https://doi.org/10.29289/259453942024V34S1004

# Machine-learning model to predict resistance to neoadjuvant chemotherapy in breast cancer

Matheus Gibeke Siqueira Dalmolin[1], Martina Lichtenfels[2], Marcelo Augusto Costa Fernandes[1],
Caroline Brunetto de Farias[3]

[1]Universidade Federal do Rio Grande do Norte, Department of Computer Engineering and Automation.
[2]Translational Research, Ziel Biosciences.
[3]Ziel Biosciences.

**Objective:** The aim of this study was to use a machine learning algorithm to identify biomarkers of resistance to neoadjuvant chemotherapy (NACT) in breast cancer (BC). **Methodology:** We evaluated microarray gene expression data of BC samples before NACT from public datasets of the Gene Expression Omnibus database. We performed differential expression analyses comparing patients who presented residual disease (RD) vs pathological complete response (pCR) to NACT in each dataset and employed a machine learning algorithm to classify genes involved in NACT resistance. Differentially expressed genes with an adjusted p-value less than 0.01 and a logFC greater than 1 or less than -1, identified in more than one analysis, were selected as potentially relevant to tumor resistance. We implemented the XGBoost algorithm, a machine--learning technique based on trees, and used the SHAP method to interpret the prediction results of the machine-learning model. **Results:** The selected datasets were GSE25066, GSE20271, and GSE20194, containing 472, 173, and 267 samples. These datasets present heterogeneous data, with different subtypes of BC and treatments used in the NACT (FACT×FECT and paclitaxel×docetaxel). Our differential expression analysis identified 39 genes for the dataset GSE25066, 28 for GSE20271, and 43 for GSE20194. The XGBoost algorithm achieved an average accuracy of 95% in classifying samples into pCR and RD. Through the SHAP, we identified the genes that most contributed to the prediction of resistance to NACT in the algorithm and found 229 genes in GSE25066, 84 in GSE20271, and 154 in GSE20194. Despite the high heterogeneity of the datasets and methodologies, we identified five genes that were common to both methods. **Conclusion:** These findings contribute to a better understanding of the mechanisms involved in intrinsic tumor resistance to NACT, highlighting the capacity of the XGBoost algorithm in predicting BC resistance, and allowing the development of personalized therapeutic strategies.

**Keywords:** breast neoplasms; neoadjuvant chemotherapy; drug resistance; gene expression; algorithms.