ABSTRACT

Lung cancer remains a major global health problem and is one of the leading causes of cancer-related mortality. One contributing factor to this condition is the application of uniform therapeutic strategies for all patients. Precision medicine is therefore required to address heterogeneity in patient responses, improve treatment effectiveness, and reduce adverse effects. Integrating proteomic data with drug molecular structures has emerged as a promising strategy for predicting drug response, while machine learning algorithms are needed to optimize predictive performance. This study aims to identify the best classification model for drug response prediction using three machine learning algorithms (Random Forest, XGBoost, and Multi-Layer Perceptron (MLP)) which are well suited for high-dimensional data. These algorithms are combined with two molecular representations (Morgan Fingerprints and Graph Neural Networks (GNN)), that transform chemical structures into numerical vectors. The GNN-based approach produces more informative features with a very low sparsity level of 8.55% compared to Morgan Fingerprints. The best-performing model is achieved by combining XGBoost with GNN representations, yielding an Area Under the Precision–Recall Curve (AUPRC) of 95.57%, and is further evaluated using a confusion matrix. The results demonstrate that integrating proteomic data with GNN-based molecular representations and the XGBoost algorithm provides more optimal drug response predictions and supports the development of precision medicine.

Keywords: lung cancer, precision medicine, proteomic, drug response, machine learning, molecular representation, Graph Neural Networks, XGBoost.