PCR AND RFS PREDICTION ON BREAST CANCER DATASET

Shahabaldin Mortazavi
Emaan Bashir
Kalu, Stephen Eke
Bushra Jalali
Mustafa Mehmood

ABSTRACT

This study develops predictive models for Pathological Complete Response (PCR) and Relapse-Free Survival (RFS) in breast cancer, utilizing a numeric dataset with 10 clinical and 107 MRI-based features. The unbalanced dataset underwent preprocessing, including missing value handling and SMOTE balancing. Feature extraction techniques like RandomForest importance, ANOVA, and Pearson correlation, along with PCA, were assessed. Models including Random Forest, XGBoost, SVM, and MLP for classification focusing on hyperparameter tuning, and SVR, Random Forest, and XGBoost for regression, were explored whose performance was assessed using F1-score, mean absolute error and cross-validation. XGBoost emerged as the most effective for PCR prediction, despite challenges with the positive class. Regression models exhibited varied performances, highlighting the importance of feature selection and outlier handling in predictive accuracy for breast cancer outcomes.
Intex Terms. - Machine Learning, Breast Cancer

1. INTRODUCTION

Breast cancer remains one of the most prevalent and challenging diseases, with early and accurate prediction of treatment outcomes like Pathological Complete Response (PCR) and Relapse-Free Survival (RFS) being crucial for effective patient management. This research delves into the development of predictive models for these critical outcomes, leveraging a dataset encompassing clinical and MRI-based features.

The study's focus extends beyond traditional analysis, employing machine learning techniques to interpret the complex patterns inherent in breast cancer data. Recognizing the limitations and potential biases in the original dataset, significant emphasis is placed on preprocessing, including innovative approaches to handle missing values and address data imbalance. It employs advanced feature extraction methods, such as RandomForest feature importance, Analysis of Variance (ANOVA), and Pearson correlation, alongside Principal Component Analysis (PCA), to enhance the predictive power of the models.Utilizing a range of classification models, the research aims to identify the most effective approach for PCR and RFS prediction.

By exploring the relationships between a wide array of clinical and numeric features extracted from MRI imaging, the aim is to create accurate predictive models for PCR and also to offer a framework for future research in the application of machine learning in Breast Cancer diagnostics and prognostics.

2. RELATED WORK

The integration of machine learning (ML) in breast cancer prognosis, marked by key studies, has significantly influenced our current research. Cruz and Wishart's (2016) study on SVM and Random Forest algorithms in breast cancer survival prediction using clinical data set foundational methodologies in the field. Boeri et al. (2020) further explored this, evaluating ML techniques like SVM and ANN, focusing on recurrence and mortality, providing insights into model predictive capabilities.

Li et al. (2021) contributed a systematic review, underscoring the expanding application of ML in breast cancer survival prediction. This review encompassed a range of ML algorithms, examining their methodological quality and performance metrics, highlighting ML's potential in health informatics. Additionally, another study by Li Y et al. (2021) developed ML models for predicting breast cancer recurrence, using algorithms like logistic regression, random forest, and XGBoost. This study emphasized the importance of feature selection and interpretation using SHAP.

These studies collectively enrich our research, demonstrating diverse methodologies in breast cancer survival and recurrence prediction using ML. They lay the groundwork for further advancements in applying ML in breast cancer prognosis.

.

3. METHOD

3.1. Data Collection

The data has been extracted from The American College of Radiology Imaging and it contains 10 clinical features and 107 MRI-based features. The number of samples in the dataset is 400. Moreover, 17 missing values were detected in the dataset.

3.2. Preprocessing

For classification regarding the prediction of PCR, all 5 missing values for the target variable were removed. All clinical features were treated as ordinal categorical, or binary based on the nature of the feature. Since it is essential to keep the order of these features, techniques like one-hot encoding were not implemented. For handling missing values in the features, median imputation was chosen as the imputation method. P was dropped.

Since the dataset is imbalanced, an oversampling technique named Synthetic Minorit Over- sampling SMOTE was used to make the dataset balanced. Before oversampling, 78% of the data was labelled as negative and 22% positive.

For regression, median imputation was used to handle the missing values. The outliers in the MRI-based numerical features were managed using a percentile-based technique where the outliers were constrained by setting them to a predetermined maximum or minimum percentile value.

3.3. Model Selection

For classification, the models that have been used are random forest, XGBoost, SVM and MLP. Since random forest and XGBoost are robust to outliers and scale of the features, normalisation (standard scaling) was used just for SVM and MLP Classifiers.

For dimensionality reduction, PCA was used, and the number of components was considered as a hyperparameter that was tuned. As a Feature Selection technique, random forest feature selection were implemented. We also considered using class weighing techniques and ANOVA for feature selection for classification task , but initial results were not favorable.

For regression, the primary models used were SVR, Random Forest, and XGBoost. MLP was initially tested but not selected due to overfitting issues. Feature selection played a critical role, with techniques such as ANOVA and Random Forest importance being used to identify the most important features. The SVR model saw improvements through the use of Pearson correlation and ANOVA, effectively reducing the feature count. Similarly, for the Random Forest and XGBoost models, the same feature selection methods were employed to identify the most significant features. Dimensionality reduction was addressed through PCA (linear and Kernel), with the number of components as an essential hyperparameter.

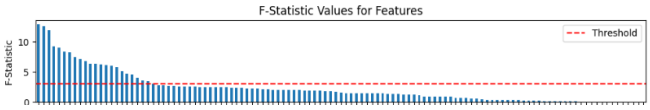


Figure 1- F-statistic values for ANOVA

3.4. Hyperparameter Tuning

For Classifiers, all the hyperparameters were tuned based on nested 3-fold cross validation. The goal was to identify optimal hyperparameter configurations for each model, ensuring robust and generalizable performance. In each fold SMOTE oversampling was implemented in order to have a balanced dataset before training. Moreover, as mentioned before, the number of components of the PCA was tuned as a hyperparameter. The most important hyperparameters to be tuned are shown in Table I.

For regression a 5-fold cross validation was used to tune the hyperparameters. GridSearchCV was applied on the train set to get the optimal values for a minimum absolute error.

TABLE I-HYPERPARAMETER TUNING

Classifier	F1-Score
SVC	Kernel, gamma, degree, coef0
Random Forest	n_estimators, min sample split, min sample leaf, max depth
XGBoost Classifier	N_estimator, min child weight, learning rate, gamma, max depth, alpha

4. EVALUATION

For classification, after tuning the hyperparameters, the result of the models based on the Macro Average F1 Score is shown in Table II-Classifiers Performance.

TABLE II-CLASSIFIERS PERFORMANCE

Classifier	F1-Score
SVC	0.43,0.56,0.51
Random Forest	0.43,0.49,0.47
XGBoost Classifier	0.54,0.55,0.48

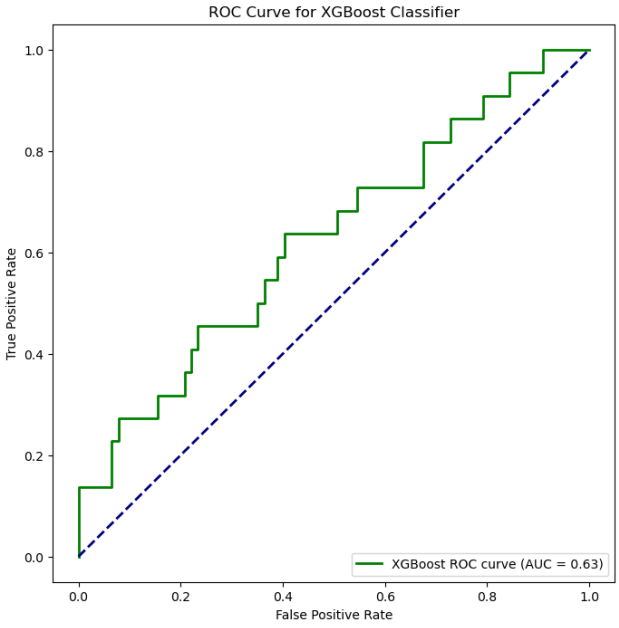
Selecting the F1 Score as the evaluation metric for the test set is driven by the imbalanced nature of the dataset. Since the primary results have not been significantly

different, a t-test was implemented in order to choose the best model. In the t-test experiment, the accuracy metric was chosen, and it has been implemented on a balanced test dataset to ensure the validity of the results. The P-value for the hypothesis test between XGBoost and Random Forest was 0.0004, which is less than 0.05. Hence by considering the fact that Mean Accuracy for XGBoost is 0.76 and for Random Forest is 0.7, the null hypothesis is rejected and it can be concluded that XGBoost has performed better. In the same way the P-Value for the comparison between XGBoost and SVM isa very negligible value of 2.19e-17 and the Mean Accuracy for this model has been 0.54. Hence, the best performing model has been XGBoost on this dataset for predicting PCR.

It is important to mention that dimensionality reduction technique (PCA) has not improved the performance of the best model. Furthermore, a feature selection technique has been implemented for the XGBoost classifier, in which the features with feature importance of more than 0.01 were selected, however, this did not yield a substantial improvement in the model's performance.

For a better and more detailed evaluation of the final model, the ROC curve is shown in the Figure2. The

Figure 2 - ROC-curve for XGBOOST



area under the curve is 0.63 for the test set and the model has been trained on the training dataset after oversampling. The test dataset is 25 percent of the whole dataset to achieve a reliable result with more that 100 samples in it.

For regression, the models were assessed using R-squared score and mean absolute error (MAE). A 5-fold cross validation was used to validate the performance of the model. The model was trained and evaluated iteratively across multiple folds to ensure its robustness and generalisation.

TABLE III - SVR PERFORMANCE

Feature Selection	**Outlier Handling**	**R2 Score**	**MAE**
No Feature Selection	No	-0.008	21.23
Pearson & ANOVA Feature Selection without PCA	No	0.001	22.13
Pearson & ANOVA Feature Selection + PCA	Yes	0.005	21.08
RFE Feature Selection without PCA	Yes	-0.007	21.1
RFE Feature Selection + PCA	Yes	-0.007	21.2
Pearson & ANOVA Feature Selection without PCA	Yes	0.004	21.2

TABLE IV - RANDOM FOREST PERFORMANCE

Feature Selection	**Outlier Handling**	**R2 Score**	**MAE**
No Feature Selection	Yes	0.008	21.14
RF Feature Selection	Yes	0.000	21.03
Linear PCA	Yes	-0.006	21.51
Kernel PCA	Yes	0.001	21.51

TABLE V - XGBOOST PERFORMANCE

Feature Selection	**Outlier Handling**	**R2 Score**	**MAE**
No Feature Selection	No	0.043	20.864
XGBoost Feature Importance	No	0.012	21.175
PCA	No	-0.021	21.405
PCA	Yes	-0.016	21.319
ANOVA Threshold 1	No	0.052	20.784
ANOVA Threshold 2	No	0.037	21.065
ANOVA Threshold 3	No	0.086	20.329

ANOVA Threshold 4	No	0.075	20.596
ANOVA Threshold 3	Yes	0.099	20.261
ANOVA Threshold 4	Yes	0.097	20.266

5. DISCUSSION

For predicting PCR, despite the fact that oversampling was implemented, our model performance for predicting the positive class has not been satisfactory. The precision and recall for the positive class have been in the range 0.19-0.27 and 0.11-0.46 for different test datasets. This is a challenge that can be encountered with gathering more positive classes. Since the number of samples is very low (400 samples), undersampling techniques were not successful.

Possible explanations for the suboptimal performance could be attributed to the complexity of the underlying patterns in the data, the presence of noisy features, or the need for more sophisticated modelling techniques. Further exploration into the intricacies of the dataset and experimentation with alternative methodologies, such as different feature engineering strategies or advanced ensemble techniques, may be warranted.

Additionally, comprehensive model evaluation, including the examination of the receiver operating characteristic (ROC) curve and area under the curve (AUC) metrics, provided a more detailed understanding of the classifier's performance. This iterative process of exploration and refinement is crucial in fine-tuning the predictive capabilities of the model for effective application in PCR prediction.

For regression, initial models showed very optimistic results which did not generalise well with test data. Various approaches were then used to achieve a reasonable bias-variance trade-off.

For the SVR model, outlier handling was particularly impactful, achieved by adjusting extreme values based on the Interquartile Range (IQR) method, which focuses on the difference between the 25th and 75th percentiles. The use of Pearson and ANOVA for feature selection was significant due to SVR's sensitivity to outliers, effectively reducing the feature count from 117 to 32. For dimensionality reduction, both PCA and ISOMAP were evaluated. PCA was chosen because of its high Proportion Variance Explained (PVE) of 0.95 and low Reconstruction Error (RE) of 0.05, whereas ISOMAP had a considerably higher RE of 11.03. Fine-tuning the SVR model included adjustments to parameters such as kernel, flexibility (C), complexity (degree), learning behaviour (gamma), and epsilon, all contributing to enhancing its overall performance.

The impact of PCA does not show a clear benefit, suggesting the data is not linearly separable. Similarly, for the Random forest model, PCA had a negative effect on performance. This increased error from the use of PCA may not be unconnected to how pca works and the nature of the dataset.

For instance, the dataset could contain certain features that show little variance but are crucial to the model's performance. This could cause PCA to rank them lower, leading to loss of information. In addition to this, Random forests are very capable of working well with higher dimensional data which may be lost due to pca.

Furthermore, for XGBoost, the use of Anova thresholds for feature selection and XGBoost's feature importance show a positive impact on the r2 score. It also caused marginal improvements in MAE scores, underscoring the importance of good feature selection. However, as in the other cases, pca does not seem beneficial, suggesting the data is highly linearly inseparable.

To summarise, feature selection and outlier handling generally improved the model while pca (the linear and kernel forms) led to worse performance. To improve the model, further feature selection and outlier handling techniques should be considered, or more data could be collected.

6. CONCLUSION

For classification tasks, the XGBoost model outperformed other models such as SVM, Random Forest and MLP in predicting PCR. Despite implementing oversampling to address the class imbalance, the positive class prediction for PCR remained unsatisfactory, with the best model (XGBoost) only achieving modest performance. Feature selection, particularly using XGBoost feature importance and ANOVA, resulted in marginal improvements

For regression tasks, initial models did not generalize well with the test data. Various approaches were tried to improve the model's performance. The Random Forest model, with outlier handling through interquartile range adjustment and feature selection, showed some promise. However, dimensionality reduction techniques like PCA did not improve the performance of the models significantly, possibly due to the nature of the dataset being linearly inseparable.

In summary, the study demonstrates the potential of machine learning models in predicting PCR and RFS in breast cancer, with XGBoost showing the best results for classification. However, models require further refinement and consideration of the dataset's characteristics for improved performance. The importance of feature selection and outlier handling in developing robust models is emphasized while acknowledging the limitations of certain techniques like PCA in the context of the dataset. Further research with more data and advanced techniques could lead to improved models for predicting breast cancer outcomes.

7. References list

Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2007 Feb 11;2:59-77. PMID: 19458758; PMCID: PMC2675494.

Li J, Zhou Z, Dong J, Fu Y, Li Y, et al. (2021) Predicting breast cancer 5-year survival using machine learning: A systematic review. PLOS ONE 16(4): e0250370. https://doi.org/10.1371/journal.pone.0250370

Zuo, D., Yang, L., Jin, Y. et al. Machine learning-based models for the prediction of breast cancer recurrence risk. BMC Med Inform Decis Mak 23, 276 (2023). https://doi.org/10.1186/s12911-023-02377-z

Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, Rovera F. Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Med. 2020 May;9(9):3234-3243. doi: 10.1002/cam4.2811. Epub 2020 Mar 10. PMID: 32154669; PMCID: PMC7196042.

Task and Weighting	Data pre-processing (10%)	Feature Selection (25%)	ML method development (25%)	Method Evaluation (10%)	Report Writing (30%)
Emaan Bashir	20%	20%	20%	20%	20%
Bushra Jalali	20%	20%	20%	20%	20%
Shahabaldin Mortazavi	20%	20%	20%	20%	20%
Mustafa Mehmood	20%	20%	20%	20%	20%
Kalu Stephen Eke	20%	20%	20%	20%	20%