POSTER - RNA AND TRANSCRIPTOMICS

EXPLAINABLE AI APPLIED IN MACHINE LEARNING FOR IDENTIFYING INFLUENTIAL GENES IN CANCER CLASSIFICATION VIA RNA-SEQ GENE EXPRESSION DATA

Matheus Dalmolin (matheusdalmolinrs@gmail.com)

Karolayne Santos De Azevedo (karolayneazevsantos@gmail.com)

Luísa Christina De Souza (luisa.souza.103@ufrn.edu.br)

Martina Lichtenfels (martinalichtenfels@hotmail.com)

Caroline Brunetto De Farias (carolbfarias@gmail.com)

Marcelo A. C. Fernandes (mfernandes@dca.ufrn.br)

Explainable AI (XAI) in machine learning (ML) algorithms can be used as a strategy for attribute selection. XAI techniques include model interpretation, explanation generation, and interactive visualization. One specific technique mentioned is SHAP, which stands for Shapley Additive Explanations. SHAP is a game-theoretic approach to explain the output of any machine learning model. In this study, we trained different ML models using RNA-Seq gene expression data to classify the five most common types of cancer in women. In addition, we evaluated whether the values of SHAP (SHAPley Additive exPlanations) can differentiate between tumor types.

The RNA-Seq data was obtained from The Cancer Genome Atlas (TCGA). We employed the Under Sampling technique to balance the number of samples for each tumor type. This approach involves undersampling the majority sets based

on the minority set, in this case, the ovary, which had 421 samples. This resulted in a total of 2,105 samples. For building the machine learning models, we selected the Decision Tree (DT), Random Forest (RF), eXtreme Gradient Boosting (XGB), and Support Vector Machine (SVM) algorithms. We used cross-validation to train and validate the four ML models. Then, we applied the SHAP method, which can identify the most relevant features in the decision-making of each ML model.

The RF, DT, XGB, and SVM models achieved accuracies of 99.82%, 98.69%, 99.37%, and 96.73%, respectively. After training the models, we used the SHAP method to obtain the global SHAP values matrix, except for SVM. Then, we performed feature selection, considering only those features that had a contribution equal to or greater than 0.1% to the model's output. The original gene expression matrix, which initially contained 21,481 genes, was reduced to 119 in the RF model, 10 in the DT model, and 81 in the XGB model, resulting in a total of 172 unique genes. This new list, containing 0.8% of the original data, was applied to the ML algorithms. RF maintained the same accuracy, while DT and XGB achieved an increase of 0.18% and 0.48%, respectively, and SVM showed a decrease of 8.25% in accuracy with only 172 genes.

We demonstrated that genes with the highest SHAP values are related to the classification. This approach allowed us to obtain information about the impact of genes on the model and the relationship between the degree of contribution (positive or negative) and the expression value of each gene. We also observed that it is possible to identify highly specific genes for all tumor types from any of the three models analyzed. Our results indicate that the SHAP method can be used as a feature selector in gene expression data to be applied in ML models. XAI can be used to search for potential relevant biomarkers for complex diseases such as cancers.