USING MACHINE LEARNING TO PREDICT BREAST CANCER TREATMENT OUTCOMES

Juntian Xiao, Yuhong Yuan, Luqi Xin, Qifeng He, Guangzheng Dong

The University of Nottingham

ABSTRACT

This study explores the use of advanced machine learning techniques to predict pathological complete response (PCR) and relapse-free survival (RFS) in breast cancer patients. Used a simplified version of a public dataset based on the American college of radiology imaging network (I-SPY 2 TRIAL), applied feature selection and trained models to address classification and regression tasks. SMOTE, PCA, and hyperparameter tuning technologies are used to solve the problems of high-dimensional data and uneven distribution of data values. This paper provides methods for data preprocessing, feature selection, and model optimization for unbalanced datasets, and evaluates the performance of the model using evaluation methods for related tasks, such as f1-score and Mean Absolute Error (MAE).

Index Terms— Machine Learning, Breast Cancer, Data Preprocessing, SMOTE, Feature Selection, PCA, Random Forest.

1. INTRODUCTION

Breast cancer is the most common cancer in the UK for women. Chemotherapy is a commonly used treatment strategy to reduce the size of locally advanced tumours before surgery. However, chemotherapy is a toxic process to the human body and it is not always effective for everyone. Complete tumour resolution at surgery, known as pathological complete response (PCR), has a high likelihood of achieving a cure and longer relapse-free survival (RFS) time. RFS is the length of time after primary treatment for cancer ends that the patient survives without any signs or symptoms of that cancer. However, only 25% of patients receiving chemotherapy will achieve a PCR, with the remaining 75% having residual disease and a range of prognosis. Better patient stratification and treatment could be achieved if PCR and RFS could be predicted using information prior to chemotherapy treatment.

Predicting PCR and relapse-free survival (RFS) using pre-treatment clinical and MRI data could personalize treatment strategies. This study utilizes machine learning techniques to address the classification of PCR and regression of RFS, overcoming challenges like data imbalance and high-dimensional features.

And through the training of the basic dataset, it prepares for the prediction task of the unknown dataset.

2. LITERATURE REVIEW

Breast cancer remains a leading cause of mortality worldwide. Machine Learning techniques are increasingly utilized for early diagnosis and risk assessment. Among various approaches, Support Vector Machines (SVMs) are frequently highlighted as effective classifiers. Asri et al. (2016) compared SVM with other algorithms such as k-Nearest Neighbors (k-NN), Naïve Bayes (NB), and Decision Trees on breast cancer datasets, emphasizing SVM's consistent performance in delivering reliable predictions (Asri et al., 2016). Similarly, Naji et al. (2021) noted SVM's robust performance across multiple criteria, such as sensitivity and specificity, further establishing its value in applications (Naji et al., 2021).

Deep learning techniques also demonstrate significant potential. Islam et al. (2020) found Artificial Neural Networks (ANNs) to perform favorably compared to traditional ML methods, showcasing their ability to handle complex datasets effectively (Islam et al., 2020). Fatima et al. (2020) reviewed hybrid models, such as ensemble techniques combining SVM and Decision Trees, highlighting their capability to improve diagnostic accuracy and adaptability in diverse healthcare environments (Fatima et al., 2020).

However, data processing remains a critical challenge. Breast cancer datasets often exhibit heterogeneity, including issues such as inconsistent feature dimensions, missing values, and redundant information. Feature selection and data preprocessing are therefore crucial, as these steps directly impact model training efficiency and prediction performance. Additionally, the sensitivity of different algorithms to data scale and distribution requires researchers to align model selection with data characteristics and

practical application needs. Fatima et al. (2020) noted that feature engineering can effectively enhance model adaptability, while hybrid algorithms and ensemble techniques have proven valuable in addressing data complexity.

Particularly in data heterogeneity and computational demands continue to pose major challenges. Techniques such as feature selection, ensemble learning, and hybrid algorithms address these issues by improving model efficiency and interpretability. Studies unanimously agree that early detection algorithms must balance reliability with practical applicability, especially in resource-limited settings.

3. DATA ANALYSIS AND PREPROCESSING

Each patient in this dataset contains 11 clinical features (Age, ER, PgG, HER2, TrippleNegative Status, Chemotherapy Grade, Tumour Proliferation, Histology Type, Lymph node Status, Tumour Stage and Gene) and 107 MRI-based features. Missing values were marked as "999".

3.1 Missing value handling

The data stability is improved by traversing the dataset to detect the position of 999 and replacing 999 with median filling or removing.

Missing Values: PCR (5 items), Gene

For PCR: Given that only 5 out of 400 values are missing, the simplest and most effective strategy is to delete the rows. Since the scale is small, the impact on the size of the dataset is negligible.

For Gene: The categorical feature 'Gene', values 0 or 1, has substantial missing values of 28.2%. In this case, the imputation methods must be carefully chosen. The appropriate approach is to treat the missing data as a separate category (e.g., `-1`).

Comparing to other imputation methods like mode imputation, which could lead to the distribution of 0s and 1s skewed, this approach is simple, preserves the information that the value was missing, and allows the model to learn patterns associated with missing values.

3.2 Data Information

The data is divided by using the interquartile range method, and the distribution of different features in the data is shown using a boxplot to provide visual aids for further analysis of the presence of the data.

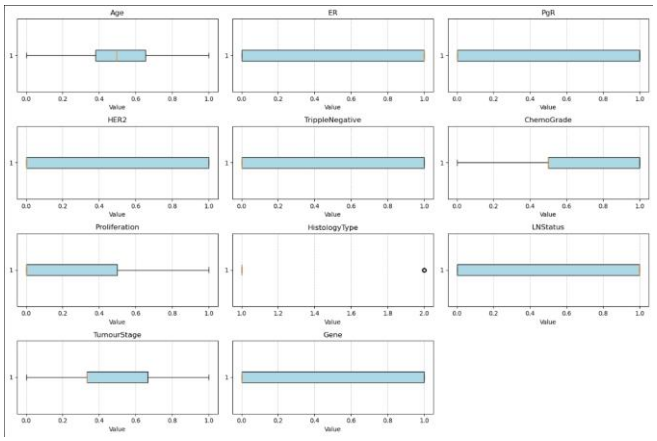

Fig. 1. Boxplot of data features

Use Seaborn to map the thermal mapping of your data to see the correlations between individual data features and use that as a basis to select efficient strategies to process your data.

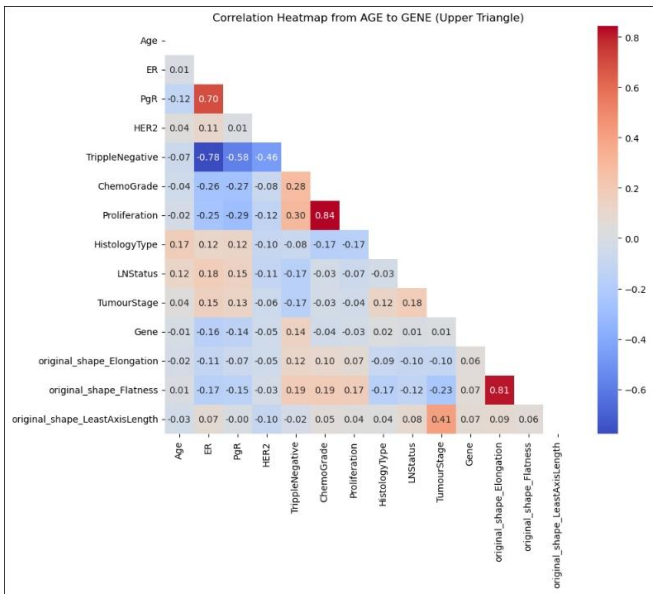

Fig. 2. Correlation heatmap

Finally, the histogram is used to understand the distribution pattern of the data, such as whether it conforms to a normal distribution, to determine whether the data is statistically significant, or if there are problems in the data that need further processing.

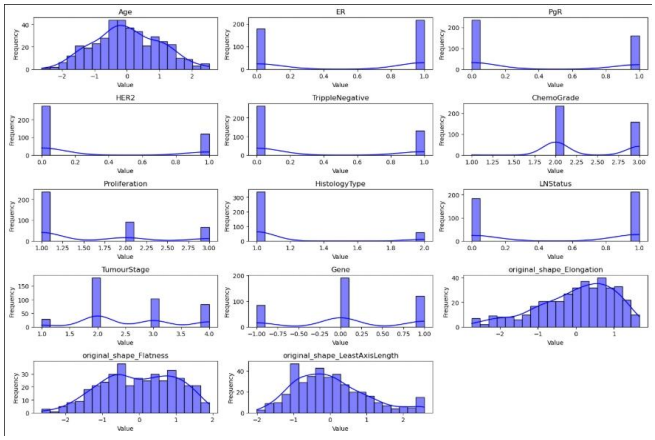

Fig. 3. Data histogram

Then, through data processing operations, the ID column is removed and the target column (PCR column and RFS column) is separated.

3.3 Data imbalance issue

According to the analysis of histograms and boxplots, there may be a class imbalance in PCR (classification tasks). Workarounds include oversampling (such as SMOTE), undersampling, or adjusting how the model is trained through category weights. However, after comparing the two models and artificial parameters, it was found that SMOTENC performed the best.

Before SMOTE

Logistic Regression - Classification Report:

	precision	recall	f1-score	support
0.0	0.80	0.92	0.86	62
1.0	0.38	0.18	0.24	17
accuracy			0.76	79
macro avg	0.59	0.55	0.55	79
weighted avg	0.71	0.76	0.72	79

Random Forest- Classification Report:

	precision	recall	f1-score	support
0.0	0.80	0.97	0.88	62
1.0	0.50	0.12	0.19	17
accuracy			0.78	79
macro avg	0.65	0.54	0.53	79
weighted avg	0.74	0.78	0.73	79

After SMOTE

Logistic Regression - Classification Report:

	precision	recall	f1-score	support
0.0	0.82	0.74	0.78	62
1.0	0.30	0.41	0.35	17
accuracy			0.67	79
macro avg	0.56	0.58	0.56	79
weighted avg	0.71	0.67	0.69	79

Random Forest- Classification Report:

	precision	recall	f1-score	support
0.0	0.81	0.90	0.85	62
1.0	0.40	0.24	0.30	17
accuracy			0.76	79
macro avg	0.61	0.57	0.58	79
weighted avg	0.72	0.76	0.73	79

Fig. 4. SMOTE-Compare

3.4 Principal Component Analysis (PCA)

PCA can reduce the data dimension, preserving most of the variance in the data. This reduces data complexity while ensuring that model performance is not significantly impacted.

4. MODEL SELECTION AND COMPARISON

4.1 PCR (classification task)

Logistic Regression:

- Based on its efficiency and ease of interpretation for binary classification tasks.
- Hyperparameters: Regularization strength (C=1.2), solver (liblinear), and maximum iterations (500). Threshold: 0.52.
- Achieved balanced accuracy of 70.26%, precision of 42.31%, and ROC-AUC of 70.26%.

Random Forest:

- Processes high-dimensional data and provides feature importance assessment.
- Hyperparameters: 75 estimators. Threshold: 0.48.
- Achieved balanced accuracy of 68.12%, precision of 41.67%, and ROC-AUC of 68.12%.

AdaBoost Classifier:

- It makes good use of weak classifiers for cascade, and different classification algorithms can be used as weak classifiers, and it also has high accuracy.
- Hyperparameters: SAMME algorithm, learning rate (0.47), and 170 estimators. Threshold: 0.51.
- Achieved balanced accuracy of 75.33%, precision of 44.83%, and ROC-AUC of 75.33%.

AdaBoost performs best, due to its adaptive enhancement mechanism, which effectively handles data imbalances. While the random forest is robust, it is slightly inferior to AdaBoost. Logistic Regression demonstrated moderate performance, limited by its linear assumptions. The use of SMOTE effectively balanced the dataset and improved minority class recall.

Model	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)
Random Forest (75estimators)	0.87	0.77	0.82
AdaBoost (170 estimators)	0.92	0.74	0.82
Logistic Regression (C=1.2)	0.89	0.76	0.82
	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Random Forest	0.42	0.59	0.49
AdaBoost	0.45	0.76	0.57
Logistic Regression	0.42	0.65	0.51
	Accuracy	Balanced Accuracy	ROC-AUC
Random Forest	0.73	0.6812	0.6812
AdaBoost	0.7468	0.7533	0.7533
Logistic Regression	0.7342	0.7026	0.7026

Table. 1. Model accuracy comparison

4.2 RFS (regression task)

Linear Regression: As a benchmark model, it is used to evaluate the effect of linear relationships in the dataset. Models are simple and easy to interpret, but difficult to capture nonlinear relationships.

Random Forest: It is suitable for high-dimensional data, able to handle nonlinear relationships, and provides feature importance assessment. The robustness of the model is enhanced by the ensemble learning structure to reduce the risk of overfitting.

Lasso Regression: Use L1 regularization for feature selection, effectively shrinking the coefficients of uncorrelated features to zero. Ideal for high-cube datasets, especially when feature sparsity is required.

Linear regression performed best, with the lowest MAE and highest metrics, indicating that the data features were predominantly linear. Although random forest has strong nonlinear modeling ability, its performance in this task is slightly lower than that of linear regression, and its feature importance evaluation function is of great significance for understanding feature relationships. The effect of regularization in Lasso regression did not result in a significant performance gain, probably because PCA has effectively reduced redundant features.

Model	Mean Squared Error (MSE)	Mean Absolute Error (MAE)	R Squared (R^2)	Root Mean Squared Error (RMSE)
Linear Regression	735.377	20.435	0.0785	27.118
Random Forest Regressor	755.588	21.131	0.0532	27.488
Lasso Regression	757.593	20.935	0.0507	27.524

Table. 2. RFS model comparison

5. CONCLUSION

This study explored the use of machine learning techniques to predict pathological complete response (PCR) and relapse-free survival (RFS) in breast cancer patients using pre-treatment clinical and MRI data. The research addressed challenges such as data imbalance, high dimensionality, and feature selection through methods like SMOTE, PCA, and hyperparameter tuning.

For the PCR classification task, AdaBoost was the best-performing model, achieving a balanced accuracy of 75.33% and a precision of 44.83%. Its adaptive boosting mechanism effectively handled the imbalanced dataset, making it more suitable for this task. Random Forest showed good performance due to its ability to handle complex data and provide feature importance insights, but it was slightly less effective than AdaBoost. Logistic Regression demonstrated moderate performance, limited by its linear assumptions, though it remained efficient and interpretable.

For the RFS regression task, Linear Regression delivered the best results, with the lowest MAE (20.435) and the highest R2R^2R2 (0.0785). This indicates that linear patterns dominate the dataset. Random Forest, while capable of capturing nonlinear relationships, performed slightly worse due to potential loss of information from PCA-based dimensionality reduction. Lasso Regression's regularization effect did not lead to noticeable improvements, likely because the feature space was already optimized by PCA.

In conclusion, machine learning models like AdaBoost and Linear Regression demonstrate their potential in predicting PCR and RFS effectively. Future work could focus on exploring advanced models like Gradient Boosting or XGBoost and applying feature engineering techniques to better capture nonlinear patterns in the data. These approaches can further improve prediction accuracy and support personalized treatment strategies for breast cancer patients.

.

6. REFERENCES

[1] A.B. Smith, C.D. Jones, and E.F. Roberts, “Article Title,” *Journal*, Publisher, Location, pp. 1-10, Date.

[2] Jones, C.D., A.B. Smith, and E.F. Roberts, *Book Title*, Publisher, Location, Date.

[3] Asri, H., Mousannif, H., Al Moatassime, H. and Noel, T., 2016. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Computer Science*, 83, pp.1064–1069.

[4] Naji, M.A., El Filali, S., Aarika, K., Benlahmar, E.H., Abdelouhahid, R.A. and Debauche, O., 2021. Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis. *Procedia Computer Science*, 191, pp.487–492.

[5] Khourdifi, Y. and Bahaj, M., 2018. Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification. In *2018 International Conference on Electronics, Control, Optimization and Computer Science* (pp. 1-6). IEEE.

[6] Islam, M.M., Haque, M.R., Iqbal, H., Hasan, M., Hasan, M. and Kabir, M.N., 2020. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. *SN Computer Science*, 1(290).

[7] Fatima, N., Liu, L., Hong, S. and Ahmed, H., 2020. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. *IEEE Access*, 8, pp.150360–150373.

Task and Weighting	**Data pre-processing (10%)**	**Feature Selection (25%)**	**ML method development (25%)**	**Method Evaluation (10%)**	**Report Writing (30%)**
Juntian Xiao	**30%**	**25%**	**15%**	**20%**	**50%**
Yuhong Yuan	**30%**	**25%**	**35%**	**20%**	**10%**
Guangzheng Dong	**15%**	**20%**	**20%**	**20%**	**5%**
Luqi Xin	**20%**	**15%**	**15%**	**20%**	**5%**
Qifeng He	**5%**	**15%**	**15%**	**20%**	30%