MACHINE LEARNING FOR DISEASE TREATMENT RESPONSE PREDICTION

Minlong Chen, Ruoyu Wen, Junjie Xia, Zhuangzhou Feng

University of Nottingham

ABSTRACT

Breast cancer is one of the cancers that belongs to the common female cancers. In this regard, the treatments for this cancer have also attracted great attention. Among them, chemotherapy is one of the common methods of cancer treatment, but it may affect the patients themselves and may not be able to produce therapeutic effects for all patients. In response, machine learning algorithms are used to predict pathological complete response (PCR) and recurrence-free survival (RFS), whereupon patients can be stratified and treated according to their condition. To solve this problem, we used Principal Component Analysis (PCA) for feature selection and then used K-fold cross validation for model evaluation using logistic regression algorithms, decision trees, support vector machines, and neural networks.

1. INTRODUCTION

Breast cancer is the most common form of cancer among women in the UK and a major health challenge globally. In 2023, approximately 297,790 new cases of invasive breast cancer are expected to be diagnosed in U.S. women. Additionally, approximately 55,720 new cases of ductal carcinoma in situ (DCIS) are expected to be diagnosed. Unfortunately, approximately 43,700 women are expected to die from breast cancer that same year (American Cancer Society, 2023). Despite advances in medical science, the treatment and management of breast cancer remains complex and multifaceted. One of the main treatment strategies is chemotherapy, especially for locally advanced tumors. The aim of this approach is to reduce the size of the tumor before surgical intervention. The effectiveness of chemotherapy varies from patient to patient, and its toxicity often poses additional health risks. In recent years, the incidence of breast cancer has increased slightly, increasing by approximately 0.5% per year. Still, overall breast cancer death rates have been declining steadily since 1989, with an overall decrease of 43% in 2020. This decline has been attributed to earlier detection through screening, increased awareness, and improved treatment (American Cancer Society, 2023).

A key indicator of successful breast cancer treatment is the achievement of pathological complete response (PCR), which is the complete regression of the tumor at the time of surgery. pCR is frequently used as an endpoint in clinical trials to assess the efficacy of neoadjuvant therapy. The higher the pCR rate in a trial, the more effective the treatment (Fayanju et al., 2018). Achieving PCR is closely associated with a higher likelihood of cure and longer recurrence-free survival (RFS) time. A large observational study found that pCR was associated with improved survival in patients with early-stage HER2+ breast cancer, with a 3-year overall survival rate of 97.7% in patients with pCR and 94.4% in patients with residual disease (An et al ., 2022). RFS is the duration after initial treatment that a patient shows no signs or symptoms of cancer. Unfortunately, current statistics indicate that only approximately 25% of patients receiving chemotherapy achieve PCR, while the majority of patients still have residual disease with variable prognosis (Huang et al., 2022). In this context, there is an urgent need to improve patient stratification and treatment planning. This need emphasizes the importance of utilizing available clinical and diagnostic information to predict PCR and RFS outcomes before chemotherapy is initiated.

The goal of this project is to utilize sophisticated machine learning algorithms to predict pathological complete response (PCR) and recurrence-free survival (RFS) related outcomes in breast cancer patients. These predictions will utilize characteristics of clinical indicators and characteristics of pre-chemotherapy magnetic resonance imaging (MRI). The overall goal is to create a predictive model that helps breast cancer patients better plan and personalize treatment strategies, potentially improving their treatment outcomes and quality of life. This tailored approach has the potential to reduce unnecessary chemotherapy toxicity in patients who are unlikely to benefit from it and provide more targeted treatment to those who are more likely to respond positively.

2. DATA PROCESSING

The data used in this paper is a simplified dataset generated by The American College of Radiology Imaging Network (I-SPY 2 TRIAL), which consists of 10 clinical features and 107 MRI-based features that were extracted from the tumor region of the MRI. of the tumor region. First, the data preprocessing includes normalization, feature extraction, dimensionality reduction and other operations. These steps allow the data to be better categorized and to find the most informative feature set, which is the one we need. This is because the performance of the classifier operation can be improved (Subasi, 2020).

2.1. CheckOutliers

For this, in the first step we first remove the rows that have missing values in PCR (outcome) and RelapseFreeSurvival (outcome), and then we use the median of each column for the other columns to fill in the missing values. In the second step we remove the outliers. This is because removing outliers from the data reduces the overall variability of the data and at the same time increases the statistical power and improves the model training process (Pedregosa et al., 2011). In this regard, we use the theory of box-and-line plot to identify outliers in columns 13 and beyond, specifically, the first quartile, third quartile, and quartiles are calculated for each column, then upper and lower bounds are defined, specifically less than $Q1 - 1.5 * IQR$ or more than $Q3 + 1.5 * IQR$ are considered as outliers, and then values exceeding the bounds are set as NaN. KNNImputer to interpolate the values as NaN. Finally these data are united with the previous 13 columns.

2.2. SMOTE-- imbalence using oversampling

Viewing the result set of PCR on the entire data reveals that the data is highly unbalanced. SMOTE is a method used to solve the problem of unbalanced classification, which has been considered as one of the most popular methods since it was proposed (Chawla et al., 2002). It has now become the benchmark method used to solve the classification imbalance problem (Sun et al., 2023). We extract the feature matrix from DataFrame excluding the columns PCR (outcome),ID,RelapseFreeSurvival(outcome).PCR (outcome) was used as the target variable. Oversampling the feature and target vectors using SMOTE yields the generated samples due to the addition of a small amount of sample data from a small number of categories to balance it with a large number of samples.

2.3. Normalization

Finally the data is normalized. This is where standardizing the data allows the data to be distributed within the same range, which has the advantage of making the gap between each feature quantity smaller (He et al., 2010). Different data will usually have different dimensional units, which can affect the assessment results, as different dimensions can lead to too large a gap in the data. And errors due to differences in data indicators need to be addressed before using PCA (Sun et al., 2016). We decided to use the method MinMaxScaler

$$X_{std} = \frac{X - X.\min}{X.\max - X.\min}$$

, which traverses the data in each column in turn to see if it is between [0,1], and if it is not in this range, the normalization operation is performed. This facilitates the adaptation of machine learning algorithms.

2.4 OverSampling

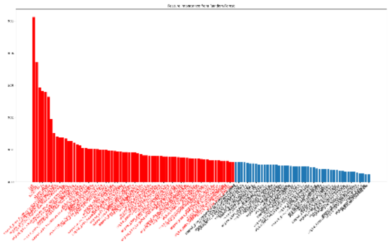
Due to the imbalanced class problem in classification task, We employ SMOTE technology which is an oversampling method that generates synthetic samples for the minority class to address class imbalance in machine learning datasets.

3. FEATURE ENGINEERING

3.1. Attribute Selection Based on Random Forest

Random Forest(RF) is a ensemble machine learning algorithm, which can select features according to the importance of feature and reduce overfitted of the model. (Saraswat and Arya 2014)Wade(Wade, Joshi et al. 2017) build the regularized RF to select the feature from HDSD data from subcortical brain surface.

There are 117 features in the disease treatment response prediction. Different features contribute different role to the train of the model. Less important feature will affect the prediction's accuracy, lead to curse of dimension due to the limited data samples and increase the complexity of the model.



This figure show the feature according to the order of importance from the biggest to smallest. For Classification task(PCR),we choose top 70 features. For Regression task, we choose top 60 features.

3.2. PCA for Feature Extraction

Pca can mapping the original feature space to the low-dimensional feature space by calculating the relationship between features, this achieve the goal of dimension reduction (Wang and Paliwal 2003).

PCA is an dimensionality reduction method which is quite frequently used in unsupervised learning. After calculating and sorting the correlation between multidimensional data groups, strongly correlated features will be merged to generate a new feature. On the one hand, PCA satisfied with the condition of minimizing the information loss, on the other hand, it is able to make the data set easier to use, simplify the data structure, completely without parameter limitation, and reduce the calculation cost of the algorithm(Hess and Hess 2018).

For classification task(PCR),we choose to reduce dimension to 40.For Regression task(RFS),we choose to reduce dimension to 16.

4. MODEL

4.1. Linear regression:

Linear regression is a commonly used method in machine learning and is a linear model used to establish a linear relationship between the independent and dependent variables. The expression is:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \beta_3 \mathrm{X}_3 + \ldots . + \beta_n \mathrm{X}_n + \varepsilon$$

In this expression, Y is the dependent variable. X_1 , X_2 ... ,X_n are the independent variables and ε is the error term.

4.2. Logistic regression

Logistic regression is a statistical learning model that is used to represent a classification model that is used to deal with the problem of classification, although the name with the word "regression" is used in this case for binary classification problems.

First of all its expression is:

$$\mathrm{P(x)} = \frac{1}{1 + \mathrm{e}^{-\mathrm{z}}}$$

P(x) is the probability that the event will occur, so the probability that the event will not occur is 1 - P(x).

Michael extensively examines the utility of a logistic regression model based on blood-based gene expression for predicting obstructive coronary artery disease. (LaValley 2008)

4.3. Neural network

A neural network model is a machine learning model created by modeling the structure and function of biological nerves. The model consists of neurons, layers, activation functions and weights. (Krose 1996)

First of all, the neurons are the most important part of the neural network because each neuron has a significant role to play, for example: accepting the input data and then generating the output data through the activation function contained in the neurons.

There are three main neural network layers, which are: input layer, hidden layer and output layer.

The weights(w) are on the edges between the neurons and it is automatically updated during the training process.

The role of the activation function is to convert the linear input into a non-linear output.

Feedforward propagation formula:

$$a_j = \sum_{i=1}^{n} w_{ij} x_i + b_j$$

$$z_j = f(a_j)$$

In the first formula a_j denotes the weighted input of the neuron, w_{ij} denotes the weights between the neurons, xi denotes the input of the neuron and b_j denotes the bias of the neuron. In the second formula z_j is the output of the neuron, $f()$ is the activation function.

4.4. SVM

The svm is a machine-supervised learning algorithm for classification and regression. The idea of this algorithm is to find the best hyperplane to separate samples of different classes. Specifically, in two dimensions, a hyperplane is a line, while in three dimensions it presents a plane. And the support vector is the point that is the closest to the hyperplane. The interval is the distance between this point of the support vector and the hyperplane. This model allows us to work with more high dimensional data. The kernel function we use in this experiment is the Gaussian kernel function (RBF).

4.5. Decision Tree

Decision Tree is a non-parameter supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Classification and regression trees are machine learning techniques widely used in our world. Classification trees suit categorical variables, measuring errors in misclassification costs. Regression trees fit continuous variables, assessing errors through squared differences.(Loh 2011)

5. OPTIMIZE MODEL WITH GRIDSEARCH

GridSearch is a hyper parameter optimization technique in machine learning that systematically explores a specified set of hyperparameters for a given model.

It exhaustively evaluates all parameter combinations by performing cross-validation, optimizing model performance based on a scoring metric. By searching through this grid of parameters, it aims to find the optimal configuration that yields the best model performance. (Jiménez, Lázaro et al. 2008)

Sun conducted a comparative analysis of machine learning algorithms for classification tasks, evaluating their performance with improved grid search algorithm to optimize SVR. (Huang, Mao et al. 2012)

6. METHOD EVALUATION

6.1. MAE

Mean Absolute Error (MAE) is a performance evaluation metric often used in regression tasks. The formula is：

$$\mathrm{MAE} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{y}_\mathrm{i} - \hat{\mathrm{y}}_\mathrm{i}|$$

first, n denotes the number of samples, y_i denotes the actual observation of the ith sample, and $\hat{y}_i$ is the predicted value of the model for the ith sample. To begin with, the MAE is relatively simple to compute by summing all the errors (absolute values) and dividing the total error by the number of samples. From this we can conclude that MAE is used to measure the average absolute difference between the actual observations and the model predictions. So we use MAE here as an evaluation metric for machine learning. Finally if the value of MAE is larger, it means that the error of the model is larger. the smaller the MAE is, the smaller the error is, which means that the accuracy of the model is more accurate.

6.2. Balanced-accuracy

Balanced-accuracy is a performance evaluation metric in categorization tasks with the formula:

$$\text{Accuracy} = \frac{\text{Specificity} + \text{Sensitivity}}{2}$$

that is divided into two parts specificity and sensitivity.

First sensitivity is used to determine reliability (Su et al., 2023, Zhang et al., 2023) with the formula:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The formula for specificity is:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

This metric takes into account the prediction accuracy for each category, and the accuracy calculation is expressed as the percentage of accurate identifications For the instances that are identified accurately, the prediction accuracy is calculated by dividing the number of all correct predictions by the number of all instances in the entire dataset (Das et al., 2023), and then the prediction accuracy is weighted and averaged, which ensures that on an imbalanced dataset one can obtain a more balanced and fairer assessment.

For this we performed Balanced-accuracy calculations on the test results and obtained a final result of 0.823838

7. DESCRIPTION OF EXPERIMENTAL RESULTS

In the data preprocessing stage, we replace the null value marked '999' in the label with np.nan. For pCR and RFS columns, as soon as a missing value (NaN) is detected, the relevant row is discarded. If there are missing values in other columns, they are filled with the median of that column.

Next, use the KNN algorithm to handle outliers, where the K value is set to 3, which means using the three nearest neighbor values to estimate missing values. We also applied box plot theory to identify outliers in the data.

In the training data set, we observed that the number of non-cured and cured samples in the pCR label was 311 and 84 respectively, indicating an obvious class imbalance. In order to solve this problem, we used SMOTE technology for oversampling to balance the number of samples in the two categories.

To ensure that certain columns in the data, such as age and high-dimensional data after column 11, are normalized, we map all values to between 0 and 1. After this series of data processing, the positive and negative ratio of data used for model training is reasonable.

In the feature selection stage, we used the random forest algorithm to train the model and output it according to the feature importance. Only the top k features were retained for experiments. In the end, the pCR classification model retained 70 features, while the RFS regression model retained 60 features.

For PCA dimensionality reduction, the pCR classification model retains 40 principal components, while the RFS regression model retains 16 principal components. This can reduce the random fluctuations of the data and retain the most important information.

In five-fold cross-validation, different algorithms will produce different results, as shown in the following figure:

CrossValMeans	CrossValStd	Model
0.720081	0.030770	Logistic Regression
0.706867	0.034009	Decision Tree
0.793785	0.070789	SVM
0.799517	0.093510	Neural Network

From the above figure, it can be concluded that the Neural Network (MLPClassifier) and support vector machine (SVM) have higher average scores than logistic regression and decision trees, although the standard errors are larger, which indicates that they have higher performance potential. Therefore, in pCR classification projects, we tend to choose the MLP model and further improve its accuracy through tuning. The final result of GridSearchCV tuning the neural network is 0.8125574.

In the five-fold cross-validation of the RFS regression model, the following results were obtained:

CrossValMeans	CrossValStd	Model
23.735836	10.178033	Linear Regression
32.065190	5.657444	Decision Tree
22.795589	10.799633	SVM
32.955459	15.602255	Neural Network

According to the data in the above figure, for the RFS regression model, SVM has the fastest performance and better effect, so it is selected for GridSearchCV tuning. The adjustment result is 22.84452859865143.

In summary, through careful data preprocessing and reasonable model selection and tuning, we have achieved satisfactory results in both classification and regression tasks. In future work, we will continue to monitor the performance of the model and further adjust model parameters as needed.

8. REFERENCES

[1]American Cancer Society, 2023. How common is breast cancer? [online] Available at: https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html [Accessed 14 December 2023].

[2]An, S.J., Duchesneau, E.D., Strassle, P.D., Reeder-Hayes, K., Gallagher, K.K., Ollila, D.W., Downs-Canner, S.M. and Spanheimer, P.M. (2022). Pathologic complete response and survival after neoadjuvant chemotherapy in cT1-T2/N0 HER2+ breast cancer. npj Breast Cancer, 8(1). doi:https://doi.org/10.1038/s41523-022-00433-x.

[3]Chawla, N.V. et al. (2002) ‘SMOTE: Synthetic Minority Over-sampling Technique’, The Journal of artificial intelligence research, 16, pp. 321 – 357. doi:10.1613/jair.953.

[4]Fayanju, O.M., Ren, Y., Thomas, S.M., Greenup, R.A., Plichta, J.K., Rosenberger, L.H., Tamirisa, N., Force, J., Boughey, J.C., Hyslop, T. and Hwang, E.S. (2018). The Clinical Significance of Breast-only and Node-only Pathologic Complete Response (pCR) after Neoadjuvant Chemotherapy (NACT): A Review of 20,000 Breast Cancer Patients in the National Cancer Data Base (NCDB). Annals of surgery, [online] 268(4), pp.591–601. doi:https://doi.org/10.1097/SLA.0000000000002953.

[5]He, M. et al. (2010) ‘Performance evaluation of score level fusion in multimodal biometric systems’, Pattern recognition, 43(5), pp. 1789 – 1800. doi:10.1016/j.patcog.2009.11.018.

[6]Huang M, O'Shaughnessy J, Zhao J, Haiderali A, Cortés J, Ramsey SD, Briggs A, Hu P, Karantza V, Aktan G, Qi CZ, Gu C, Xie J, Yuan M, Cook J, Untch M, Schmid P, Fasching PA. Association of Pathologic Complete Response with Long-Term Survival Outcomes in Triple-Negative Breast Cancer: A Meta-Analysis. Cancer Res. 2020 Dec 15;80(24):5427-5434. doi: 10.1158/0008-5472.CAN-20-1792. Epub 2020 Sep 14. PMID: 32928917.

[7]Pedregosa, F. et al. (2011) ‘Scikit-learn: Machine Learning in Python’, Journal of machine learning research [Preprint]. doi:10.5555/1953048.2078195.

[8]Saraswat, M. and K.V. Arya, Feature selection and classification of leukocytes using random forest. Medical & Biological Engineering & Computing, 2014. 52(12): p. 1041-1052.

[9]Su, W. et al. (2023) ‘A subgroup dominance-based benefit of the doubt method for addressing rank reversals: A case study of the human development index in Europe’, European journal of operational research, 307(3), pp. 1299 – 1317. doi:10.1016/j.ejor.2022.11.030.

[10]Subasi, A. (2020) ‘Chapter 2’, in Practical machine learning for data analysis using Python. London England: Academic Press, pp. 27 – 89.

[11]Sun, W. et al. (2016) ‘A Method for Developing Biomechanical Response Corridors based on Principal Component Analysis’, Journal of biomechanics, 49(14), pp. 3208 – 3215. doi:10.1016/j.jbiomech.2016.07.034.

[12]Wade, B.S.C., et al., Machine learning on high dimensional shape data from subcortical brain surfaces: A comparison of feature selection and classification methods. Pattern Recognition, 2017. 63: p. 731-739.

[13]Zhang, C. et al. (2023) ‘A Combined Weighting Based Large Scale Group Decision Making Framework for MOOC Group Recommendation’, Group decision and negotiation, 32(3), pp. 537 – 567. doi:10.1007/s10726-023-09816-2.

[14]Hess, A. S. and J. R. Hess (2018). "Principal component analysis." Transfusion 58(7): 1580-1582.

[15]Loh, W. Y. (2011). "Classification and regression trees." Wiley interdisciplinary reviews: data mining and knowledge discovery 1(1): 14-23.

[16]Wang, X. and K. K. Paliwal (2003). "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition." Pattern recognition 36(10): 2429-2439.

[17] Jiménez, Á. B., J. L. Lázaro and J. R. Dorronsoro (2008). Finding optimal model parameters by discrete grid search. Innovations in Hybrid Intelligent Systems, Springer: 120-127

[18] Huang, Q., J. Mao and Y. Liu (2012). An improved grid search algorithm of SVR parameters optimization. 2012 IEEE 14th International Conference on Communication Technology, IEEE.

[19] Krose, B. (1996). An introduction to neural networks.

[20] LaValley, M. P. (2008). "Logistic regression." Circulation 117(18): 2395-2399.

Weighing	Data process	Feature Selection	ML method development	Method Evaluation	Report
Feng	25%	35%	30%	25%	35%
Xia	25%	35%	35%	25%	15%
Chen	25%	20%	25%	25%	35%
Wen	25%	10%	10%	25%	15%