: China Medical Science Press, 2009.
[27] N. V. Chawla, K. W. Bowyer and L. O. Hall, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.
[28] S. Rodda, U. S. R. Erothi, "Class imbalance problem in the network intrusion detection systems," in *2016 Int. Conf. on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, India, pp. 2685–2688, 2016.
[29] Y. Liu, N. V. Chawla and M. P. Harper, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech & Language*, vol. 20, no. 4, pp. 468–494, 2006.