ABSTRACT

Precision medicine for breast cancer requires accurate predictive models to address the heterogeneity of patient responses to standard therapies. One potential approach to capture this variability is the use of molecular profiling. Proteomic data has demonstrated high predictive value, determining the most optimal machine learning architecture to leverage it remains a challenge. This study aims to optimize drug response prediction architectures by comparing the effectiveness of classical and modern drug feature representation methods. A comparative evaluation was conducted on ten model architectures, combining two feature extraction representation methods (Morgan Fingerprint and Graph Neural Network) with five machine learning algorithms (LightGBM, XGBoost, Random Forest, SVM, and MLP). The dataset comprised 47 breast cancer cell lines and 239 drugs, validated using the Group Shuffle Split (80:20) method to prevent data leakage. Model performance was assessed using the Area Under the Precision-Recall Curve (AUPRC) as the primary metric. The results demonstrated that boosting-based algorithms (LightGBM and XGBoost) consistently outperformed others. Graph-based feature representation (GNN) proved to provide richer topological information compared to Morgan Fingerprints in tree-based models. The most optimal model architecture identified was the combination of GNN features with the XGBoost algorithm, achieving the highest AUPRC score of 0.9844. These findings confirm that integrating graph-based molecular representations with gradient boosting algorithms is an effective strategy for enhancing pharmacogenomic modeling accuracy on proteomic data.

Keywords: Drug Response Prediction, Breast Cancer, Proteomic, Pharmacogenomic, Group Shuffle Split, Machine Learning, Graph Neural Network.