PREDICTIVE MODELLING OF PATHOLOGICAL COMPLETE RESPONSE CLASSIFICATION AND RELAPSE FREE SURVIVAL REGRESSION IN CANCER PATIENT.

C.H. Boey, *J.H.S. Loo,* *J.J.J. Swee,* *K. Rudra,* *Q.H. Chai*

University of Nottingham
{hfycb3, hcyjl7, hfyjs8, hcykr2, hfyqc1}@nottingham.ac.uk

ABSTRACT

This study explores machine learning methods to enhance the accuracy of predicting Pathological Complete Response (PCR) and Relapse-Free Survival (RFS) in chemotherapy-treated breast cancer patients. Analysing a dataset comprising 10 clinical and 107 magnetic resonance imaging features across 400 patients, the research aims to provide additional prognosis information, aiding doctors in personalised treatment recommendations. The findings contribute to the broader landscape of personalised medicine, facilitating informed decision-making for cancer patients.

Index Terms**—** Machine Learning (ML), Breast Cancer, Pathological Complete Response (PCR), Relapse-Free Survival (RFS), Classification, Regression

1. INTRODUCTION

This study aims to improve PCR and RFS predictions in breast cancer patients using machine learning methods. It analyses a publicly available dataset from The American College of Radiology Imaging Network which consists of clinically measured features and 107 MRI-based features across 400 patients. The results aim to contribute to personalised medicine research, enabling doctors to tailor their recommendations and provide cancer patients with additional prognosis information.

Breast cancer patients worldwide have a low survival rate of around 50%, with a mortality rate of 58%. The disease occurs when abnormally growing cells form tumours, governed by genes in the cell nucleus. Changes in these genes can contribute to the development of breast cancer (Dhillon & Singh, 2020).

Hence, the main research objective is to compare prediction models for Pathological Complete Response (PCR) and Relapse Free Survival (RFS) in breast cancer patients. The goal is to determine which model is the most effective and find its optimal parameter settings. Accurate prediction of PCR and RFS has value in providing guidance for treatment decisions and more precise prognostic information.

2. RELATED WORK

2.1. Regression

The study (Tahmassebi et al., 2019) explores the use of multiparametric magnetic resonance imaging (mpMRI) for early prediction of relapse-free survival (RFS) in breast cancer patients. It extracts 23 features from 38 breast cancer patients and uses eight ML classifiers. The study found that LR performed best in RFS prediction, with an AUC of 0.83. This literature offers insights on regression models for addressing the continuous nature of RFS within data.

(Mihaylov et al., 2019) tests and compares various ML techniques with a specific focus of predicting accurate survival time. The selection of ML models used for this study includes Linear Support Vector Regression (LSVR), Lasso Regression (LSR), Kernel Ridge Regression (KRR), K-neighbourhood regression (KNR), DT, and Multi-layer perceptron regression (MLP). After integrating two distinct datasets of clinical information and genomic profiles revolving around breast cancer, the ML methods were applied post normalisation. Amongst the best performing models for accurately predicting survival time were LSVR, LSR, KRR, KNR, and DT regression.

2.2. Classification

A study explored ML techniques to predict the survival rate of patients with breast cancer. The classification methods showed performance in terms of accuracy ranging from 0.80 to 0.93 across test sets (Tapak et al., 2019). However, since survival prediction is the focus in this biomedical application, a classifier with higher sensitivity is preferred. Therefore, the result in terms of sensitivity showed that AdaBoost achieved a minimum sensitivity of 0.61 while SVM and LDA achieved the maximum sensitivity value of 0.73 each (Tapak et al., 2019). Overall the results indicated that SVM outperformed other ML techniques as it considered both positive and negative likelihood ratios when predicting patient survival.

Hence further investigation was conducted, revealing that a study implemented a voting ensemble comprising six classifiers. The researchers combined different classifiers, like DT, SVM, LR, LDA, Naive Bayes and KNN algorithms to make predictions. The findings indicate that the soft voting classifier achieved the prediction accuracy of 94.03%

when two classifiers (DT and SVM) or three classifiers (DT, SVM and LR) were combined (Kurian & Jyothi, 2023).

Ghasemieh et al., (2023) introduced a Stacking Ensemble Learner (SEL) that utilises ensemble learning to optimise prediction performance. The study demonstrates that incorporating learning into the SEL model overcomes the limitations associated with training classification models which tend to be unstable.

3. CLASSIFICATION METHODOLOGY

3.1. Data Preprocessing for Classification

The importance of each feature was analysed using ANOVA, and it was plotted, as seen in the second graph, and the scale of f-scores were low, this dictates that the presence of missing values introduced noise, affecting the importance scores, resulting in all of the features appearing less significant. Therefore, for handling these missing values, rows that included '999' were dropped, and the importance of each feature was once analysed again using ANOVA, this time the scale of f-scores were way higher, inferring that removing missing values increases the importance scores for a number of features. This was because the analysis was done on a cleaner dataset, providing a more accurate measure of f-scores. To find out the best value of k, the threshold line was increased to where the graph took a significant spike in gradient (elbow point), y =8.5, accumulating to 9 features.

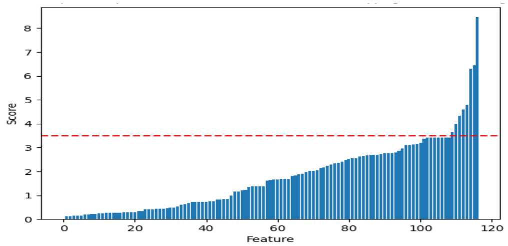

Figure 1: Sorted Feature Importance Plot for ANOVA without dropping rows with missing value

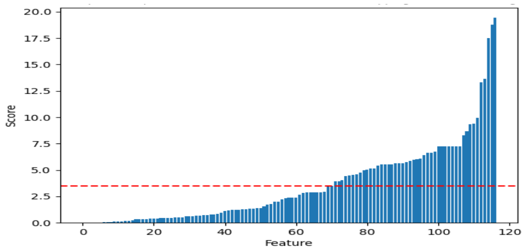

Figure 2: Sorted Feature Importance plot for ANOVA after dropping rows with missing value

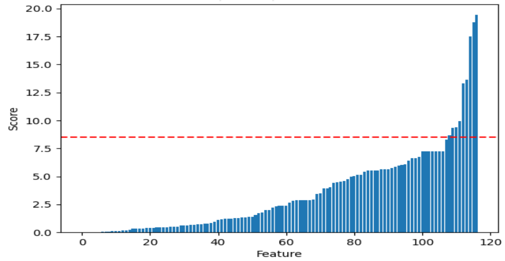

Figure 3: Sorted Feature Importance Plot for ANOVA

Now the data can be split into training and test sets, the test size is set to 0.3, where 30% of the data would be used for testing, and 70% would be used for training the model.

3.2. Model Selection for Classification

Model	Description
Logistic Regression	The logistic model produces a curve that only has values between 0 and 1, this is to estimate the probability of a binary outcome (0/1), depending on the dataset's features.
AdaBoost	A classifier is an ensemble learning approach method for classifying tasks. It functions by combining the forecasts of numerous basic learners to engineer a robust classifier.
SVM	A model for classifying tasks by mapping data to a high-dimensional feature space and categorising the points; this works even with linearly inseparable data.
Voting Classifier	An ensemble learning method that combines the predictions of multiple base models to make a final classification decision. Where hard voting represents the choice of majority class and soft voting, where the probabilities of the class are averaged.

Table 1: Potential Models Selected for PCR Classification Task

4. REGRESSION METHODOLOGY

4.1. Data Preprocessing for Regression

Before feature selection could take place, a better understanding of the dataset was needed. In a glance, it consisted of 119 features which included 1 output feature - RelapseFreeSurvival (outcome). In the dataset, 10 out of 400 rows (2.5%) contained missing data. It was important that this missing data is handled to enhance model accuracy and stability.

After temporarily removing rows with missing data, a feature importance analysis was performed on the dataset using ANOVA. It was found that missing values were not found in highly important features, as shown in the figure below. Features with missing values were highlighted in red to show this finding.

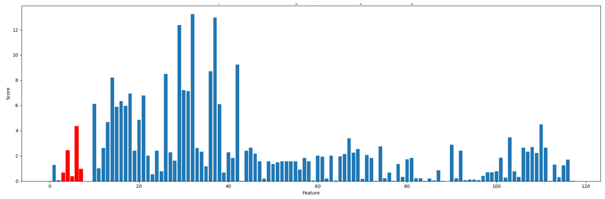

Figure 4: Feature Importance Plot for ANOVA

From this, median imputation was the selected approach to impute the missing data, as it was found that this method had the performance similar to more complex models (Berkelmans *et al.*, 2022), considering that the features containing missing values did not exhibit high importance. Additionally, median imputation proved effective in mitigating the impact of outliers.

As shown in Figure 4, there were many irrelevant features, and therefore feature selection was needed to improve model performance. From this, ANOVA was reused to stay consistent with past findings. Sorting the f-scores of the features after ANOVA, it was decided that features to be selected should have an f-score of over 3.5, based on the distribution of f-scores as shown in Figure 5. This score was chosen based on the significant change of the gradient of f-score values at that point. The number of features found above the threshold was found to be 21 features. Therefore, this number of features was selected to train the models.

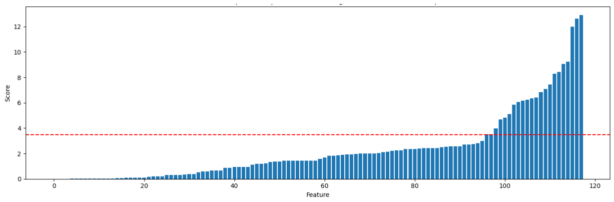

Figure 5: Sorted Feature Importance Plot for ANOVA after mean Imputation

4.2. Model Selection for Regression

To address the prediction task, several regression models were considered. The selected models aim to capture the complex relationships within the training dataset and provide near optimal predictions for relapse-free survival. The following regression models were chosen as shown in Table 2:

Model	Description
Random Forest Regressor	A powerful ensemble method that leverages decision trees to capture intricate patterns in data.
Support Vector Regressor	Utilises support vector machines to find the optimal hyperplane for regression tasks.
LASSO Regressor	Applies L1 regularisation to encourage sparsity in the model coefficients.
Ridge Regressor	Implements L2 regularisation to prevent overfitting by penalising large coefficients.
AdaBoost Regressor	Boosting algorithm that combines multiple weak learners to form a strong predictive model.

Table 2: Potential Models for RFS Regression Task

5. MODEL EVALUATION

The evaluation of classification and regression models were conducted to identify the models with the lowest performance metrics. The process is detailed in Table 3:

Model Evaluation Feature	**Model Type**	
	Classification	**Regression**
Hyperpara-meter Tuning	**Grid Search** - This explores a range of parameter combinations for each model. The best combination of parameters were used for the final evaluation of each model.	
Training and Testing	Models were trained on the training set and evaluated on the testing set to quantify their predictive performance.	
Performance Metric	**Balanced Accuracy and Precision** were chosen as the performance metrics.	**Mean absolute error (MAE)** was chosen as the primary performance metric due to its interpretability and sensitivity to prediction errors.
Predicted Output	**PCR**	**RFS**

Table 3: Model Evaluation Methods Breakdown for Classification and Regression Tasks

6. CLASSIFICATION RESULTS DISCUSSION

Model	Parameters	Balanced Accuracy	Precision
Logistic Regression	max_iter=200000, class_weight={0: 0.51, 1: 0.75}, solver='liblinear', random_state=52, C=5.5	**70.14%**	**76.92%**
AdaBoost Classifier	estimator=LogisticRegression(class_weight={0: 0.425, 1: 0.8}, solver='liblinear', learning_rate=0.1, n_estimators=1000	**73.43%**	**70.59%**
Support Vector Classifier	kernel='linear', C=17.0, class_weight={0: 0.54, 1: 0.805}, decision_function_shape='ovo', probability=True	**69.08%**	**66.67%**
Voting Classifier	estimators=[('SVM', svm_model), ('Log Reg', logreg_model)], voting='hard'	**70.68%**	**83.33%**

Table 4: Classification Balanced Accuracy and Precision Scores with their Optimal Parameters

As seen in Table 4, Adaboost Classifier and Voting Classifier performed the best among the rest of the models in terms of Balanced Accuracy, whilst Logistic Regression and Voting Classifier performed the best in terms of Precision. A pattern can be spotted here, where models that had their estimators set to Logistic Regression, generally performed well for both categories, this is due to the nature of it being simple and robust, making it less prone to overfitting. Apart from that, it is less sensitive to outliers. However, based on their confusion matrices, AdaBoost Classifier produced the most False Positives, which is critical since misclassifying patients might result in them receiving needless treatment. However, a significant number of False Negatives might result in a recommendation against chemotherapy, which would be damaging, especially if the patient may benefit from treatment. Therefore, AdaBoost Classifier was chosen due to the fact that it has a balanced approach to False Positive and False negatives whilst still maintaining a high balanced accuracy and precision.

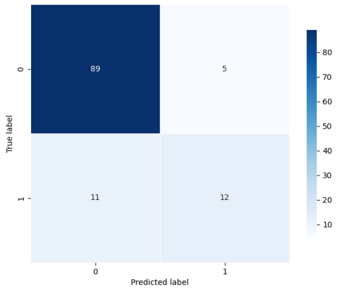

Figure 6: AdaBoost Classifier Confusion Matrix

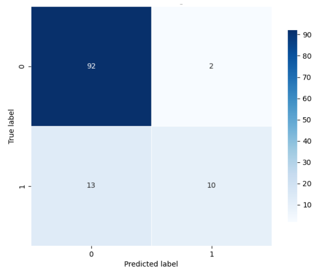

Figure 7: Voting Classifier Confusion Matrix

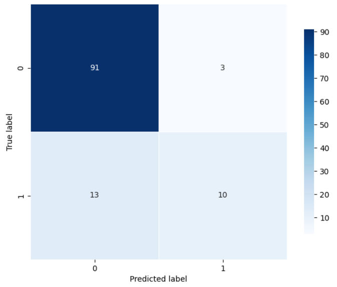

Figure 8: Logistic Regression Classifier Confusion Matrix

7. REGRESSION RESULTS DISCUSSION

In the evaluation phase, the analysis revealed that both the LASSO and Random Forest models emerged as the top-performing candidates during the evaluation process, with the lowest MAEs as seen in Table 5.

Model	**Parameters**	**MAE**
Random Forest	bootstrap=True, max_depth=1000, max_features=sqrt, min_samples_leaf=1, min_samples_split=10, n_estimators=100	**20.43**
SVR	C=0.001, gamma=auto, kernel=sigmoid	21.81
LASSO	alpha=1, max_iter=10000	20.84
Ridge	alpha=1000	21.15
AdaBoost	learning_rate=0.01,loss=linear,n_estimators=100	20.99

Table 5: Regression Methods Mean Absolute Error Scores

However, a closer examination of the residual plots depicting residual values versus predicted values uncovered distinct patterns.

The residual plot for the LASSO model showed clear clusters in a specific range of predicted values as per Figure 10, which showed heteroscedasticity, as the residuals are not evenly spread across the x-axis. This showed that the model's variance of errors were not constant. In contrast, the residual plot for the Random Forest model showed a more dispersed and uniform distribution of residuals across the entire range of predicted values as shown in Figure 9, making its predictions more reliable than LASSO.

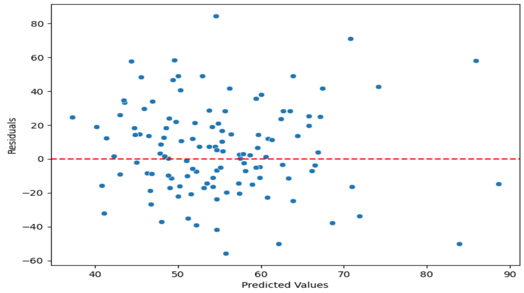

Figure 9: Predicted against Residual Values for Random Forest Regressor

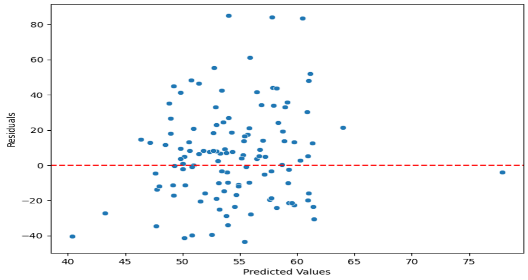

Figure 10: Predicted against Residual Values for Lasso

The lack of clear patterns in the Random Forest residual plots indicates a more robustness in capturing a diverse set of relationships within the data. Based on these observations, the Random Forest regressor is a more versatile and effective choice for predicting relapse-free survival in the given dataset.

8. CONCLUSION

In conclusion, the best model chosen for the prediction of PCR was the AdaBoost Classifier as it was able to strike a balance between correctly classifying negative and positive classes achieving the highest balance accuracy keeping in mind the consequences of a misclassification. The final model chosen for the prediction of RFS values was the RandomForest Regressor as it achieved the smallest mean absolute error while also proving to be a more robust model by being resistant to heteroscedasticity as shown in the residual plots detailed above.

With that, these findings will be able to help doctors in making an informed decision regarding breast cancer patients' need for further treatment and chemotherapy.

9. REFERENCES

Dhillon, A., & Singh, A. (2020). eBreCaP: extreme learning-based model for breast cancer survival prediction. *IET Systems Biology*, *14*(3), 160–169. https://doi.org/10.1049/iet-syb.2019.0087

Tahmassebi, A., Wengert, G. J., Helbich, T. H., Bago-Horvath, Z., Alaei, S., Bartsch, R., Dubsky, P., Baltzer, P., Clauser, P., Kapetas, P., Morris, E. A., Meyer-Baese, A., & Pinker, K. (2019). Impact of Machine Learning With Multiparametric Magnetic Resonance Imaging of the Breast for Early Prediction of Response to Neoadjuvant Chemotherapy and Survival Outcomes in Breast Cancer Patients. *Investigative Radiology*, *54*(2), 110–117. https://doi.org/10.1097/rli.0000000000000518

Mihaylov, I., Nisheva, M., & Vassilev, D. (2019). Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies. Information, 10(3), 93. https://doi.org/10.3390/info10030093

Tapak, L., Shirmohammadi-Khorram, N., Amini, P., Alafchi, B., Hamidi, O., & Poorolajal, J. (2019). Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clinical Epidemiology and Global Health*, *7*(3), 293–299. https://doi.org/10.1016/j.cegh.2018.10.003

Kurian, B., & Jyothi, V. L. (2023). Breast cancer prediction using ensemble voting classifiers in next-generation sequences. *Soft Computing*. https://doi.org/10.1007/s00500-023-08658-z

Ghasemieh, A., Lloyed, A., Bahrami, P., Vajar, P., & Kashef, R. (2023). A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients. *Decision Analytics Journal*, *7*, 100242. https://doi.org/10.1016/j.dajour.2023.100242

Berkelmans, G.F.N. et al. (2022) ‘Population median imputation was noninferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice’, Journal of Clinical Epidemiology, 145, pp. 70–80. Available at: https://doi.org/10.1016/j.jclinepi.2022.01.011.

Group Number : 12

Task and Weighting	**Data pre-processing (10%)**	**Feature Selection (25%)**	**ML method development (25%)**	**Method Evaluation (10%)**	**Report Writing (30%)**	Signature
Boey Chun Hong **(20299272)**	***20%***	***33.34%***	***33.33%***	**26.66%**	***20%***	
Japhia Loo Hyean Shyn **(20129351)**	***20%***	***33.33%***	***33.34%***	**26.67%**	***20%***	
Joshua Swee Jung Jie **(20297500)**	***20%***	***33.33%***	***33.33%***	**26.67%**	***20%***	
Keshaav Suhash Rudra **(20415777)**	***20%***	***0%***	***0%***	**10%**	***20%***	Keshaav
Chai Qin Hui **(20307184)**	***20%***	***0%***	***0%***	**10%**	***20%***	