Machine Learning for Predicting Pathological Complete Response and Relapse-Free Survival in Breast Cancer Patients Using Clinical and Radiomic Features

1st Farhan Shaikh
Department of Computer Science
University of Nottingham
Nottingham, UK

2nd Mohamed Ramiz Latif
Department of Computer Science
University of Nottingham
Nottingham, UK

3rd Abdulla Al-Ali
Department of Computer Science
University of Nottingham
Nottingham, UK

[redacted]

4th Ghalia Khan
Department of Computer Science
University of Nottingham
Nottingham, UK

5th Muhammad Saad
Department of Computer Science
University of Nottingham
Nottingham, UK

[redacted]

***Abstract*—Breast cancer treatment planning requires accurate prediction of treatment response and survival outcomes. This study presents machine learning approaches to predict pathological complete response (pCR) and relapse-free survival (RFS) using clinical features and magnetic resonance imaging (MRI) radiomic features from 400 breast cancer patients. For pCR prediction, we evaluated multiple preprocessing strategies, baseline models, and addressed class imbalance using SMOTE, achieving a balanced accuracy of 0.639 (±0.071) with an MLP classifier. For RFS prediction, we employed feature selection and regression models, achieving a mean absolute error (MAE) of 20.29 days using an optimized Random Forest regressor with 30 selected features. Both approaches demonstrate the effectiveness of combining clinical and radiomic features for personalized treatment planning.**

***Index Terms*—Breast cancer, pathological complete response, relapse-free survival, machine learning, radiomics, classification, regression**

I. Introduction

Breast cancer is the most common cancer among women worldwide, with chemotherapy being a commonly used treatment strategy to reduce tumor size before surgery [1]. Pathological complete response (pCR), defined as complete tumor resolution at surgery, has a high likelihood of achieving a cure and longer relapse-free survival (RFS). However, only approximately 25% of patients receiving chemotherapy achieve pCR, with the remaining 75% having residual disease [2]. RFS, the length of time after primary treatment without cancer recurrence, is essential for treatment planning and patient stratification, with recurrence rates of 10-15% within 5 years of diagnosis [3].

Traditional prognostic factors include ER, HER2 status, and tumor grade. Recent advances enable extraction of quantitative radiomic features from MRI scans, capturing tumor heterogeneity and providing additional prognostic information [5], [15].

This study addresses two complementary challenges: predicting pCR (binary classification) and RFS (regression) using a combination of 11 clinical features and 107 MRI-derived radiomic features from 400 patients in the I-SPY 2 TRIAL dataset. The pCR task presents significant class imbalance (25% positive), requiring specialized techniques such as SMOTE to address the skewed distribution. The RFS task faces high dimensionality ($p = 118$ features) relative to sample size ($n = 400$), with $n/p \approx 3.4$, necessitating feature selection to prevent overfitting. We employ advanced preprocessing techniques, feature selection, multiple baseline models, and class balancing strategies to build robust predictive models for both tasks.

II. Methods

A. Dataset

The dataset consists of 400 breast cancer patients with 121 features per patient: 11 clinical features (Age, ER, PgR, HER2, TripleNegative status, ChemoGrade, Proliferation, HistologyType, LymphNodeStatus, TumourStage, and Gene) and 107 radiomic features extracted from MRI using PyRadiomics [5]. Missing values were encoded as 999 in the original dataset. For pCR prediction, the target is binary (0 = no pCR, 1 = pCR achieved), with approximately 300 patients (75%) classified as pCR negative and 100 patients (25%) achieving pCR. For RFS prediction, the target is continuous (days until recurrence or censoring). The dataset was split 80/20 for train/test evaluation with stratification for pCR and random state 42 for reproducibility.

TABLE I
DATASET CHARACTERISTICS

Characteristic	Value
Total Patients	400
Clinical Features	11
Radiomic Features	107
Total Features	121
pCR Negative (Class 0)	∼300 (75%)
pCR Positive (Class 1)	∼100 (25%)
Train/Test Split	80/20 (stratified)
Training Samples	320
Test Samples	79 (pCR) / 80 (RFS)

B. Data Preprocessing

1) pCR Prediction Pipeline: Rows with pCR outcome equal to 999 (ambiguous labels) were removed from the dataset. All remaining instances of 999 were treated as missing values and converted to NaN for downstream imputation. Given the heterogeneous nature of the features, we implemented separate preprocessing pipelines for different feature types:

Numerical Features: Age and all 107 radiomic features (prefixed with 'original_') were processed through a pipeline that included: (1) **Outlier Capping:** Custom transformer using the Interquartile Range (IQR) method with a factor of 1.5 to cap outliers, preventing data leakage by computing limits only on training data. For a feature x, outliers beyond $[Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR}]$ were capped, where Q_1 and Q_3 are the first and third quartiles, and $\text{IQR} = Q_3 - Q_1$; (2) **Standard Scaling:** Z-score normalization to standardize feature distributions, transforming each feature x to $z = \frac{x-\mu}{\sigma}$, where μ and σ are the mean and standard deviation computed on training data.

Ordinal Features: ChemoGrade, Proliferation, HistologyType, and TumourStage were processed using: (1) **Imputation:** Most frequent value imputation for missing data; (2) **One-Hot Encoding:** Binary encoding with first category dropped to avoid multicollinearity.

Binary Features: ER, PgR, HER2, TripleNegative, LNStatus, and Gene were processed using most frequent value imputation. All transformations were combined using a `ColumnTransformer` to ensure consistent preprocessing across all models and prevent data leakage.

2) RFS Prediction Pipeline: Missing values (999) were converted to NaN and imputed using median-based SimpleImputer (robust to outliers) and KNNImputer (k=5). Three scaling approaches were compared: no scaling, StandardScaler ($z = \frac{x-\mu}{\sigma}$), and RobustScaler ($z = \frac{x-\text{median}(x)}{\text{IQR}(x)}$). RobustScaler demonstrated superior performance. Median imputation performed comparably to KNN, suggesting independent feature imputation was sufficient.

C. Feature Selection (RFS Task)

With $p = 118$ features and $n = 400$ samples ($n/p \approx 3.4$), we face the curse of dimensionality. Feature selection was essential to reduce overfitting and improve generalization. Per assignment requirements, ER, HER2, and Gene features were always retained.

Four feature selection methods were evaluated: (1) **Mutual Information:** Filter method measuring statistical dependency $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$ between features X and target Y, capturing non-linear relationships; (2) **Lasso Regression:** Embedded method using L1 regularization to minimize $\frac{1}{2n}\|y - X\beta\|^2 + \lambda\|\beta\|_1$, where λ controls sparsity, driving irrelevant feature coefficients to zero [8], effective for handling multicollinearity; (3) **Random Forest Importance:** Embedded method measuring feature importance via mean decrease in impurity [6], handling non-linear relationships and feature interactions; (4) **PCA Hybrid:** Dimensionality reduction keeping all 11 clinical features and applying PCA to 107 radiomic features, preserving clinical interpretability while compressing radiomics through linear transformation $Y = XW$, where W contains principal components.

Feature counts from 10 to 50 were evaluated using 5-fold cross-validation. Random Forest importance with 30 features achieved the best performance, reducing MAE from 20.93 (all 118 features) to 20.48, representing a 2.2% improvement.

D. Model Selection and Evaluation

1) pCR Prediction Models: Four baseline models were evaluated: (1) **CatBoost:** Gradient boosting with built-in categorical handling [11] (iterations=200, learning_rate=0.03, depth=6); (2) **LightGBM:** Fast gradient boosting [10] (num_leaves=31, max_depth=1, learning_rate=0.1, n_estimators=100); (3) **XGBoost:** Optimized gradient boosting [9] (n_estimators=100, max_depth=6, learning_rate=0.3); (4) **MLP:** Multi-Layer Perceptron neural network with forward propagation $\mathbf{h}^{(l)} = \sigma(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)})$, where σ is the activation function (ReLU: $\sigma(x) = \max(0,x)$ or tanh: $\sigma(x) = \tanh(x)$), $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are weights and biases at layer l, with hidden_layer_sizes=(100, 50), activation='relu', solver='adam'.

Learning curves (40%, 60%, 80%, 100% training data) showed improvement with more data. LDA (maximizing $\frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$) decreased performance, suggesting discriminative information is lost in dimensionality reduction.

For the MLP model, RandomizedSearchCV was used with 20 iterations to optimize: hidden layer sizes [(50,50), (100,50), (150,100)], activation function ['relu', 'tanh'], L2 regularization (alpha) [0.0001, 0.001, 0.01], learning rate [0.0005, 0.001, 0.005], and maximum iterations [300, 500]. The search was performed with LDA preprocessing and balanced accuracy as the scoring metric.

Given class imbalance (75% negative, 25% positive), SMOTE [4] ($k = 3$) was integrated, generating synthetic samples as $\mathbf{x}_{\text{new}} = \mathbf{x}_i + \delta \times (\mathbf{x}_{zi} - \mathbf{x}_i)$ where $\delta \in [0,1]$.

Models were evaluated using 5-fold stratified cross-validation with balanced accuracy $\text{BalAcc} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right)$.

2) RFS Prediction Models: Seven regression models were evaluated across all preprocessing pipelines: (1) Linear models: Ridge regression minimizing $\|y - X\beta\|^2 + \alpha\|\beta\|_2^2$ [12], Lasso with L1 regularization, ElasticNet combining L1 and L2 penalties; (2) Tree-based: Random Forest [6], Gradient Boosting [16]; (3) Non-parametric: Support Vector Regression (SVR) [13] with ϵ-insensitive loss, K-Nearest Neighbors (KNN) [14] using $\hat{y} = \frac{1}{k}\sum_{i \in N_k(x)} y_i$ where $N_k(x)$ are the k nearest neighbors. Random Forest with RobustScaler preprocessing achieved the best baseline performance (CV MAE: 20.93 ± 1.45), outperforming StandardScaler and unscaled approaches.

Random Forest hyperparameters were optimized via RandomizedSearchCV (50 iterations) followed by GridSearchCV. Advanced models (XGBoost [9], LightGBM [10], CatBoost [11]) and ensembles (Voting: $\hat{y} = \frac{1}{M}\sum_{i=1}^{M} \hat{y}_i$, Stacking) were evaluated but did not outperform optimized Random Forest. Evaluation used 5-fold cross-validation with MAE: $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$.

III. Results

A. pCR Prediction Results

Exploratory analysis confirmed class imbalance (75% negative, 25% positive). LDA decreased performance across all models (Table II), confirming discriminative information is lost in dimensionality reduction.

TABLE II
pCR: Model Performance with LDA (Test Set Balanced Accuracy)

Model	Balanced Accuracy
MLP	0.513
XGBoost	0.483
CatBoost	0.521
LightGBM	0.521

RandomizedSearchCV for MLP with LDA preprocessing identified optimal parameters: hidden_layer_sizes=(150, 100), activation='tanh', alpha=0.01, learning_rate_init=0.005, max_iter=300, solver='adam', learning_rate='adaptive'. This configuration achieved a CV balanced accuracy of 0.629.

Integrating SMOTE into the MLP pipeline significantly improved performance. The final model configuration combined: preprocessing pipeline (outlier capping, scaling, encoding), SMOTE oversampling (k=3), and optimized MLP. The final MLP+SMOTE model achieved: CV balanced accuracy of 0.639 (±0.071) and test balanced accuracy of 0.591.

The confusion matrix on the test set ($n = 79$) showed 55 true negatives (TN), 7 false positives (FP), 12 false negatives (FN), and 5 true positives (TP). The model demonstrated good recall for the majority class ($\text{Recall}_0 = \frac{\text{TN}}{\text{TN+FP}} = 0.89$) but lower recall for the minority class ($\text{Recall}_1 = \frac{\text{TP}}{\text{TP+FN}} = 0.29$), reflecting the persistent challenge of class imbalance even with SMOTE. The precision for the positive class was $\text{Precision}_1 = \frac{\text{TP}}{\text{TP+FP}} = 0.42$, indicating that when the model predicts pCR, it is correct approximately 42% of the time.

TABLE III
pCR: Final Model Performance Metrics

Metric	Value
CV Balanced Accuracy	0.639 ± 0.071
Test Balanced Accuracy	0.591
Test Accuracy	0.759
Precision (Class 0)	0.82
Recall (Class 0)	0.89
F1-Score (Class 0)	0.85
Precision (Class 1)	0.42
Recall (Class 1)	0.29
F1-Score (Class 1)	0.34

B. RFS Prediction Results

Table III summarizes baseline model performance across preprocessing pipelines. Random Forest with RobustScaler achieved the lowest MAE (20.93 ± 1.45), outperforming StandardScaler and unscaled approaches. SimpleImputer with median strategy performed comparably to KNNImputer, suggesting that for this dataset, independent feature imputation was sufficient.

TABLE IV
RFS: Baseline Model Performance (CV MAE)

Model	S+Rob	K+Rob	S+Z
Random Forest	**20.93**	21.05	20.96
Grad. Boosting	22.09	22.01	22.09
SVR	21.69	21.72	21.69
KNN	24.38	22.32	22.00
Ridge	33.21	25.38	33.77
Lasso	30.58	21.07	30.58

Feature selection significantly improved model performance. Random Forest importance with 30 features reduced MAE from 20.93 (all 118 features) to 20.48, representing a 2.2% improvement. Table IV compares feature selection methods, showing that Random Forest importance outperformed Mutual Information (20.52), Lasso (20.58), and PCA Hybrid (20.65).

TABLE V
RFS: Feature Selection Method Comparison

Method	Features	CV MAE
Baseline (All features)	118	20.93
Mutual Information	30	20.52
Lasso	30	20.58
RF Importance	**30**	**20.48**
PCA Hybrid (15 components)	26	20.65

Hyperparameter optimization further reduced MAE to 20.29. The optimal Random Forest configuration was: n_estimators=250, max_depth=5, min_samples_split=7, min_samples_leaf=5, max_features='sqrt', bootstrap=False, criterion='absolute_error'. Advanced gradient boosting models (XGBoost, LightGBM, CatBoost) and ensemble methods (Voting and Stacking) were evaluated but did not outperform the optimized Random Forest.

The final model achieved: cross-validation MAE of 20.29 ± 1.42 days and test set MAE of 20.81 days. The close agreement between CV and test MAE (difference ¡ 0.6 days) indicates good generalization without overfitting.

IV. Discussion

Our results demonstrate that combining clinical and radiomic features with appropriate preprocessing can effectively predict both pCR and RFS. Key findings include:

Preprocessing: For pCR prediction, the use of separate pipelines for numerical, ordinal, and binary features, combined with outlier capping to prevent data leakage, provided a robust foundation for model training. The IQR-based outlier capping was particularly important given the presence of extreme values in radiomic features. For RFS prediction, RobustScaler outperformed StandardScaler, likely due to its robustness to outliers in medical survival data. Simple median imputation performed comparably to KNN imputation, suggesting that for this dataset, feature independence assumptions were reasonable.

Feature Selection (RFS): Reducing features from 118 to 30 improved performance, confirming the curse of dimensionality with N/p ratio of 3.4. Random Forest importance was most effective, likely because it captures non-linear relationships and feature interactions that linear methods (Lasso) or filter methods (Mutual Information) miss. The PCA Hybrid approach preserved clinical interpretability but performed slightly worse, suggesting that the full radiomic feature space contains valuable discriminative information.

Model Selection: For pCR prediction, MLP with SMOTE outperformed tree-based models (CatBoost, LightGBM, XGBoost), suggesting that neural networks can effectively learn complex non-linear relationships in the high-dimensional feature space. The tanh activation function and deeper architecture (150, 100) provided better capacity for learning discriminative patterns compared to shallower networks. For RFS prediction, Random Forest outperformed linear models (Ridge, Lasso, ElasticNet) and other tree-based methods (Gradient Boosting). Its ability to handle non-linear relationships, feature interactions, and provide feature importance made it well-suited for this high-dimensional, small-sample problem.

Class Imbalance (pCR): SMOTE significantly improved model performance, increasing CV balanced accuracy from baseline models. However, the test set performance (0.591 balanced accuracy) indicates that the model still struggles with the minority class, achieving only 29% recall for pCR-positive cases. This suggests that additional techniques such as class weights, different SMOTE parameters, or ensemble methods might further improve performance. The precision for the positive class (0.42) indicates moderate confidence when predicting pCR.

Dimensionality Reduction: LDA did not improve performance for pCR prediction, likely because it reduces the feature space too aggressively and loses discriminative information present in the original high-dimensional radiomic features. This finding suggests that the full feature dimensionality is beneficial for neural network models.

Limitations: The dataset size (400 patients) limits the complexity of models that can be reliably trained. For pCR, severe class imbalance (25% positive) and the test set performance gap (CV: 0.639 vs. Test: 0.591) suggest some overfitting, though the difference is within reasonable bounds. For RFS, the small sample size may make cross-validation performance optimistic. External validation on independent datasets would strengthen the findings. The persistent challenge of class imbalance in pCR prediction, reflected in lower recall for the minority class, suggests that future work should explore additional balancing techniques, cost-sensitive learning, or ensemble methods.

V. Conclusion

This study presents machine learning pipelines for predicting pCR and RFS in breast cancer patients using clinical and radiomic features. For pCR prediction, we achieved a balanced accuracy of 0.639 using cross-validation and 0.591 on the test set with an MLP classifier combined with SMOTE oversampling. For RFS prediction, we achieved a MAE of 20.29 days using an optimized Random Forest regressor with 30 selected features. Both approaches demonstrate the value of combining clinical biomarkers with quantitative imaging features, potentially supporting personalized treatment decisions.

Future work should explore additional balancing techniques for pCR, cost-sensitive learning, deep learning approaches, and validation on larger, multi-institutional datasets to improve generalizability and clinical applicability.

Acknowledgment

This work uses data from the I-SPY 2 TRIAL. We acknowledge the use of PyRadiomics for feature extraction.

References

[1] Esserman, L. J., et al. "Pathologic complete response predicts recurrence-free survival more effectively by cancer subset: results from the I-SPY 1 TRIAL–CALGB 150007/150012, ACRIN 6657." *Journal of Clinical Oncology*, vol. 30, no. 26, pp. 3242-3249, 2012.

[2] Cortazar, P., et al. "Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis." *The Lancet*, vol. 384, no. 9938, pp. 164-172, 2014.

[3] Yu, Y., Ren, W., He, Z., et al. "Machine learning radiomics of magnetic resonance imaging predicts recurrence-free survival after surgery and correlation of LncRNAs in patients with breast cancer: a multicenter cohort study." *Breast Cancer Research*, vol. 25, no. 132, 2023.

[4] Chawla, N. V., et al. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[5] van Griethuysen, J. J. M., et al. "Computational Radiomics System to Decode the Radiographic Phenotype." *Cancer Research*, vol. 77, no. 21, pp. e104-e107, 2017.

[6] Breiman, L. "Random Forests." *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[7] Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[8] Tibshirani, R. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267-288, 1996.

[9] Chen, T., and Guestrin, C. ”XGBoost: A Scalable Tree Boosting System.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.

[10] Ke, G., et al. ”LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] Prokhorenkova, L., et al. ”CatBoost: unbiased boosting with categorical features.” *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[12] Hoerl, A. E., and Kennard, R. W. ”Ridge regression: Biased estimation for nonorthogonal problems.” *Technometrics*, vol. 12, no. 1, pp. 55-67, 1970.

[13] Drucker, H., et al. ”Support vector regression machines.” *Advances in Neural Information Processing Systems*, vol. 9, 1996.

[14] Cover, T., and Hart, P. ”Nearest neighbor pattern classification.” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.

[15] Aerts, H. J., et al. ”Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach.” *Nature Communications*, vol. 5, no. 4006, 2014.

[16] Friedman, J. H. ”Greedy function approximation: a gradient boosting machine.” *Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

CONTRIBUTION TABLE

TABLE VI
INDIVIDUAL CONTRIBUTION TO ASSIGNMENT TASKS

Member	**Data Analysis**	**Preproc. & Imput.**	**Feature Select.**	**Model Dev.**	**Report Writing**
Muhammad Saad	20%	20%	20%	20%	20%
Abdulla Al-Ali	20%	20%	20%	20%	20%
Galia Khan	20%	20%	20%	20%	20%
Farhan Shaikh	20%	20%	20%	20%	20%
Mohamed Ramiz Latif	20%	20%	20%	20%	20%