BREAST CANCER TREATMENT RESPONSE PREDICTION USING *SVM* AND *RANDOM FOREST*

Ceren Dinç [redacted]

University of Nottingham

ABSTRACT

Breast cancer is the most prevalent type of cancer in women. Pathological complete response (pCR) and relapse-free survival (RFS) are metrics used to measure the effectiveness of breast cancer treatment. After trialling a range of machine learning models, a support vector machine was chosen to predict pCR and Random Forest for RFS. When testing a balanced classification accuracy of 0.807 for pCR prediction and mean absolute error of 20.601 for RFS prediction were achieved.

***Index Terms*—** Breast cancer, pathological complete response, relapse-free survival, machine learning, support vector machine, random forest

1. INTRODUCTION

Breast cancer is the most often diagnosed cancer and leading cause of cancer death in women globally [1]. Pathological complete response (pCR) and relapse-free survival time (RFS) are common metrics for assessing the effectiveness of breast cancer treatment and quantifying survival. A pCR positive outcome indicates "the lack of all signs of cancer in tissue samples removed during surgery or biopsy after treatment" [2] (denoted by 1 in the data). Patients achieving pCR after neoadjuvant chemotherapy tend to have a favourable prognosis [3]. RFS is the "the length of time after primary treatment for a cancer ends that the patient survives without any signs or symptoms of that cancer" [4]. This paper details two machine learning models for predicting breast cancer treatment response based on a number of clinical and magnetic resonance imaging (MRI) based features: a classification model for pCR and a regression model for RFS.

2. RELATED WORK

Several studies have been conducted that apply Machine Learning methods to predicting breast cancer survival. Fatima et al. [5] conducted a meta-analysis, analysing the effectiveness of different machine learning models in predicting pCR for breast cancer patients. Their analysis has shown that SVM has been the most successful method. Gonzalez-Castro et al. [6] applied several algorithms to predict 5-year breast cancer recurrence from health data and concluded that extreme gradient boosting achieves the best average results, outperforming Decision Trees, Gradient Boost, Linear Regression and Deep Neural Network for all evaluation metrics. On the other hand, Lauritzen el al. [7] identified recurrent breast cancer patients in national health registries using machine learning and concluded that the performance of the RF model was superior to the performance of the LR model regardless of data source. Alongside machine learning techniques, they also looked at data mining techniques, with these techniques having a very high accuracy.

3. METHODOLOGY

3.1. Approach

Before developing any machine learning model, it is crucial to ensure the dataset is clean, consistent and properly structured. Effective data pre-processing is essential to prevent biases, inaccuracies and inefficiencies during training. By reducing noise through removal of redundant features and inconsistent samples, the machine learning algorithm can perform to an optimal standard.

The key steps in the data preprocessing phase include data cleaning, addressing missing values, handling anomalous samples, and dimensionality reduction. These steps are discussed in detail in sections 3.2-3.6

The chosen machine learning algorithms were applied to a portion of the processed dataset and subsequently optimised. For classification (predicting pCR), these were support vector machine (SVM), decision tree (DT), random forest (RF), linear regression (LR), naive bayes (NB) and multilayer perceptron (MLP). For regression (predicting RFS), the evaluated models were SVM, RF, LR and MLP. Their performance was evaluated and compared, to yield one optimal model for each task. The model development process is outlined in section 3.7, while the evaluation and justification for selecting the optimal models are presented in section 4.

3.2. Dataset

The original dataset is comprised of 400 samples and 121 features. The "ID" column serves as the unique identifier for each sample. The target variables for the classification and regression tasks are pCR and RFS, respectively. The dataset includes 11 clinical features: Age; ER (oestrogen receptor); PgG (proliferation grade); HER2 (human epidermal growth factor 2); triple negative status; chemotherapy grade; tumour proliferation; histology type; lymph node status; tumour stage; and gene. Additionally, there are 107 MRI-based features.

3.3. Missing values

Five samples in the dataset had missing pCR target values. Imputing these values without medical expertise was deemed inappropriate due to the potential impact on model performance. Additionally, having a single, unified dataset simplifies data preprocessing and model development. As a result, all five samples were removed from the dataset despite their RFS values not missing.

Missing values in the clinical features were imputed using the mode, stratified by pCR outcomes. This approach ensures the imputed data aligns with the distribution of each pCR group. Table 1 summarises the missing value counts in the clinical features, along with their mode values for each pCR outcome. There were no missing values in the MRI-based features.

Imputation was preferred to dropping the records with missing values, as the affected samples accounted for 87 out of 395 (22.0 %) samples. Removing them would have resulted in a substantial reduction in dataset size and a significant loss of valuable data.

3.4. Anomaly detection

Skewness was assessed for the MRI-based features, and many were found to be significantly above or below zero. Consequently, modified Z-scores (using the median and median absolute deviation) were used for univariate outlier detection as this method is more robust for skewed data.

Analysis of the histograms of modified Z-scores for each feature revealed spurious records where *original_ngtdm_Coarseness* had a value of 1,000,000 - this value was a clear outlier as it significantly exceeded the observed range of the other values (0.00024 to 1.34063). Isolating these samples for further inspection revealed additional abnormalities in other feature values, which were inconsistent with the rest of the dataset. Consequently, these thirteen samples were removed to preserve data integrity.

An Isolation Forest was used to calculate anomaly scores for the remaining samples in the dataset. A histogram of these scores is seen in Figure 1. Instead of removing all 24 records with negative scores, a threshold of -0.05 was selected based on visual inspection of the histogram.

Clinical feature	Count missing values pCR 0, 1	Mode value pCR 0, pCR 1
PgR	0, 1	0, 0
HER2	0, 1	0, 0
TrippleNegative	0, 1	0, 0
Chemograde	2, 1	2, 2
Proliferation	1, 1	1, 1
HistologyType	2, 1	1, 1
LNStatus	1, 0	1, 0
Gene	55, 30	0, 1

Table 1. For clinical features with missing values, the count of missing values and the mode value for that feature for each of pCR 0 and pCR 1

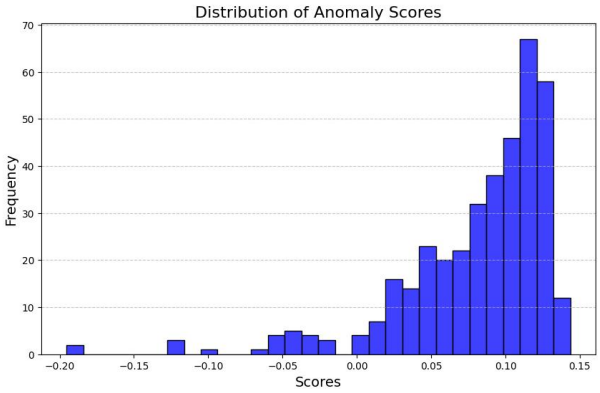

Fig. 1. Distribution of anomaly scores

3.5. Oversampling

After addressing the anomalies, 371 samples remained, with a pCR negative to positive ratio of 3.75:1. Imbalances in training data can often bias models toward the majority class, leading to suboptimal performance on minority class predictions. Oversampling is a common method to address this imbalance by increasing the representation of the minority class; however this carries risks in medical contexts. In this dataset, pCR-negative samples represent the majority class, aligning with real-world distributions of pCR outcomes. Due to the medical implications of false-positive predictions (missed pathological cases), oversampling of the minority class was deemed inappropriate. Training a model on data reflective of real-world conditions ensures better generalisation to unseen data, minimising bias and overfitting. As a result, oversampling was not performed to preserve the inherent characteristics crucial to produce a reliable and clinically applicable model.

3.6. Dimension Reduction & Feature Selection

The training data contained 107 continuous features derived from MRI scans of the tumour. This high-dimension feature space is therefore sparse in data samples, leading to reduced precision during model training. This challenge, commonly referred to as the 'curse of dimensionality' problem, is mitigated through techniques such as dimensionality reduction and feature selection.

The MRI features were first scaled using min-max normalisation and subsequently principal component analysis (PCA) was performed. Covariance between features was captured through eigenvectors, enabling dimensionality reduction while retaining as much of the original data variance as possible. PCA reduced the feature space to 8 dimensions, capturing 89.7 % of the original variance.

The optimal number of principal components was determined by identifying the largest perpendicular distance between the normalised cumulative variance curve (scaled to [0,1]) and the line y=x. This point represents a threshold where the rate of increase in explained variance diminishes relative to the addition of components, ensuring that the minimum number of components is selected to capture the maximum variance.

Although PCA assumes a normal distribution, it is not uncommon to apply the technique when the multivariate normality assumption is not met [8]. Following imputation and the removal of anomalies, the Henze-Zirkler test for multivariate normality was conducted to evaluate the data distribution. The test yielded a statistic of 1484 with a p-value less than 0.1, indicating that the data is not normally distributed. Therefore, while the principle components are uncorrelated, they are not independent, which limits their interpretability [9].

In addition to dimensionality reduction, the ten most significant features were identified to minimise irrelevance in the training data. The ER, HER2 and Gene features were retained as these have been shown to be clinically significant in predicting pCR and RFS outcomes [10]. The remaining features were selected using feature importances computed using random forest (RF). RF is a useful tool for feature selection because it is effective in high dimensional spaces, and has intrinsic feature ranking. RF can capture complex, non-linear feature interactions resulting in more robust feature ranking. The top 10 features identified by RF: **Gene**; **ER**, **HER2**, and $\mathbf{PCA_{0\text{-}6}}$.

3.7. Machine Learning Model Design

The following machine learning algorithms were evaluated for the classification task: SVM, DT, RF, LR, NB, and MLP. The following machine learning algorithms were evaluated for the regression task: SVM, RF, LR, and MLP. The selected models were chosen based on their general popularity in regression tasks and their demonstrated effectiveness on small datasets.

Component	Captured variance	Top 3 features
PCA_0	0.391	glcm difference variance; glcm ldmn; glcm contrast
PCA_1	0.169	shape mesh volume; shape voxel volume; shape least axis length
PCA_2	0.135	firstorder 90th percentile; firstorder root mean squared; firstorder mean absolute deviation

Table 2. The most importance features for the classification task

The dataset was initially split into 70:30 train:test sets, stratified by PCR outcome for the classification task. The training set was used to optimise the model through hyperparameter tuning, measured using five-fold cross-validation. A grid search was applied to models with multiple hyperparameters, identifying the optimal combination based on a chosen scoring metric. *Balanced accuracy score* (BA) was used for the classification task and *mean absolute error* (MAE) for the regression task.

The test set was reserved to evaluate the performance of the optimised model using the same scoring metrics applied during tuning (BA and MAE). The best-performing models were selected primarily based on these results, which are presented in Tables 3 and 4. However, additional considerations were made based on the analysis of the predictions, as discussed in section 4.

4. EVALUATION

The balanced accuracy scores of the classification models are in Table 3. Although the logistic regression (LR) and random forest (RF) models achieved higher balanced accuracy scores compared to the SVM (Support Vector Machine), the SVM demonstrated better recall for pCR-negative outcomes (0.863 compared to 0.75 for both RF and LR). Prioritising recall over precision or overall accuracy is critical to minimise the risk of missed pathological cases (false-positives), which could lead to over-prediction of cancer-free cases. The confusion matrices for the SVM, RF, and LR models are seen in tables 5, 6, and 7, respectively.

The final SVM model had a linear kernel with a regularisation term of 1.

The mean absolute error scores for the regression models are seen in Table 4. While none of the models achieved a satisfactory performance in terms of MAE, differences in the distribution of predicted values set them apart. The top-performing models were SVM and RF, with MAE scores of 20.601 and 20.972, respectively.

Model	Balanced classification accuracy on test
SVM	0.807
Decision tree	0.782
Random forest	0.812
Logistic regression	0.833
Naive Bayes	0.790
MLP	0.705

Table 3. Results for classification models

Model	Mean absolute error on test
SVM	20.601
Random forest	20.972
Linear regression	21.612
MLP	22.183

Table 4. Results for regression models

Despite SVM's marginally better performance, its predicted RFS values were constrained within a narrow range (54.234 to 54.658), which is unexpected given the dataset's full range (0.0 to 144.0). This likely stems from the Gaussian distribution of RFS values in the training data, which biases the model towards the mean. Consequently, significant prediction inaccuracies are masked by the majority being close to their true targets, highlighting issues with the model's generalisability.

In contrast, the RF model's predicted RFS values were more acceptable, ranging from 37.830 to 78.050. The predicted RFS values also demonstrated a Gaussian distribution comparable to the full training dataset, suggesting improved generalisability and robustness.

The final Random Forest model has 50 trees, no max depth, a minimum of 1 data point allowed in a leaf node and the square root of total number of features as the max number of features considered for splitting a node.

5. DISCUSSION

Dimension reduction with PCA captures more variance of the data in fewer features than feature selection alone. However, a primary limitation of this approach is that it results in a less interpretable model, as feature details must be inferred from the PCA vectoriser. As a result, the top three features for each component were identified by calculating PCA loadings (eigenvectors) and ranking them by their absolute eigenvalues. Table 2 summarises the most important features for the classification task derived using this method. However, it should be noted that due to the underlying data not having multivariate normality, the components are not independent. GLCM features (PCA0) and tumour size (PCA1) being important features is consistent with previous studies [11].

The SVM, Random Forest and Logistic Regression classification models performed best in terms of overall accuracy for predicting pCR. However, given that the preferable outcome is not over-predicting pCR positive outcomes, SVM was chosen as the 'best model' because it had higher recall than the other models despite being the third best performing for overall accuracy.

The MAE of the best regression model (RF) was 20.97, which is considerably large (14 %) when compared to the true RFS value range (0.0-144.0). This significant error poses challenges for its application in clinical settings. Even assuming the MAE is constant across all samples, an uncertainty of approximately 21 months is insufficient. For instance, if the prediction is used to schedule screening appointments for relapse testing, the appointment could be schedules years too late, jeopardising timely interventions. Furthermore, this study did not address factors such as race and marital status, which have been shown to influence survival [12]. Overall, RFS prediction is a less active field of research compared to pCR prediction, therefore, there is not a baseline against which to assess the performance of the regression model.

Actual / Predicted	Negative	Positive
Negative	76	12
Positive	6	18

Table 5. Confusion matrix for SVM classification

Actual / Predicted	Negative	Positive
Negative	66	22
Positive	3	21

Table 6. Confusion matrix for random forest classification

Actual / Predicted	Negative	Positive
Negative	66	22
Positive	2	22

Table 7. Confusion matrix for logistic regression classification

6. CONCLUSION

In summary, the SVM classifier demonstrated high accuracy (balanced accuracy of 0.807) in predicting pCR outcomes, this aligns with existing research. Conversely, while Random Forest was the best performing regression model for predicting relapse-free survival (MAE of 20.972), its performance would be inadequate for clinical application.

7. REFERENCES

[1] F. Bray, J. Ferlay, , I. Soerjomataram, R. L. Siegel, L. A. Torre, and A Jemal, "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. Dec 2018. https://doi.org/10.3322/caac.21492

[2] National Cancer Institute. (2024, Dec. 12). Definition of pathologic complete response [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pathologic-complete-response

[3] P. Cortazar, L. Zhang, M. Untch, K. Mehta, J. P. Costantino, N Wolmark, H Bonnefoi, D Cameron, L Gianni, P Valagussa, et al, "Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis," Lancet, vol. 384, no. 9938, pp. 164-172, Jul 2014.

[4] National Cancer Institute. (2024, Dec. 12) Definition of relapse-free survival [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/relapse-free-survival

[5] N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and their analysis," IEEE Access, vol. 8, pp. 150360-150376, Aug 2020.

[6] L. González-Castro et al, "Machine Learning Algorithms to Predict Breast Cancer Recurrence Using Structured and Unstructured Sources from Electronic Health Records," Cancers, vol. 15, no. 10, pp. 2741, May 2023. DOI: https://doi.org/10.3390/cancers15102741.

[7] A. D. Lauritzen, T. Berg, M. B. Jensen, M. Lillholm, and A. Knoop, "Identifying recurrent breast cancer patients in national health registries using machine learning," Acta Oncologica, vol. 62, no. 4, pp. 350–357, Apr 2023.

[8] I.T. Jolliffe, "Principal Component Analysis for Special Types of Data," Springer Series in Statistics. Springer, New York, NY. 2002. pp. 338-372. https://doi.org/10.1007/0-387-22440-8_13

[9] D. Kim, SK. Kim, "Comparing patterns of component loadings: Principal Component Analysis (PCA) versus Independent Component Analysis (ICA) in analyzing multivariate non-normal data," Behav Res, vol. 44, pp. 1239–1243, Feb 2012. https://doi.org/10.3758/s13428-012-0193-1

[10] B.C. Calhoun, L.C. Collins, "Predictive markers in breast cancer: An update on ER and HER2 testing and reporting,", Seminars in Diagnostic Pathology, vol. 32, no. 5, pp. 362-369, Sep 2015.

[11] M. L. Marinovich, F. Sardanelli, S. Ciatto, E. Mamounas, M. Brennan, P. Macaskill, L. Irwig, G. von Minckwitz, N. Houssami, "Early prediction of pathologic response to neoadjuvant therapy in breast cancer: Systematic review of the accuracy of MRI," Breast, vol. 21, no. 5, pp. 669-677, Oct 2012.

[12] M. Y. M. Naser, D. Chambers and S. Bhattacharya, "Prediction Model of Breast Cancer Survival Months: A Machine Learning Approach," SoutheastCon 2023, Orlando, FL, USA, 2023, pp. 851-855, doi: 10.1109/SoutheastCon51012.2023.10115220.

A. CONTRIBUTION TABLE

Task and Weighting	Data pre-processing (10%)	Feature Selection (25%)	ML method development (25%)	Method Evaluation (10%)	Report Writing (30%)
Ceren Dinç	50%	21%	25%	5%	10%
Sylwia Sajdak	50%	21%	15%	0%	20%
Effie Menzies	0%	16%	10%	45%	30%
Jaber Ahmed	0%	21%	25%	25%	20%
Dominic Binks	0%	21%	25%	25%	20%