Theory and Practice of Science and Technology
2025, VOL. 6, NO. 6, 125-130
DOI: 10.47297/taposatWSP2633-456925.20250606

Application of the Random Forest-MLP Algorithm in the Recognition of Esophageal Cancer

Zhixiang Xu*, Wenpei Zhang, Mingming Li, Weihao Li, Weihong Chen

School of Mathematics and Statistics, Kashi University, Kashi, Xinjiang, 844000, China

ABSTRACT

This research focuses on the application of the Random Forest (RF) and Multilayer Perceptron (MLP) combined model in the field of esophageal cancer recognition. The incidence of esophageal cancer is relatively high in China, and accurate prediction is of great significance for improving the survival rate of patients. With the help of 85 types of esophageal cancer data provided by the Kaggle platform, which cover rich information such as clinical features, pathological diagnosis results, treatment methods and prognostic data, this study first uses random forest to carry out feature screening and accurately identifies the key features that have a significant impact on esophageal cancer prediction. Then, these selected features are input into the MLP model for training and verification. The experimental results show that the RF-MLP model significantly outperforms the traditional decision tree, support vector machine and adaptive boosting algorithm in terms of accuracy (as high as 98.746%) and precision (reaching 100%), fully demonstrating its great potential in the early diagnosis of esophageal cancer, building a solid and effective technical support for the risk prediction of esophageal cancer, and effectively promoting the development process of related research and clinical application.

KEYWORDS

Esophageal cancer; Random forest; Multilayer perceptron; Machine learning

1 Introduction

Esophageal cancer is a malignant tumor that threatens human health. There are more than 600,000 new cases and more than 370,000 deaths in the world each year. China is one of the countries with high incidence. Because of its hidden early symptoms, accurate prediction is of great significance. In 1940, Wu Yingkai 's team completed the first case of esophageal cancer resection + gastroesophageal reconstruction in China. However, the conventional prediction model has limitations, and machine learning has brought breakthroughs to this end.

Nowadays, machine learning has brought a new direction for the prediction of esophageal cancer, and many scholars have carried out research. Li Housheng[1]used the improved algorithm to optimize the random forest model to improve the prediction accuracy ; wu Yuefeng et al.[2] summarized the existing models and discussed the difficulties and methods ; zhuang et al.[3] confirmed that the deep learning method is better in accuracy and other indicators ; xu Jinye et al.[4] found that the XGBoost model has better predictive performance for thoracic esophageal squamous cell carcinoma after surgery ; li Jiayi et al.[5] explored the application of intelligent medicine in tumor diagnosis and treatment ; jiang Lianghui and Zhou Changhong[6] analyzed the application and prospect of DL in the diagnosis of digestive tract diseases. Deng et al.[7] successfully applied convolutional neural network to diagnose esophageal cancer images ; wang et al.[8] used artificial bee colony algorithm to improve the prediction accuracy ; li Baohua et al.[9] carried out multivariate regression analysis to verify the stability of the model ; sheng et al[10] predicted the early mortality of liver metastasis of esophageal cancer ; thavanesan et al.[11] predicted the treatment pathway ; navamayooran et al.[12] identified key decision variables ; macomber et al[13] predicted pathological complete remission ; jorn Jan van de Beld et al[14] predicted postoperative complications ; Xue-Feng L et al.[15] predicted pathological remission based on radiomics, and promoted the development of esophageal cancer prediction research from multiple dimensions.

By comparing various machine learning methods, this paper finds that the random forest-neural network combination has the best effect, which not only improves the prediction accuracy, but also provides support for academic research and clinical practice of esophageal cancer risk prediction, and promotes the development of related research and treatment.

2 RF-MLP model

2.1 Random forest

Random forest is an ensemble learning algorithm based on decision tree, which is widely used in the field of machine learning. Its core principle mainly depends on the self-help sampling technology, assuming that the original training data set is, through self-help sampling, from the original training data set to continuously generate multiple sub-data sets with significant differences. The process of self-sampling can be abstracted as follows : for the original data set of size, a sample is randomly selected to form a new sub-data set every time, and the process is repeated to obtain a sub-data set.

In subsequent operations, an independent decision tree is constructed for each sub-dataset. In the construction process, each node randomly selects some features for splitting, and the feature selection is usually measured according to the information gain. The formula is :

$$IG(D,A)=H(D)-\sum_{v\in Values(A)}\frac{|D_v|}{|D|}H(D_v) \tag{1}$$

It is precisely because of this randomness in node feature selection that there is a certain degree of difference between each final decision tree, which avoids the excessive similarity between decision trees. When the random forest is used for prediction, the voting method is used for the classification problem, that is :

$$\hat{y}=\operatorname{argmax}_{c\in C}\sum_{i=1}^{N}I(\widehat{y_i}=c) \tag{2}$$

For the regression problem, the average method is used, and the formula is :

$$\hat{y}=\frac{1}{N}\sum_{i=1}^{N}\widehat{y_i} \tag{3}$$

Random forest has significant advantages, which can efficiently process high-dimensional data, reduce the risk of over-fitting, have good anti-noise ability, and evaluate the importance of features. It measures the importance of features by calculating the average impure reduction caused by features in all decision trees. Let F be a feature, the specific calculation formula is as follows :

$$Importance(F)=\sum_{t=1}^{N}\frac{1}{|D_t|}\sum_{(x,y)\in D_t}[impurity(node)-\sum_{j=1}^{k}\frac{|D_{t,j}|}{|D_t|}impurity(D_{t,j}) \tag{4}$$

Random forests perform well in many fields such as medical diagnosis, financial risk prediction, and image recognition, providing an efficient and accurate solution for the analysis and prediction of complex data.

2.2 MLP neural network

Multilayer perceptron neural network (MLP) is of great significance in the history of neural network development. In 1961, Frank Rosenblatt proposed MLP, which laid the foundation of classical neural network architecture. However, early MLP training was difficult. It was not until 1986 that Rumelhart, Hinton and Williams proposed a back propagation algorithm to achieve a breakthrough. The algorithm uses the chain rule to optimize the weight parameters, and its core principle can be explained by mathematical formulas. For the error term of the lth layer neuron :

$$\delta^{(l)}=(W^{(l+1)T}\delta^{(l+1)})\odot\dot{f}(net^{l}) \tag{5}$$

It has laid a solid foundation for the learning algorithm of the MLP, enabling a leap-forward improvement in the training performance of the MLP.

The structure of the MLP is distinctly characterized and is mainly composed of an input layer, hidden layers, and an output layer in an orderly manner. Among these, the hierarchical architecture is its core characteristic. Assume that the input layer is $x\in R^n$,and the linear transformation of the first hidden layer (the 1-th hidden layer) is as follows:

$$net^{(l)}=W^{(L)}\cdot out^{(l-1)}+b^{(l)} \tag{6}$$

After being processed by the activation function f(·),the output is:

$$out^{(l)}=f(net^{(l)}) \tag{7}$$

This layered architecture gives MLP a strong nonlinear mapping capability through a composite nonlinear transformation, which can be expressed as :

$$\hat{y}=f_L(\cdots f_2(f_1(w_1x+b_1)+b_2)\cdots+b_L) \tag{8}$$

More importantly, MLP has adaptive learning ability. In the training process, the prediction error is measured by the loss function, and the weight parameters are updated by the gradient descent method :

$$W^{(l)}\leftarrow W^{(l)}-\eta\cdot\frac{\partial L}{\partial W^{(l)}},b^{(l)}\leftarrow b^{(l)}-\eta\cdot\frac{\partial L}{\partial b^{(l)}} \tag{9}$$

This process allows MLP to adaptively adjust weights according to data characteristics, optimize learning, and play a key role in data analysis in the fields of image recognition and natural language processing.

2.3 RF-MLP combination model

The RF-MLP combined model deeply integrates the advantages of random forest and multi-layer perceptron, and provides an efficient solution for data prediction. At runtime, the features are extracted by random forest and the key feature set is screened, and then input into the MLP model to train the prediction. This combination takes advantage of random forest dimensionality reduction and feature screening capabilities, combines the advantages of MLP nonlinear fitting, and builds accurate models based on high-quality features. It performs well in complex prediction and classification tasks, and promotes the development of machine learning.

3 Data sets

3.1 Sources of data

The esophageal cancer data used in this study were from the Kaggle platform, covering 85 kinds of data such as clinical characteristics, pathological diagnosis, treatment methods and prognosis of patients, providing strong support for the study.

3.2 Data preprocessing

In order to ensure data quality and analysis effectiveness, this study implements rigorous data preprocessing. The missing values are processed by the mean filling method to ensure data integrity. In view of the large number of variables in the data set, the feature that the absolute value of the correlation with the target variable is less than 0.1 is discarded to achieve dimensionality reduction. The processing results are shown in Figure 1 :

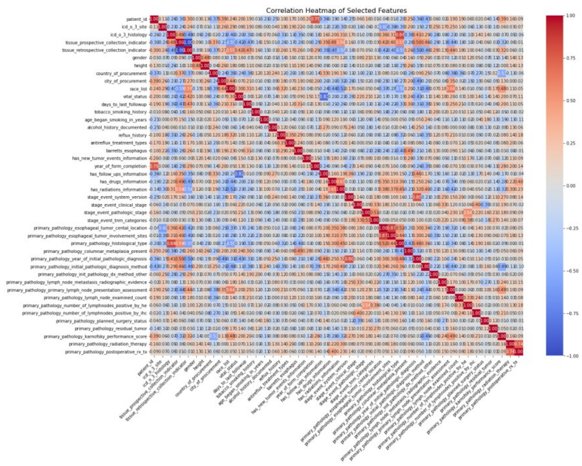

Figure 1 Correlation heat map of selected features

The correlation heat map shown in Figure 1 intuitively presents the results of feature screening, clearly shows the correlation between the features, and provides a strong visual basis for data preprocessing decision-making.

4 The establishment of RF-MLP model

4.1 Establishment of feature engineering

The identification of esophageal cancer is affected by many factors, and the data related to it contains extremely complex and diverse information. It covers the patient 's follow-up information, such as ' has _ new _ tumor _ events _ information ' and ' days _ to _ last _ followup ' ; it also includes vital signs such as ' vital _ status ' and so on, which are characterized by obvious diversity and complexity.

In order to improve the prediction accuracy of esophageal cancer, random forest is used to construct feature engineering for feature selection, and key feature combinations are selected. Only features with importance greater than 0.02 are retained, as shown in Figure 2.

It can be seen from the result chart that a total of 18 features with importance greater than 0.02 were identified.

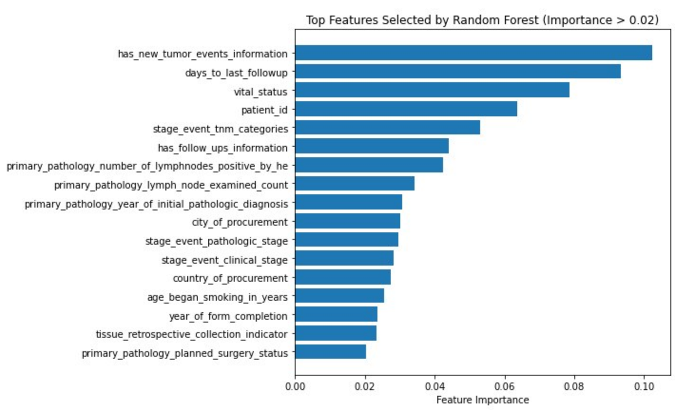

Figure 2 Main features of random forest selection

Among them, ' has _ new _ tumor _ events _ information ' is the most critical and may play a decisive role in the prediction of esophageal cancer. The importance of 'days _ to _ last _ followup ' and 'vital _ status ' is the second, which is an important part of the prediction system. Features such as ' patient _ id ', ' stage _ event _ tnm _ categories ' and ' has _ follow _ ups _ information ' also have reference value. Features such as ' primary _ pathology _ lymph _ node _ examined _ count ' are slightly less important, but still help to improve the accuracy of esophageal cancer prediction.

The partial dependency graph of the first two features is shown in Figure 3 :

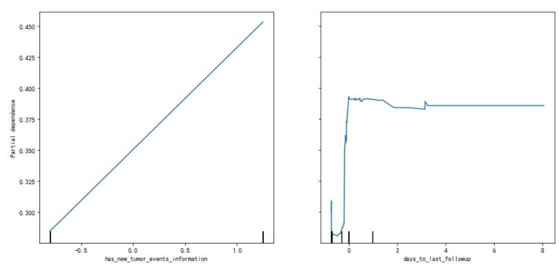

Figure 3 The partial dependency graph of the first two features

It can be seen from Fig. 3 that the left feature is positively correlated with the prediction (the higher the value, the higher the prediction probability), and the right short-term follow-up (0-2 days) is sensitive to the influence, and the long-term (> 2 days) tends to be stable, showing the dependence pattern of the feature on the model prediction.

4.2 Model training and validation

In the stage of model training and verification, the data set after dimensionality reduction is divided into 70 % training set and 30 % test set by stratified sampling method to evaluate the generalization ability and accuracy of the model, and to balance the training adequacy and verification effectiveness. In the training process, the multi-layer perceptron (MLP) is selected as the neural network model, and a variety of techniques are used to analyze the training process in depth. See Figure 4 :

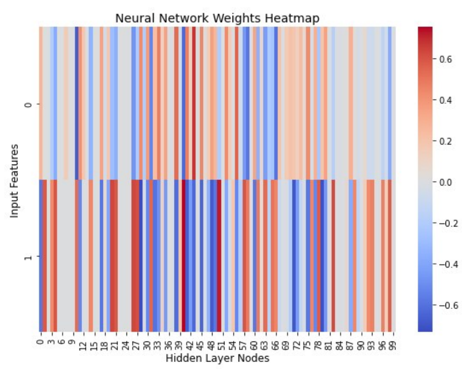

Figure 4 MLP network weight heat map

Fig. 4 The neural network weight heat map intuitively shows the weight distribution of input features during network learning. Through the color distribution of the heat map, the influence degree and direction of each input feature on the hidden layer nodes can be quickly understood, which provides a key basis for analyzing the model learning mechanism.

The decision boundary of the MLP of feature 1 is shown in Figure 5 :

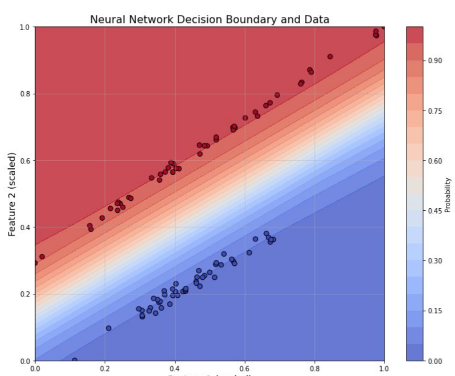

Figure 5 Decision boundary graph

Decision boundary visualization is the key to understand the model classification mechanism and evaluate the performance. In the study of neural networks, it helps to gain insight into the model learning process. It can be seen from the decision boundary diagram in Fig.4 that the nonlinear shape representation model has the ability to capture complex data patterns. The visualization results show that the model can accurately distinguish data points in most regions, and the fitting effect is good, which provides a basis for model evaluation and optimization, and further verifies its effectiveness in esophageal cancer recognition.

4.3 Comparison of models

In order to evaluate the performance of different classification models on the data set, we selected four commonly used machine learning models for comparison : Decision Tree, Support Vector Machine, SVM, Neural Network and AdaBoost. We apply each model to the training set and evaluate its performance on independent test sets.

We trained each model and calculated the following important performance indicators : Accuracy, Recall, Precision, F1 Score and AUC.These indicators can fully reflect the comprehensive performance of the model in the classification task. The ROC curve of the test set is shown in Figure 6 :

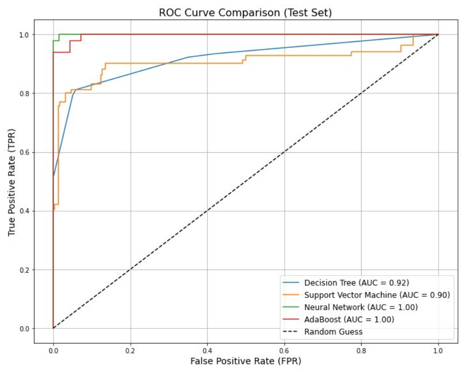

Figure 6 ROC curve comparison chart

From the comparison of the ROC curves of the four models in Figure 6, it can be seen that the neural network model is particularly prominent. The curve rises almost vertically from the lower left corner to TPR close to 1, and then extends horizontally to the right. It has the best performance among the four curves, fully demonstrating excellent classification ability and effectively distinguishing positive and negative examples.

After completing the training of each model in the training set, the performance evaluation was performed on the independent test set. The key indicators such as Accuracy, Recall, Precision, F1 Score and AUC were calculated. The results are shown in Table 1 and Figure 7 :

Table 1 Multi-model comparison data table

model types	accuracy	recall	precision	F1	AUC
Decision Tree	0.89548	0.81176	0.88462	0.84663	0.91804
Support Vector Machine	0.90635	0.80000	0.92643	0.85859	0.90489
Neural Network	0.98746	0.96471	1.00000	0.98204	0.99970
AdaBoost	0.96990	0.91529	1.00000	0.95577	0.99673

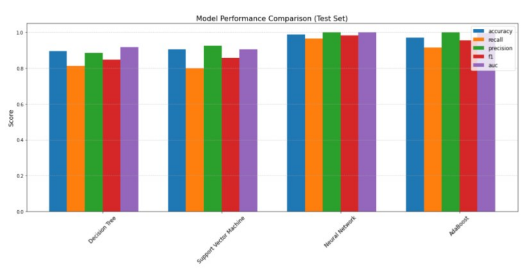

Figure 7 Model performance comparison (test set)

Further combining the model performance comparison results of Table 1 and Figure 7, the performance indicators of the four models are similar, but the neural network model (MLP) performs better overall. Therefore, this study selected it as the core model of esophageal cancer recognition, laying a solid foundation for subsequent research and clinical application, helping early and accurate diagnosis of esophageal cancer and promoting the development of medical diagnosis technology.

5 Conclusion

This study combines the advantages of RF and MLP, uses random forest to screen key features, inputs MLP model for deep training, and constructs RF-MLP esophageal cancer recognition prediction model. When the model is trained and verified, the data set is divided by stratified sampling, trained by MLP neural network, and the model mechanism is analyzed by weight heat map and decision boundary map. The results show that the model fits well with complex esophageal cancer data and can effectively capture data patterns. The prediction accuracy of the model is 98.746 % and the accuracy is 100 %. Compared with the traditionalalgorithm, the prediction ability of the model is better, and it has more accurate prediction ability in esophageal cancer recognition.

About the Author

These authors contributed equally to this work and should be considered co-first authors.
* Corresponding Author: Zhixiang Xu, 17709962087@163.com

References

[1] Li Housheng. Research on survival risk prediction model of esophageal cancer based on machine learning and membrane computing [D]. Zhengzhou University of Light Industry, 2024.

[2] Wu Yuefeng, Wang Qi, Wu Ming. Current status and prospects of machine learning technology in the field of esophageal cancer research [J]. Chinese Journal of Thoracic and Cardiovascular Surgery, 2022.

[3] Zhuang Shuai Shuai, Hua Jie, He Xiaopu, etc. Application of deep learning methods in classification and recognition of esophageal diseases [J]. Journal of Gastroenterology and Hepatology, 2023.

[4] Xu Jinye, Zhou Jianghui, Liu Shengwei, etc. Application of machine learning model in postoperative survival risk stratification of thoracic esophageal squamous cell carcinoma [J].Chinese Journal of Thoracic and Cardiovascular Surgery, 2022.

[5] Li Jiayi, Liu Ke, Shi Linlin, etc. Application progress of intelligent medicine in early diagnosis and treatment of esophageal cancer [J]. Esophageal diseases, 2022.

[6] Jiang Lianghui, Zhou Changhong. Research progress on the application of deep learning technology in gastrointestinal endoscopy diagnosis [J]. Journal of Qingdao University (Medical Edition), 2020.

[7] Deng Genqiang, Lv Zhouping, Zhu Miaomiao, etc. Design of esophageal cancer image intelligent diagnosis system based on convolutional neural network [J].China Digital Medicine, 2018.

[8] Wang Yingcong, Yan Jun, Sun Junwei, et al. Artificial bee colony algorithm based on the division of exploration and exploitation and its application in the prediction of esophageal cancer [J/OL].Engineering science and technology, 1-13 [2024-12-29].

[9] Li Baohua, Xiao Xun, Zhang Min, etc. Study on risk factors and predictive models of esophageal cancer [J].Practical Hospital Clinical Journal, 2024.

[10] Sheng Y ,Zhang L ,Hu Z , et al. Prediction of Early Mortality in Esophageal Cancer Patients with Liver Metastasis Using Machine Learning Approaches[J].Life,2024.

[11] Thavanesan N , Parfitt C , Bodala I , et al. Machine learning models for curative and palliative oesophageal cancer treatment pathway prediction[J].European Journal of Surgical Oncology,2024.

[12] Navamayooran T ,Ganesh V ,Timothy U .A Machine Learning approach to identifying key decision variables within the Oesophageal Cancer MDT[J].European Journal of Surgical Oncology,2023.

[13] Macomber M ,Samareh A , Chaovalitwongse W ,et al. Prediction of Pathologic Complete Response to Neoadjuvant Chemoradiation in the Treatment of Esophageal Cancer Using Machine Learning[J].International Journal of Radiation Oncology, Biology, Physics,2016.

[14] Beld D V J J ,Crull D ,Mikhal J , et al. Complication Prediction after Esophagectomy with Machine Learning[J].Diagnostics,2024.

[15] Xue-Feng L , Qi-Feng W , Jin-Yi L , et al. 177 ct radiomics based on machine learning predicting pathologic complete response after neoadjuvant chemoradiotherapy for esophageal cancer [J].Diseases of the Esophagus,2021.