Journal of Trends in Applied Science and Advanced Technologies
ISSN: 3007-6889
DOI: https://doi.org/10.61784/asat3011

THE PREDICTION AND INFLUENCING FACTORS OF BREAST CANCER RECURRENCE BASED ON RANDOM FOREST

Xi Yang*, WenBei Zheng, WenYun Xia
Guangxi Normal University, Guilin 541000, Guangxi, China.
Corresponding Author: Xi Yang, Email: 249971674@qq.com

Abstract: Breast cancer (BC) is one of the most common malignant tumors in women. In 2022, it has become the second most common cancer after lung cancer. Although medical technology has made great progress in recent years and the survival rate of breast cancer patients has been greatly improved, according to research, about 40% of patients still relapse after treatment. Constructing a breast cancer recurrence prediction model and finding factors that affect breast cancer recurrence are of great significance for clinical treatment and prolonging patient survival. This study used the TCGA dataset and randomly divided the patients into training and test sets in a ratio of 8:2. Seven algorithms, including decision tree, logistic regression, support vector machine, K nearest neighbor, random forest, neural network, and adaptive boosting, were used to construct the model, and the performance of each model was evaluated. The results showed that the random forest model had the best effect, with an accuracy of 97.77%, a sensitivity of 94.23%, a specificity of 99.21%, a false positive rate of 0.79%, an F1 of 96.08%, and an AUC value of 96.72%. The features obtained by the model classification were ranked according to their importance. The top three features were: Age_at_Initial_Pathologic_Diagnosis_nature2012, lymph_node_examined_count and number_of_lymphnodes_Postive_by_he. The model provides more robust feature importance analysis results, providing an important reference for clinicians in breast cancer recurrence risk assessment and individualized treatment decision-making.

Keywords: Breast cancer; Random forest; Recurrence prediction; Influencing factor

1 INTRODUCTION

According to a report released by IARC (International Agency for Research on Cancer), approximately one in five people worldwide will develop cancer in their lifetime, and cancer prevention has become one of the most significant public health challenges of the 21st century. Breast cancer is one of the most common malignant tumors in women, in 2022, it has become the second most common cancer after lung cancer. Every year, more than 300,000 women in China are diagnosed with breast cancer, and the age of onset is becoming younger and younger[1]. In 2020, the number of new cases of breast cancer in China exceeded 420,000, and the number of deaths exceeded 100,000, ranking first in the world[2]. Breast cancer has a high mortality rate. Early diagnosis of cancer and active treatment are the most effective ways to reduce deaths from malignant tumors. However, most patients' deaths are not caused by the primary tumor, but by tumor recurrence or metastasis. Some patients are at risk of recurrence after initial treatment. Recurrence will increase the original patient's clinical manifestations and greatly increase the difficulty of treatment. The highest risk of recurrence is within 5 years after treatment. Although medical technology has made great progress in recent years and the survival rate of breast cancer patients has been greatly improved, according to research, about 40% of patients still experience recurrence after treatment[3]. The occurrence and development of tumors involve many factors[4,5], and breast cancer recurrence may be related to biological characteristics such as tumor size, lymph node metastasis, degree of differentiation, estrogen receptors and human epidermal growth factor receptors. In addition, factors such as patient age, duration of disease, and failure to completely eliminate tumor cells after initial treatment may also affect the risk of recurrence.

In recent years, with the rapid development of information technology, a series of technologies such as big data and artificial intelligence have become hot topics in all walks of life. With the support of big data and the continuous integration of computer science and technology and other disciplines, the new generation of information technology has become an important driving force for social progress. Machine learning (ML), as one of the core technologies of artificial intelligence, has achieved certain results in cloud computing, biomedicine and other fields. How to combine machine learning with big data to generate value has received more and more attention from the society and has become a hot topic in the field of "big data + artificial intelligence"[6]. In this complex context, machine learning technology plays an important role in predicting breast cancer recurrence. Machine learning methods can more accurately capture complex patterns and potential relationships in data, thereby improving the accuracy of predicting breast cancer recurrence. Therefore, how to use machine learning methods to evaluate and improve the prognosis of breast cancer and find simple, efficient and easy-to-observe influencing factors to predict the risk of postoperative recurrence in breast cancer patients is of great significance for clinical treatment and prolonging patient survival.

2 THEORETICAL OVERVIEW

Random Forest (RF) is an ensemble learning algorithm based on decision trees. Ensemble learning refers to learning multiple estimators through training. When prediction is required, the results of multiple estimators are integrated as the final output through a combiner, thereby improving the versatility and robustness of a single estimator.

Random Forest is actually a bagging algorithm with decision trees as estimators. It uses random sampling, random feature selection and other techniques to construct multiple decision trees and combine these decision trees for classification or prediction. Random Forest constructs each decision tree by randomly selecting features and samples, and integrates the prediction structure of multiple trees through a voting mechanism, thereby reducing the risk of overfitting[7]. This ensemble learning method can usually provide higher accuracy than a single model[8] and significantly improve prediction accuracy.

Specifically, Random Forest learns and trains multiple decision trees through self-service sampling technology and makes aggregate predictions. For regression problems, k samples are randomly selected with replacement from the original training set, and k samples are trained separately to generate k decision tree models. Finally, the results of the k decision trees are combined according to the simple averaging method to form the result[9].

3 DATA SOURCE

The data used in this study came from the BRCA in the TCGA dataset. The inclusion criteria for the data in this study were: (1) pathologically confirmed recurrence of breast cancer after the initial diagnosis; (2) primary tumor site: breast; (3) age > 18 years old. A total of 1,247 patient data were collected. The data mainly consisted of two parts: clinical data and survival data. The clinical data included basic patient information (age, gender, menopausal status, etc.), tumor characteristics (size, pathological classification, stage, etc.), molecular biological test results (ER, PR, etc.), treatment regimen, and other variables (new tumor site, distant metastasis, etc.). Survival data mainly included overall survival (OS), disease-specific survival (DSS), disease-free survival (DFI), and progression-free survival (PFI).

4 DATA PROCESSING

The data set initially contains 182 variables, among which the key target variable is new_tumor_event_after_in itial_treatment, which means the occurrence of new tumors after initial treatment. Before model building and a nalysis, the data are processed as follows:

First, some variables are deleted. The records of the target variable are not empty in 1007 samples, indicating that this part of the data has complete records of subsequent tumor events. For the 240 samples where the target variable is empty, they are removed from the data set to ensure the completeness and accuracy of the analysis. Some variables without records and variables with more missing values are also removed. After this step, the number of variables in the data set is reduced to 77. Further review found that there are still a large number of variables that only record a single data point, that is, these variables are not recorded in most samples or the variable values are the same. These variables have no analytical value and are also deleted from the data set. Finally, combined with literature research, 17 variables that are highly correlated with breast cancer recurrence and have data records are preliminarily screened out, including 1 target variable, and 16 characteristic variables containing basic patient information, tumor characteristics and treatment related information.

Secondly, encode the variables appropriately. In the construction of the prediction model for breast cancer recurrence, since the model cannot directly interpret and process text data, it is necessary to have an appropriate encoding strategy for all non-numerical variables so that they can be effectively recognized and used by the model. The encoding process involves not only conventional digital conversion, but also the clinical significance expressed by the variables, especially when representing levels such as disease severity or treatment response. For these variables, natural numbers are used to encode in sequence, referred to as "sequential coding". This method divides the levels according to the health risk or deterioration reflected by the variables; binary variables are encoded with 0 and 1; for continuous variables, standardized processing is used.

Then, adjust the proportion of the target variable and interpolate the missing values. When statistically analyzing the target variable, the target variable data shows that the ratio of "yes" and "no" is 108:899. This distribution is extremely unbalanced and directly affects the learning effect of the model. In order to improve the fitting ability of the model, the records with a target variable of "no" and a large number of missing values in the sample are deleted, and the ratio of the target variable "yes" and "no" is adjusted to 108:639. The CART method was used to predict the small number of missing values that still existed in the sample. The oversampling technique was further used to increase the sample size to 900, and the ratio of "yes" in the target variable was appropriately increased. Finally, the ratio of yes and no in the target variable was adjusted to 261:639. This not only helps to avoid the bias of the model to the majority class, but also improves the recognition ability of the minority class, thereby enhancing the prediction accuracy and reliability of the model in practical applications.

Finally, for the study of factors affecting breast cancer recurrence, statistical tests and RFECV methods were used for feature selection. First, the Chi-squared Test was performed on the categorical variables to evaluate the independence between each variable and the target variable; at the same time, for the continuous variables, the T test was used to evaluate whether the mean difference between it and the target variable was statistically significant, so as to confirm the predictive value of the continuous variable. Through the Chi-square test and T test, the categorical variables and continuous variables related to breast cancer recurrence can be effectively screened out, which also provides reliable statistical support for the research results and enhances the persuasiveness of the research conclusions. Finally, 14 variables were selected from the initial variables in conjunction with the RFECV method, as shown in Table 1.

Table 1 Variable Information

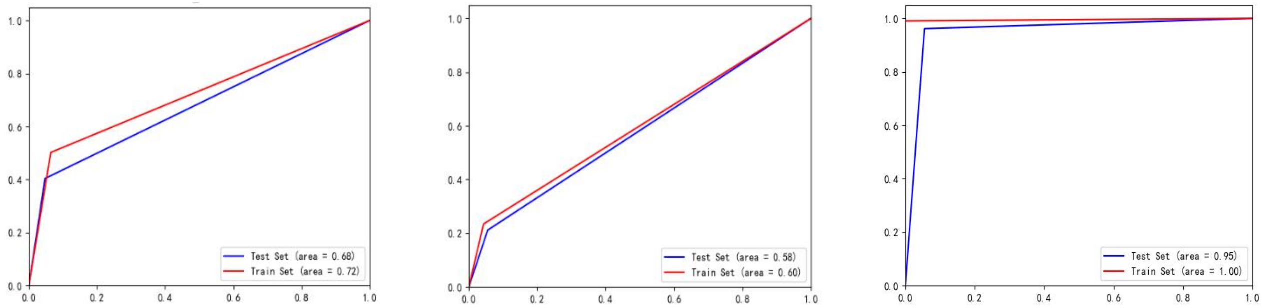
category	variable name
target variable	new_tumor_event_after_initial_treatment
patient basic information	Age_at_Initial_Pathologic_Diagnosis_nature2012 menopause_status
tumor characteristics	PAM50Call_RNAseq histological_type pathologic_T pathologic_N pathologic_M
molecular biology test results	breast_carcinoma_estrogen_receptor_status breast_carcinoma_progesterone_receptor_status lab_proc_her2_neu_immunohistochemistry_receptor_status
treatment related information	breast_carcinoma_surgical_procedure_name radiation_therapy
others	lymph_node_examined_count number_of_lymphnodes_positive_by_he

5 MODEL ANALYSIS

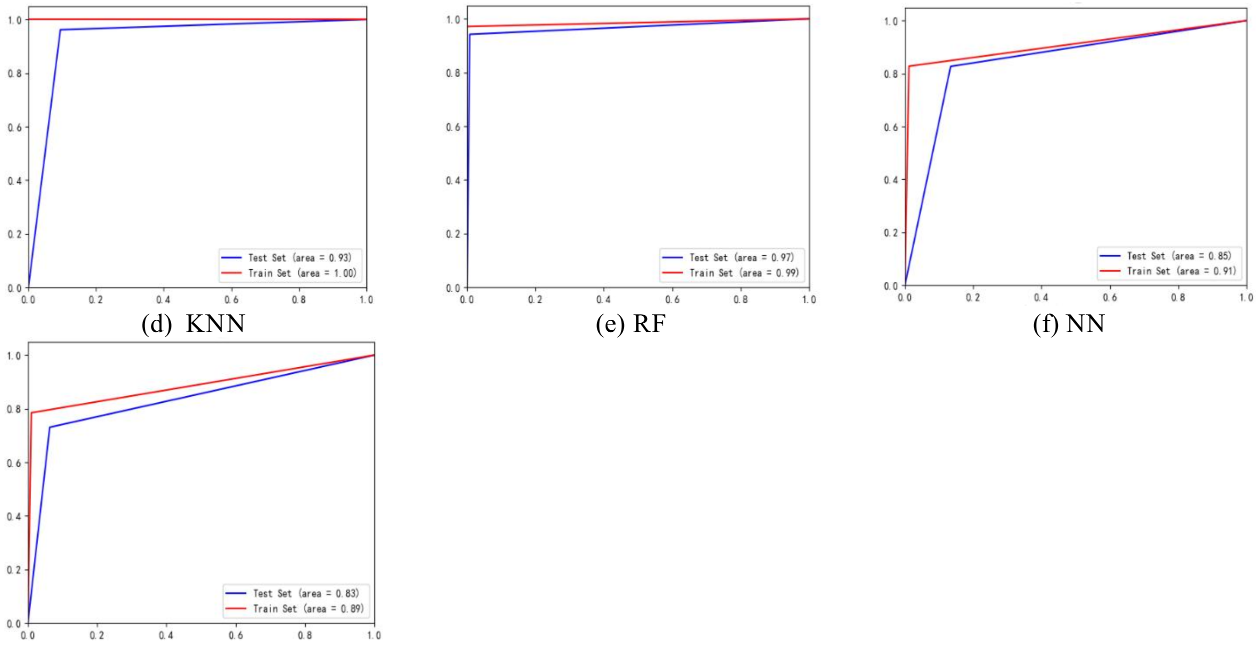
The samples were divided into training set and test set in the ratio of 8:2, and seven methods such as decision tree, support vector machine, and random forest were used to construct the model to find the best model in the study of breast cancer recurrence prediction, and the accuracy, sensitivity, specificity, false-positive rate, F1, and AUC indicators of each model were combined, and it can be found in Table 2 and Figure 1 that the random forest performs the best among the seven models, and the AUC value of the model is 0.9672, so the random forest model is the best model for classification and prediction of breast cancer.

Table 2 Classification Performance Evaluation Indicators of Each Model

Model	Accuracy	Sensitivity	Specificity	False_Positive_Rate	F1	AUC
Decision Tree	0.7933	0.4038	0.9528	0.0472	0.5316	0.6783
LR	0.7318	0.2115	0.9449	0.0551	0.3143	0.5782
SVM	0.9497	0.9615	0.9449	0.0551	0.9174	0.9532
KNN	0.9218	0.9615	0.9055	0.0945	0.8772	0.9335
RF	0.9777	0.9423	0.9921	0.0079	0.9608	0.9672
NN	0.8547	0.8269	0.8661	0.1339	0.7679	0.8465
Adaboost	0.8771	0.7308	0.9370	0.0630	0.7755	0.8339



(a) Decision Tree (b) LR (c) SVM



(d) KNN (e) RF (f) NN

(g) Adaboost

Figure 1 ROC Curves of Each Model

The confusion matrix of the random forest model on the training and test sets is shown in Figure 2, from which the model's classification effectiveness and misclassification can be visualized. In the training set, the random forest model correctly classified 512 negative observations and 203 positive observations, and produced 0 false positives and 6 false negatives; in the test set, it correctly classified 126 negative observations and 49 positive observations, and produced 1 false positive and 3 false negatives. Overall, the model performed well in classification accuracy and stability, which provides strong support for its application in practical operations.

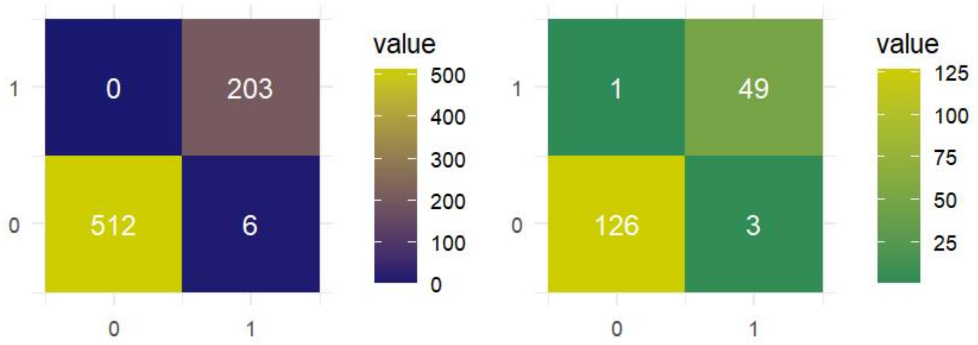


Figure 2 Confusion Matrix

The ranking of the importance of features that affect breast cancer recurrence based on the random forest classifier is shown in the Figure 3. It can be seen that the top three feature indicators are Age_at_Initial_Pathologic_Diagnosis _nature2012, lymph_node_examined_count and number_of_lymphnodes_postive _by_he. Except for the top three, the other features have a small gap in feature ranking.

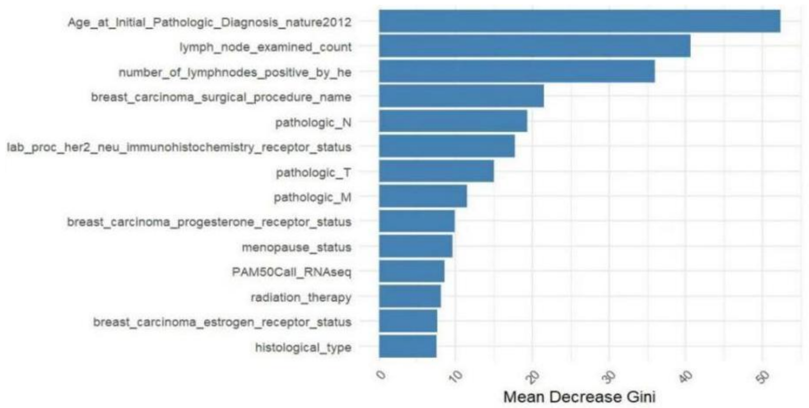


Figure 3 The Optimal Number of Features for the Model

6 CONCLUSION

With the rapid development of artificial intelligence (AI) in clinical cancer research and application, cancer prediction performance has reached a new height. In particular, machine learning and deep learning technologies use a large amount of rich medical data to diagnose cancer, predict patient prognosis and provide treatment methods. This study used the random forest model to predict breast cancer recurrence and its influencing factors, and the results showed good stability and accuracy. For the study of influencing factors of breast cancer recurrence, the features classified by the model were sorted according to their importance, and it was found that Age_at_Initial_Pathologic_Diagnosis_nature2012, lymph_node_examined_count and number_of_lymphnodes_postive_by_he were important factors affecting breast cancer recurrence. It is recommended that doctors focus on these factors during treatment and follow-up, and provide more accurate treatment and follow-up for patients with older age at initial pathological diagnosis, more lymph nodes and more positive lymph nodes, so as to reduce the recurrence rate of breast cancer. At the same time, these factors can also be used as one of the indicators for breast cancer prevention and screening, helping to detect and diagnose breast cancer early and improve treatment effect and survival rate. In subsequent studies, we should consider including patients from different regions and different medical backgrounds, increase the sample size through multi-center cooperation, and improve the reliability of the results. We should also optimize the follow-up strategy, regularly and continuously track the health status of patients, ensure the integrity and reliability of follow-up data, further optimize the model, provide a scientific basis for the realization of precision medicine, and assist clinicians in making more accurate decisions in formulating personalized diagnosis and treatment plans.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Huang Xiaolin, He Ningning, Chen Shuzhen, et al. Retrospective study on the changes of PRL levels in female breast cancer patients aged 30 to 50 years old from Zhanjiang. Smart Health, 2018, 4(22).

[2] Rainey Linda, Eriksson Mikael, Trinh Thang, et al. The impact of alcohol consumption and physical activity on breast cancer: The role of breast cancer risk. International journal of cancer, 2020.

[3] Whitaker K D, Sheth D, Olopade O I. Dynamic contrast enhanced magnetic resonance imaging for risk-stratified screening in women with BRCA mutations or high familial risk for breast cancer: are we there yet? Breast Cancer Res Treat, 2020, 183(2).

[4] Lepucki A, Orlińska K, Mielczarek-Palacz A, et al. The Role of Extracellular Matrix Proteins in Breast Cancer. Journal of Clinical Medicine, 2022, 11(5):1250

[5] Heitmeir B, Deniz M, Janni W, et al. Circulating Tumor Cells in Breast Cancer Patients:A Balancing Act between Stemness, EMT Features and DNA Damage Responses. Cancers(Basel), 2022, 14(4): 997

[6] Bilski J, Smolag J. Parallel architectures for learning the RTRN and Elman dynamic neural networks. IEEETransactions on Parallel and Distributed Systems, 2015, 26(9): 2561.

[7] Wu Ying. Research on urban waterlogging risk assessment and prediction based on machine learning. Beijing University of Civil Engineering and Architecture, 2024. DOI:10.26943/d.cnki.gbjzc.2024.000067.

[8] Xun L, Peng Z, Yichen L, et al. Influencing Factors and Risk Assessment of Precipitation-Induced Flooding in Zhengzhou, China, Based on Random Forest and XGBoost Algorithms. International Journal of Environmental Research and PublicHealth, 2022, 19(24): 16544.

[9] Liu Chao. Regression Analysis: Methods, Data and Application of R. Beijing: Higher Education Press, 2019.