Interpretable prediction of breast cancer therapy response using Shapley-Based explanations

Mingxuan Fan, Nikolai Egorov, Zhixin Mao

1 Project

1.1 Scientific Question

Is it possible to predict the response of breast cancer patients to neoadjuvant chemotherapy by integrating multi-omics data and using machine learning methods with Shapley-Based explanations?

1.2 What inspired you to work on this project?

Breast cancer has always been a serious and complex problem. Moreover, the treatment of breast cancer, whether it's chemotherapy or targeted therapies, often have bad responses. Therefore, we find it meaningful to predict the outcome of therapy, in order to better customize therapy methods for patients and increase their survival rates. The method we consider uses multi-omics data that combines different features, which would provide a more robust and comprehensive prediction.

1.3 What diseases/organisms/tissues would you study?

We study breast cancer, specifically the tumor ecosystems of patients that undergo chemotherapy or targeted therapy.

1.4 What types of data will you analyze?

We will analyze DNA and RNA sequencing data—including somatic mutations, copy-number profiles (ASCAT-derived purity and chromosomal instability), and transcriptomic (RNA-seq) expression—and complement these molecular layers with quantitative digital‑pathology features.

1.5 What computational/statistical methods do you plan to use to analyze the data?

Predictive Modeling: Logistic Regression/Random Forest

Model Interpretability: Shapley Additive Explanations(SHAP)

Statistical analysis: AUC, Fisher's exact Test, PCA

1.6 List of related work

- Sharma, A., Debik, J., Naume, B. et al. Comprehensive multi-omics analysis of breast cancer reveals distinct long-term prognostic subtypes. Oncogenesis 13, 22 (2024). https://doi.org/10.1038/s41389-024-00521-6
- Yao, K., Tong, CY. & Cheng, C. A framework to predict the applicability of Oncotype DX, MammaPrint, and E2F4 gene signatures for improving breast cancer prognostic prediction. Sci Rep 12, 2211 (2022). https://doi.org/10.1038/s41598-022-06230-7
- Li, J., Li, S., Zhang, D. et al. Machine learning-based integration develops relapse related signature for predicting prognosis and indicating immune microenvironment infiltration in breast cancer. Sci Rep 15, 19773 (2025). https://doi.org/10.1038/s41598-025-03423-8

1.7 How does your project differ from related work?

Although the authors stated that feature selection had already been performed, we will re-analyze the raw data and redefine the features for our models. The feature selection workflow will include quality control steps, such as removing columns with a high proportion of missing values. Additionally, we may perform principal component analysis (PCA) on numerical features to identify those that explain the most variance. For categorical data, we will apply Fisher’s exact test—for example, to examine which mutations are enriched in pCR patients versus non-pCR patients. To reduce the risk of overfitting, Lasso regularization may also be applied during model training.

Furthermore, we use explainable AI to capture patient-specific features. For any individual sample, the model decomposes the predicted probability of pCR into Shapley contributions from each feature. Such personalized explanations can help clinicians understand “why” the model recommends a given treatment path for that patient, building trust and aiding multidisciplinary team discussions.

2 Data

2.1 Where will you get your data from?

Data is available on github:

1) cclab-brca/neoadjuvant-therapy-response-predictor
2) micrisor/NAT-ML

3) Supplementary materials presented in the study.

2.2 How many samples are available?

- Patients enrolled: 180
- Pre-treatment biopsies collected: 168
- DNA and RNA profiled:
    - Shallow whole-genome sequencing: 168 samples
    - Whole-exome sequencing: 168 samples
    - RNA sequencing: 162 samples
- Digitized pathology slides: 166 cases
- Response assessed (RCB classification): 161 cases

In summary, the study used 168 tumour biopsy samples for multi-platform (multi-omics) profiling. For final response analysis (RCB classification), 161 cases were included.

2.3 Is the data multi-omics?

Yes, it contains information on somatic mutations (WES), transcriptomics (RNA-seq) and genomics(CNVs) .

2.4 In what format is the data available?

Copy number variation data calculated by the ASCAT tool: .tsv

Tumor purity data estimated by ASCAT: .tsv

The genomic fragment information output by ASCAT includes copy number segmentation:.tsv

Original count data of RNA sequencing related to digital pathology: .tsv

Immune-related analysis results: .tsv

Signature data of mutation characteristics: .tsv

2.5 Do you expect to spend a significant amount of time on pre-processing?

No, all the data provided on the Github is processed.

2.6 Are you sure there are no restrictions on data access?

The original data such as RNA-seq has been uploaded to EGA(European Genome-phenome Archive) with the number: EGAS00001004582, which is under controlled access. But the de-identified data is public on Github.

3 Summary of Project Plan

We aim to analyze the relationship between multi-omics tumor characteristics and treatment response of breast cancer, specifically focusing on the binary outcome of pCR and residual disease after neoadjuvant chemotherapy (complete remission vs. residual lesions). To achieve this, we will integrate genomic, transcriptomic, neoantigen, and digital pathology features into a unified machine-learning framework to predict the effects of cancer treatment.

We will use the compressed and pre-processed dataset provided by the authors of the original paper for training and validating our models. We will integrate genomic (whole-exome sequencing-based mutation and copy number features, HRD scores), transcriptomic (RNA-seq derived gene expression profiles, iC10 subtypes), neoantigen landscape (HLA typing, HLA LOH, neoantigen load), and digital pathology features (immune infiltration, lymphocyte density) into a unified machine learning framework for response prediction.Our focus will be on three types of classifiers: random forests, SVM and logistic regression. To optimize model parameters, we will perform five-fold cross-validation on the discovery cohort. Performance will be evaluated using metrics such as AUC, accuracy, sensitivity, and specificity for predicting pathological complete response (pCR).

For all of the three machine learning methods, we will compute SHAP values to quantify the contribution of each feature to individual predictions. This will enable the identification of key multi-omic drivers—such as high tumor mutational burden (TMB), lymphocyte density, and specific gene-expression patterns—and generate patient-level “feature reports” explaining why a given biopsy is predicted to result in pCR or residual disease.

By making model decisions interpretable at the individual level, this approach aims to support more informed clinical decisions and ultimately improve treatment quality in breast cancer and possibly apply to other cancer types.

4 Timeline

Week1: Data collection and cleaning finished, data exploration started
Week2: Exploratory analysis finished, some modelling/visualizations
Week3: Data modelling & Machine Learning (random forest, logistic regression, SVM)
Week4: Multimodal data integration, more modelling & ML, literature search for annotation with biological knowledge (IC10 Classification, HRD Score)
Week5: Shapley values interpretation
Week6: Wrap up all analyses, prepare for presentation (model benchmarking, interpretation plots)