Comparative Analysis of Genomic and Machine Learning Predictors of Chemotherapy Response in HER2-Negative Breast Cancer

Barbara Popławska, June 2025

Abstract

The study focuses on predicting chemotherapy response in patients with HER2-negative breast cancer, based on gene expression data. The project reimplements the genomic predictor rcb01 developed by Hatzis et al. and compares its performance to machine learning XGBoost models. Analyses were conducted separately for ER-positive and ER-negative subgroups due to their different biological characteristics. The rcb01 predictor showed higher precision and AUC in both groups. XGBoost performed well in the ER− group, but its performance in the ER+ group was limited, likely due to lower chemosensitivity and class imbalance. Gene selection revealed no overlap between ER+ and ER− models, suggesting different molecular drivers of response.

1. Introduction

Breast cancer is one of the most common cancers that affect women worldwide. It is also one of the leading causes of cancer-related deaths. [2] The majority of cases in this group are HER2-negative tumours. HER2 (also known as ERBB2) is one of four receptors of the ERBB receptor family. Those are tyrosine receptors which are anchored in the cell membrane. HER2 plays a key role in transducing intracellular signals by forming dimers with other ERBB receptors, which leads to the activation of various signalling pathways. [3] Activation of HER2 leads to activation of pathways regulating important cellular processes, including proliferation, differentiation, and angiogenesis. This can lead to uncontrolled cell growth. Overexpression of HER2 is observed in some tumours, including breast cancer, and represents a key target for anti-HER2 therapies. [3] In contrast, HER2-negative tumours lack this overexpression, which means that HER2 does not play a major role in driving tumour growth, and therefore HER2-targeted therapies are ineffective in treating this type of cancer. [2, 3]

HER2-negativity is one of the defining criteria for classifying molecular subtypes of breast cancer. Other key markers are the presence or absence of estrogen receptors (ER) and progesterone receptors (PR) in the cancer cells. [2,3] ER and PR receptors are intracellular nuclear receptors that, when bound to the appropriate hormones, can regulate gene expression. The estrogen receptor (ER) belongs to the nuclear receptor family and is activated by estrogens. ER can act via the classical genomic pathway by binding estrogen, and creating the estrogen-ER complex, which binds to DNA sequences, initiating transcription of genes involved in various biological processes, such as proliferation or differentiation. ER can also function via a non-genomic mechanism, where estrogen binds to ER receptors located at the cell membrane leading to the activation of many signalling pathways. ERα isoform is particularly important for mammary gland development and breast cancer biology. Its activity stimulates cell proliferation and may contribute to carcinogenesis.

PR receptors are regulated by estrogens, therefore their expression often depends on ER activity. PR also acts as a transcription factor after binding to progesterone, regulating genes involved in the cell cycle, differentiation and hormonal response. [4] The presence of ER-positive (ER+) and PR-positive (PR+) status is considered as one of the most important prognostic and predictive factors in breast cancer. These tumours are typically biologically less aggressive and respond to hormonal therapies. Therefore, this type of cancer is associated with a better clinical prognosis. ER-negative/PR-negative (ER–/PR–) tumours are often more aggressive and lack eligibility for endocrine therapy [4]

Tumours that do not express all three receptors (ER, PR or HER2) are called triple-negative tumours (TNBC). TNBC usually has a very aggressive clinical course and high metastatic potential. It cannot be treated with either hormone therapy or anti-HER2 targeted therapies. In such cases, the primary treatment option is chemotherapy. However, due to its non-specific mode of action (attacking all dividing cells) chemotherapy is associated with increased toxicity and a less safety profile, which, combined with the aggressive course of the disease for TNBC, leads to higher mortality rate. [2,3]

Chemotherapy based on taxanes (e.g. docetaxel or paclitaxel) and anthracyclines (e.g. doxorubicin) is the standard treatment for HER2– breast cancer, especially in patients with ER–/HER2– tumors. The therapy is used in both neoadjuvant and adjuvant settings. Neoadjuvant treatment involves chemotherapy before surgery to reduce the size of the tumour and assess its sensitivity to treatment. Adjuvant treatment, on the other hand, is a therapy used after surgery to eliminate invisible micro metastases and reduce the risk of disease recurrence. [1, 5, 6]

Taxanes act by stabilizing microtubules, which prevents their depolymerization and lead to stop of mitosis, and apoptosis of the tumour cell. Anthracyclines act through multiple mechanisms, including DNA intercalation or inhibition of topoisomerase II, that causes DNA damage and cell death. Studies have shown that adding anthracycline to taxane-based chemotherapy significantly improves breast cancer treatment outcomes, reducing the risk of recurrence by an average of 14%. [5, 6]

This project is based on the study by Hatzis et al., which aimed to develop genomic predictors of response to neoadjuvant chemotherapy in patients diagnosed with HER2-negative breast cancer. The chemotherapy regimens included taxane and anthracycline-based therapy. As part of the analysis, three genomic predictors were developed: chemo-sensitivity predictor, a chemo-resistance predictor, and the SET index. Based on these predictors, patients were classified with the treatment plans. [1] In my project, I aimed to replicate the chemo-sensitivity predictor and compare its performance with a machine learning-based model to assess which method give the best result in this case.

2. Materials and Methods

2.1 Data

The data used in this project come from the public GSE25066 collection available in Gene Expression Omnibus (GEO) - GSE25066_series_matrix.txt.gz. This is a file combining data from two cohorts: GSE25065 and GSE25055. Data from the first came from multiple clinical trial centres in Peru and Spain, and the second from the USA. Response was assessed at the end of neoadjuvant treatment.

Due to technical convenience, the file was initially processed manually as plain text. Irrelevant metadata such as centre and patient identifiers were removed. This file was processed using file_processing script. From this file I created two separate collections, the first containing clinical data, and the second was a gene expression matrix. Gene expression was measured by the authors using the Affymetrix Human Genome U133A microarray, which included 22,215 probes (I removed the control probes 'AFFX-' from the data during processing). Expression values were already log2-transformed and normalized using the Robust Multi-array Average (RMA) method. The clinical dataset included 20 variables describing patient and tumour characteristics, such as HER2 status, or estrogen receptor (ER) expression. The datasets contained a total of 508 patients, however, after filtering the samples for HER2-negative tumours, I qualified 485 samples for analysis.

2.2 Hatzis model analysis

In the first part of this project, I aimed to reproduce the analyses conducted by Hatzis et al., the reimplementation of the genomic predictor designed to identify patients likely to achieve a pathological complete response (pCR) or minimal residual disease (RCB-I) following neoadjuvant chemotherapy, referred to as the RCB 0/1 class (rcb01). As in the original methodology, I first divided the patients into two groups, ER+ and ER−, and I conducted analyses of these groups separately.

To reconstruct the rcb01 predictor, I used the publicly available gene weight file GSE25066_Genelist_weights.txt.gz created by authors. This file contains coefficients assigned to individual Affymetrix probes that were used in the construction of genomic predictors. For each sample, a predictive score was calculated as a weighted sum of gene expression levels (the sum of products of expression values and their corresponding weights across all relevant probes). The calculated score values were then compared with the actual response to treatment.

To ensure reproducibility, the data were split into training and validation sets using a separate split (the `train_test_split` function from the scikit-learn library), which ensures that the proportions of classes in both sets are preserved. Then, the Hatzis predictor (rcb01_score) is applied to the data, and patients are classified as "responders" or "non-responders" based on the given score thresholds.

In the original study, the predictor rcb01 was binarily classified with a threshold set at 0. However, the exact details of the normalization and preprocessing methods of gene expression data were not described in detail in the original publication. I observed that applying the threshold of 0 did not provide effective separation between responders and non-responders. Therefore, to enable a reliable classification on the available data, the classification thresholds for the ER+ and ER− groups were tuned. These thresholds were selected based on an analysis of the distribution of rcb01_scores in both response groups (responders vs. non-responders), visualized using boxplots, to maximize the separation between them. For the ER− group, the threshold was set at -9, while for the ER+ group it was set at 118. Patients whose rcb01_score were above the threshold were classified as "responders", otherwise as "non-responders". The final step was to evaluate the performance of this predictor on the validation set. The evaluation was done using various metrics and these results are also visualized using ROC curves and boxplots.

2.3 XGBoost model analysis

In the next step, I moved on to the implementation of the XGBoost model. The samples were split into training and test sets in a ratio of 80/20, using the train_test_split function from the scikit-learn library. Stratified splitting was applied to ensure that the proportion of response classes in both sets was preserved, which is crucial in unbalanced data. To focus on the most informative genes, feature selection was performed using the SelectKBest method with the mutual_info_classif scoring function. The top n genes (default 100) with the highest mutual information scores were selected, using only data from the training dataset.

The XGBoost classifier was implemented within an imblearn.pipeline.Pipeline to streamline preprocessing and modeling steps. To assess the impact of class imbalance (large part of the data are non-responders), the analysis was performed in two variants: with SMOTE applied on the training set (generating synthetic samples for the minority class) and without SMOTE. The model was configured with GPU acceleration, evaluated using logloss, and its hyperparameters (max_depth, learning_rate, n_estimators, colsample_bytree and min_child_weight) were optimized via grid search combined with cross-validation. For classification threshold optimization, the F2-score metric (fbeta_score(beta=2)) was used as the objective function. This metric places greater emphasis on recall (sensitivity), which is clinically relevant, because minimizing false negatives reduces the risk of failing to identify patients who would benefit from the chemotherapy.

The performance of the trained models was evaluated on an unseen test set. The threshold for binary classification was determined in a two-step procedure. First, the threshold was determined by maximizing the F2-score on the Precision-Recall curve on the test set. Then, the selected threshold was verified visually by analysing boxplots showing the distribution of predicted probabilities for the actual response classes. At the end threshold for ER- was 0.21, and for ER+ 0.14.

The evaluation results included several metrics. Additionally, the models were also assessed on ROC and Precision-Recall curves. The scikit-learn and scipy.stats libraries were used to calculate statistical metrics, and the matplotlib and seaborn libraries to generate graphs.

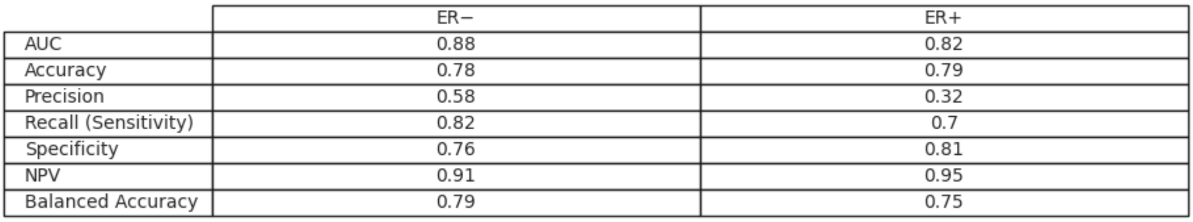
In the final step, genes identified as important predictors in both ER+ and ER− models were analysed. Probe IDs from the Affymetrix microarray were mapped to gene symbols using the annotation file HG-U133A.na36.annot.csv downloaded from Thermo Fisher Scientific page.

3 Results

3.1 Hatzis model analysis

Unfortunately, in their publication, the authors primarily reported performance metrics for the overall genomic test that they developed (rcb01 being one of its components). The only metrics strictly related to rcb01 are PPV and NPV given for the training and validation cohorts, although these results were not separated depending on ER status. In the validation cohort, the authors obtained PPV equal to 42% and NPV equal to 81%.

The reimplementation of the rcb01 predictor conducted in this project allowed for a more detailed evaluation of its performance, separately for the individual groups, ER+ and ER−. The metrics are presented in the table below (Table 1). Differences in predictive values were seen depending on the ER status suggesting that the rcb01 results may vary between biologically distinct tumour types.



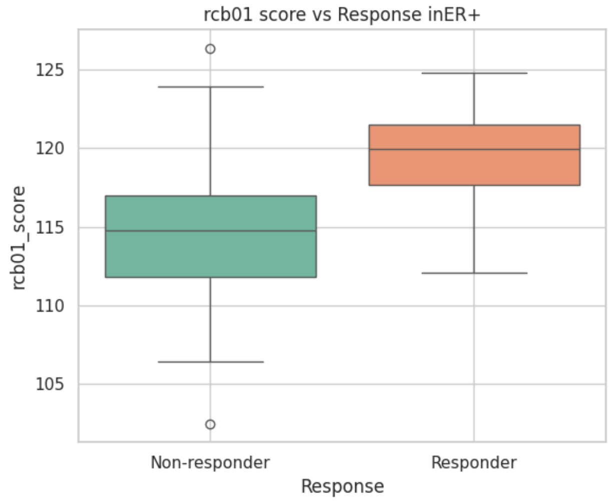
	ER−	ER+
AUC	0.88	0.82
Accuracy	0.78	0.79
Precision	0.58	0.32
Recall (Sensitivity)	0.82	0.7
Specificity	0.76	0.81
NPV	0.91	0.95
Balanced Accuracy	0.79	0.75

Table 1: Performanced metrics for implemented predictor rcb01 from Hatzis et al. for the ER- and ER+ groups.

When comparing the results for both groups, the model showed overall better predictive ability for patients with ER-negative (ER-) breast cancer, because of a higher AUC and better precision. For patients with ER-positive (ER+) breast cancer, the model achieved an AUC of 0.82, with a sensitivity of 0.70 and a specificity of 0.81. Despite the high NPV (0.95), indicating that the model is highly accurate in identifying nonresponders, the precision (PPV) was only 0.32, indicating moderate accuracy in detecting true responders.

This limitation can be partially explained by the classification threshold applied in the ER+ group. The threshold was set to 118 based on the box plot (Figure 1) aiming to classify most women who would benefit from treatment. Selecting the appropriate threshold was proved difficult due to the scores for the two groups being insufficiently separated. Therefore, to make sure that most women receive treatment that can help them, there is a need to qualify for it a large proportion of women who will not benefit from the treatment.

Nevertheless, the model shows a clear separation between the two outcome groups, allowing for meaningful discrimination. The ROC curve (Figure 2) also supports this conclusion. It rises steeply at low false positive rates, indicating high sensitivity in the early classification thresholds. The presence of the curve significantly above the random line (y = x) confirms that the predictor has significant diagnostic value in this subgroup.



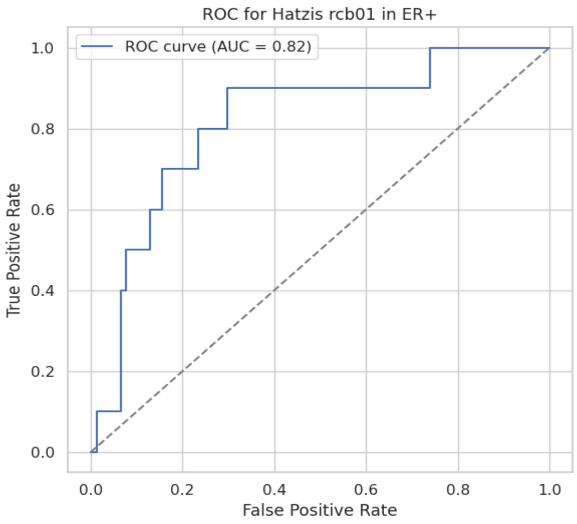
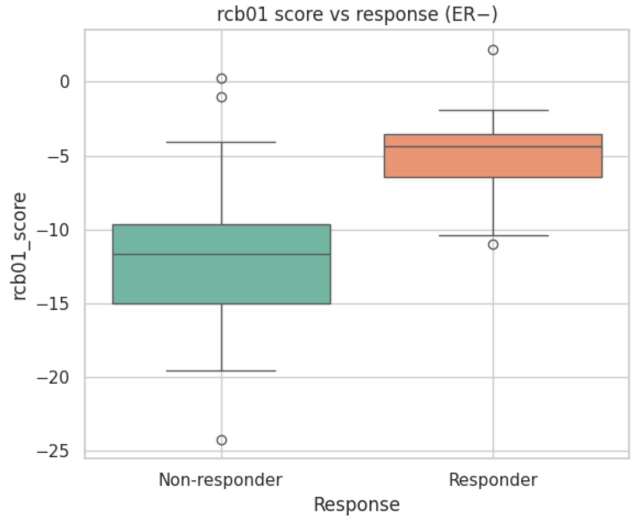


Figure 1: Boxplot showing the distribution of rcb01 scores in ER-positive (ER+) patients, separated by treatment response (responders vs. non-responders).
Figure 2: Receiver Operating Characteristic (ROC) curve illustrating the performance of the rcb01 predictor in the ER-positive (ER+) subgroup.

For patients with ER-negative (ER−) breast cancer, the rcb01 predictor showed even better classification performance: AUC was 0.88, sensitivity 0.82, specificity 0.76, and PPV reached 0.58, which is high in comparison to ER+. The classification threshold, similarly, to the previous one, was set based on the graph to -9 (Figure 3). In contrast to the ER+ subgroup, the rcb01 scores for responders and non-responders in the ER− group were more clearly separated, which enabled more accurate threshold selection.

The results, similarly, to the analysis of the graph, suggest that the model preforms well with both detecting respondents and minimizing the number of false positives. However, it should be noted that the model metrics are significantly influenced by the biology of ER tumors. Although ER- is typically more aggressive and has poorer long-term prognosis, it tend to be more sensitive to chemotherapy. Due to that, the amount of responders is higher in this group, which positively influences the performance of rcb01 classifier. [1,4]



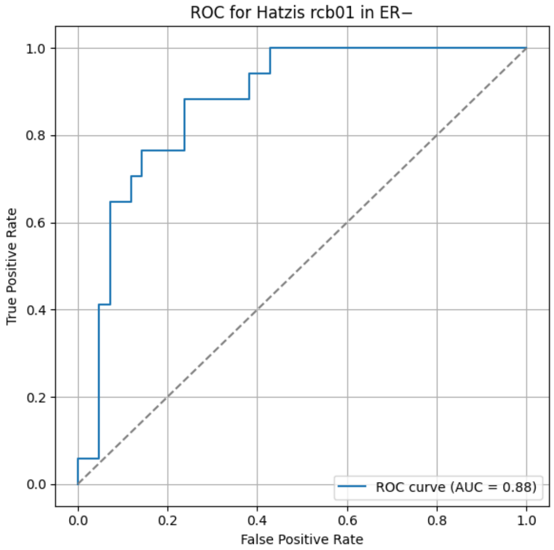


Figure 3: Boxplot showing the distribution of rcb01 scores in ER-negative (ER-) patients, separated by treatment response (responders vs. non-responders).
Figure 4: Receiver Operating Characteristic (ROC) curve illustrating the performance of the rcb01 predictor in the ER-negative (ER-) subgroup.

3.2 XGBoost model analysis

For the ER-negative (ER-) group, an XGBoost model was developed and evaluated. Selecting the most informative genes, was performed using a method based on mutual information, which measures how strongly the expression level of each gene is associated with treatment response. This method can detect both linear and non-linear relationships which is good for biology data. 100 genes were selected from the expression dataset, and used as an input features. Due to class imbalance, since there were fewer responders than non-responders, two models were trained: one using SMOTE to synthetically balance the training set, and one without SMOTE. In both cases, hyperparameter optimization was performed using grid search over 324 parameter combinations with 4-fold cross-validation, using the F2 score as the metric.

In the model trained with SMOTE, the best parameter set included: max_depth = 3, learning_rate = 0.2, n_estimators = 100, colsample_bytree = 1.0, and min_child_weight = 1. This model achieved an AUC of 0.81, an accuracy of 0.66, sensitivity of 0.64, specificity of 0.67, precision of 0.50, NPV of 0.78, and balanced accuracy of 0.65. These results suggest that the model performed well in identifying responders while controlling the false positive rate. In contrast, the version without SMOTE performed worse: the AUC dropped to 0.55, sensitivity to 0.45, precision to 0.42, and balanced accuracy to 0.56. These results highlighted how important class balance can be.

The decision threshold was selected based on the Precision-Recall curve, to maximize the F2 score. The optimal threshold was 0.081, however, visual inspection of the predicted probability (Figure 5) indicated that a threshold, around 0.2 could better distinguish the groups. The ROC curve (Figure 6) confirmed that model has good discrimination performance, rising above the random line (y = x), while the Precision-Recall curve (Figure 7) showed stability in precision across recall values, supporting the use of the model in imbalanced clinical contexts.

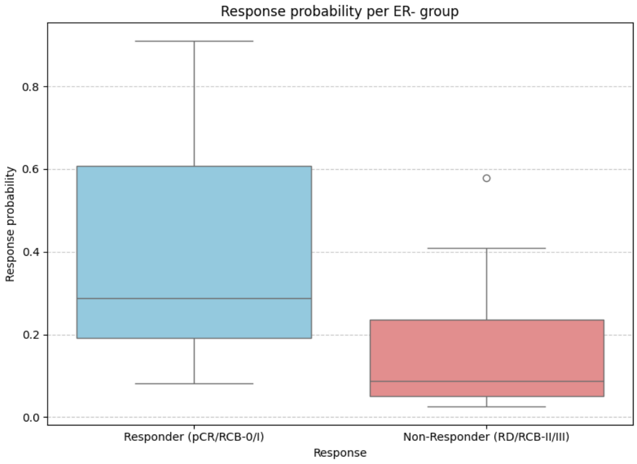
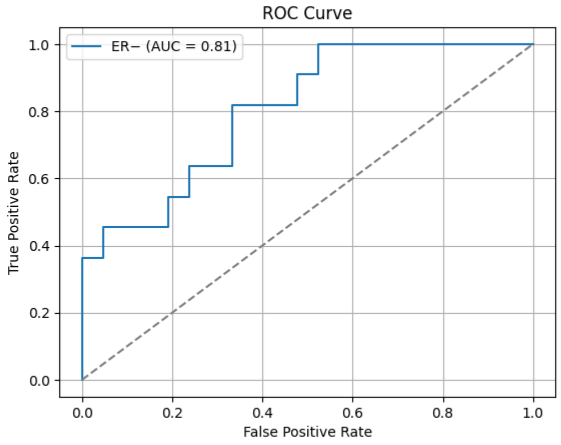


Figure 5: Boxplot of predicted response probabilities for responders and non-responders in the ER- group (XGBoost with SMOTE).



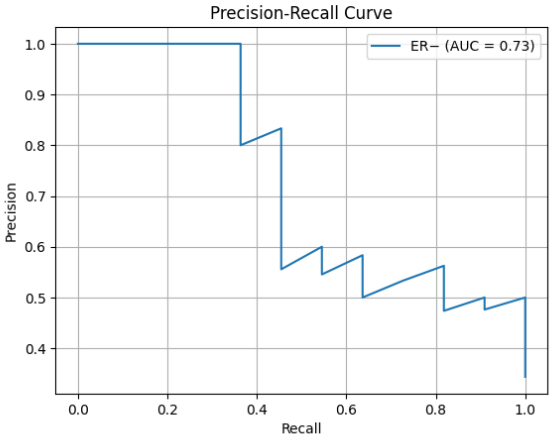


Figure 6: ROC curve for the XGBoost model trained with SMOTE in the ER- group.
Figure 7: Precision-Recall curve for the XGBoost model with SMOTE in the ER- group.

In the ER-positive (ER+) group, the development of the XGBoost model proved more challenging due to a strong class imbalance — the number of non-responders was 251, which was a lot higher than for responders, which was 29. This shows a known clinical characteristic of ER+ tumors, which typically respond poorly to chemotherapy. Thats why two classification strategies were applied. First one, was also mentioned in Hatzis et al. paper and it created a model distinguishing RCB-I from RCB-II/III (without pCR and RD labels), allowing for a more biologically meaningful classification in this context. Second, was a model that included all labels (pCR/ RCB-I vs. RD/ RCB-II/III). In both cases, SMOTE was used to oversample the smaller class in the training set, and feature selection was performed using mutual information to identify the 100 most informative genes.

First model, with excluded labels used the following hyperparameters: colsample_bytree = 0.6, learning_rate = 0.2, max_depth = 3, n_estimators = 100, min_child_weight = 3. Performance on the test set was moderate: AUC = 0.67, accuracy = 0.56, sensitivity = 0.50, and precision = 0.28. These results shows limited classification performance. The boxplot of predicted probabilities (Figure 8) shows very little separation between responders and non-responders, indicating that even though model captures some signal relevant to chemotherapy response it’s not very good.

The ROC curve (Figure 9) confirms moderate classification performance, with the curve lying above the diagonal. However, the precision-recall curve (Figure 10) shows a more limited profile (AUC = 0.50), suggesting that the model struggles to distinguish true positives with high confidence.

The second model achieved: AUC = 0.53, accuracy = 0.73, but only recall = 0.17 and precision = 0.11. It shows a tendency to overpredict the dominant class (non-responders). In comparison the first model provided a better balance between sensitivity and precision, making it better for clinical interpretation, where minimizing false negatives is critical.

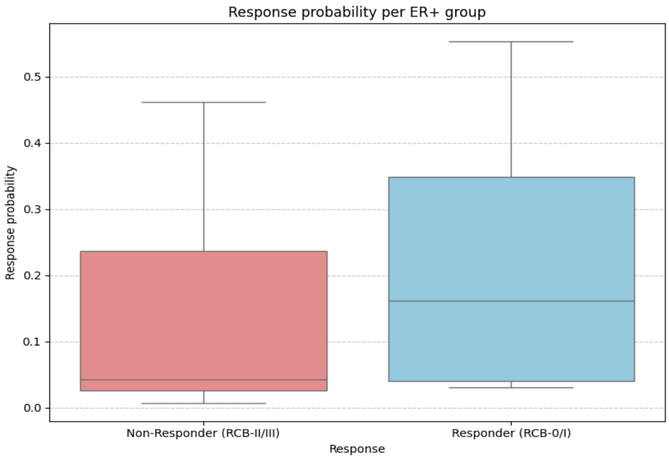
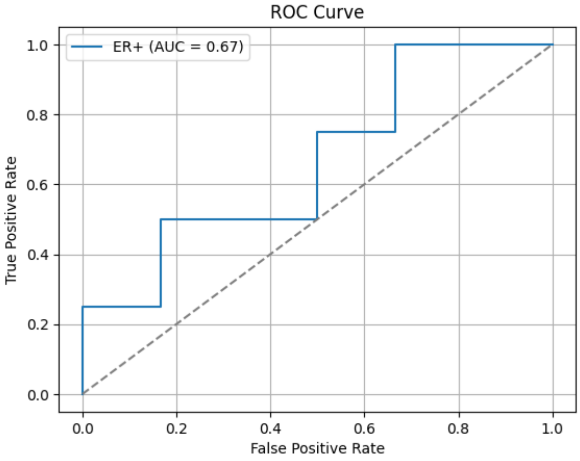


Figure 5: Boxplot of predicted response probabilities for responders and non-responders in the ER+ group for the XGBoost model trained with SMOTE, without data labelled as 'pCR' or 'RD'.



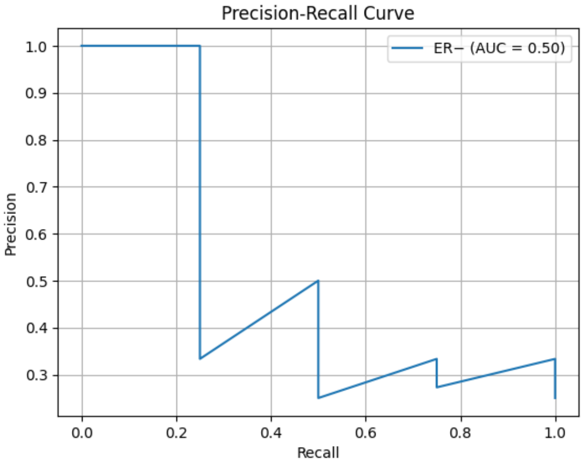
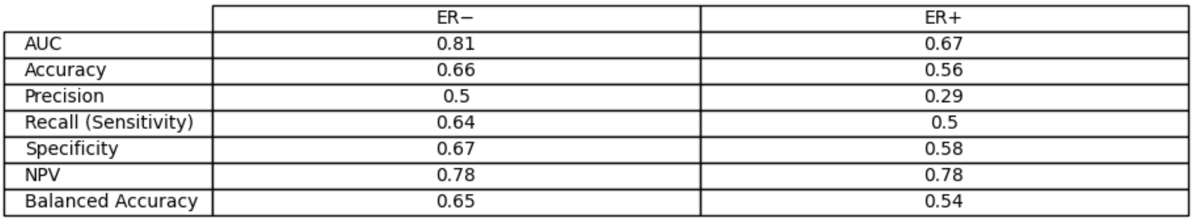


Figure 6: ROC curve for the XGBoost model trained with SMOTE, without data labelled as 'pCR' or 'RD' in the ER+ group.

Figure 7: Precision-Recall curve for the XGBoost model with SMOTE without data labelled as 'pCR' or 'RD' in the ER+ group.

When comparing model performance between ER− and ER+ subgroups, the XGBoost model achieved better results in the ER− group (Table 2). It showed higher AUC (0.81 vs. 0.67), recall (0.64 vs. 0.50), precision (0.50 vs. 0.29), and balanced accuracy (0.65 vs. 0.54). These results suggest that prediction of chemotherapy response based on gene expression is more effective in ER− tumors, probably due to their higher chemosensitivity and genomic response patterns.



	ER−	ER+
AUC	0.81	0.67
Accuracy	0.66	0.56
Precision	0.5	0.29
Recall (Sensitivity)	0.64	0.5
Specificity	0.67	0.58
NPV	0.78	0.78
Balanced Accuracy	0.65	0.54

Table 2: Performance metrics of the XGBoost model, evaluated separately for ER− and ER+ patient subgroups.

As an additional part of the project, I attempted to analyse the genes that were used for both ER- and ER+ models. Interestingly, no overlap was observed between the sets of 100 top-ranked genes selected for the ER^+ and ER^- groups, based on mutual information with treatment response. This shows that the predictive gene sets differ entirely between the two cancer types.

4 Discussion

This study confirmed that both the original rcb01 predictor and machine learning models perform differently depending on cancer ER status. Reimplementation of the rcb01 score showed better performance in the ER- group. It had higher AUC and precision, as well as better separation between responders and non-responders groups. In contrast, performance in the ER+ group was limited, probably due to the lower chemosensitivity of these tumours and class imbalance. XGBoost models also showed this difference: the ER- model achieved strong results (AUC = 0.81, recall = 0.64), while the ER+ models performed worse. This highlights the need for separate modelling in breast cancer, especially when working with biologically different subtypes.

When comparing both approaches, XGBoost model and rcb01 prodictor from Hatzis et al. paper, in ER- group, rcb01 outperformed XGBoost in all key metrics (AUC: 0.88 vs. 0.81; precision: 0.58 vs. 0.50). In ER+, the difference was even bigger (AUC: 0.82 vs. 0.67; precision: 0.32 vs. 0.29). This suggest that while machine learning offers flexibility and adaptability, biological predictors like rcb01, built on annotated gene sets, can still outperform data-driven models if there is limited amount of sample. Probably XGBoost could achieve similar or better results if more data was available, as ML models require enough data. This is especially visible in the ER+ group, where part of data was removed. Model had quite poor results even though when predicting ER- group, model had similar technical capacity but performed way better.

Interestingly, there was no overlap in the top predictive genes between ER+ and ER- models, which suggest that different molecular mechanisms are responsible for each group. The lack of shared genes supports the need for separate models for ER groups.

In the future, to improve ML model performance, especially in ER+ group, additional synthetic data augmentation techniques can be tried to expand the training set. Such approaches may help to overcome the limitations of working with small, imbalanced dataset. From the clinical perspective, it may be beneficial to combine the ML models with biological predictors. It may potentially improve the model generalisation and robustness.

5 References

1. Hatzis, Christos et al. “A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer.” JAMA vol. 305,18 (2011): 1873-81. doi:10.1001/jama.2011.593

2. Harbeck, N., Penault-Llorca, F., Cortes, J. et al. Breast cancer. Nat Rev Dis Primers 5, 66 (2019). https://doi.org/10.1038/s41572-019-0111-2

3. Hsu, J.L., Hung, MC. The role of HER2, EGFR, and other receptor tyrosine kinases in breast cancer. Cancer Metastasis Rev 35, 575–588 (2016). https://doi.org/10.1007/s10555-016-9649-6

4. Deroo BJ, Korach KS. Estrogen receptors and human disease. J Clin Invest. 2006 Mar;116(3):561-70. doi: 10.1172/JCI27987. PMID: 16511588; PMCID: PMC2373424.

5. Anthracycline-containing and taxane-containing chemotherapy for early-stage operable breast cancer: a patient-level meta-analysis of 100 000 women from 86 randomised trials
Braybrooke, Jeremy et al.
The Lancet, Volume 401, Issue 10384, 1277 - 1292

6. Jean-Marc Nabholtz, Alessandro Riva, Taxane/Anthracycline Combinations: Setting a New Standard in Breast Cancer?, The Oncologist, Volume 6, Issue S3, June 2001, Pages 5–12, https://doi.org/10.1634/theoncologist.6-suppl_3-5