Evaluation of Machine Learning Methods That Identify Colorectal Lesions with Microbiota-Associated Biomarkers

B. D. Topçuoğlu[1], J. Wiens[2], M. T. Ruffin[3], P. D. Schloss[1]

[1] Department of Microbiology and Immunology, University of Michigan
[2] Department of Computer Science and Engineering, University or Michigan
[3] Department of Family Medicine and Community Medicine, Penn State Hershey Medical Center

INTRODUCTION

- In the microbiome field, use of machine learning (ML) is often flawed. There is a lack of clarity and consistency on which methods are used and how these methods are implemented.
- This is also the case for studies that used microbiome data to detect colorectal cancer (CRC) lesions.
- To showcase a reliable ML pipeline and to shed light on how ML model selection can affect modeling results, we performed an empirical analysis comparing 7 different ML models using the same CRC dataset.

METHODS

1. Dataset: 490 patients (261 CRC, 229 healthy)
   - Fecal 16S rRNA sequences are features.
   - Colonoscopy results are labels (SRN or not)
2. Model: Binary prediction task with L2-regularized logistic regression, L1 and L2 support vector machines (SVM) with linear and radial basis function kernels, a decision tree, random forest and extreme gradient boosted decision tree (XGBoost).

RESULTS

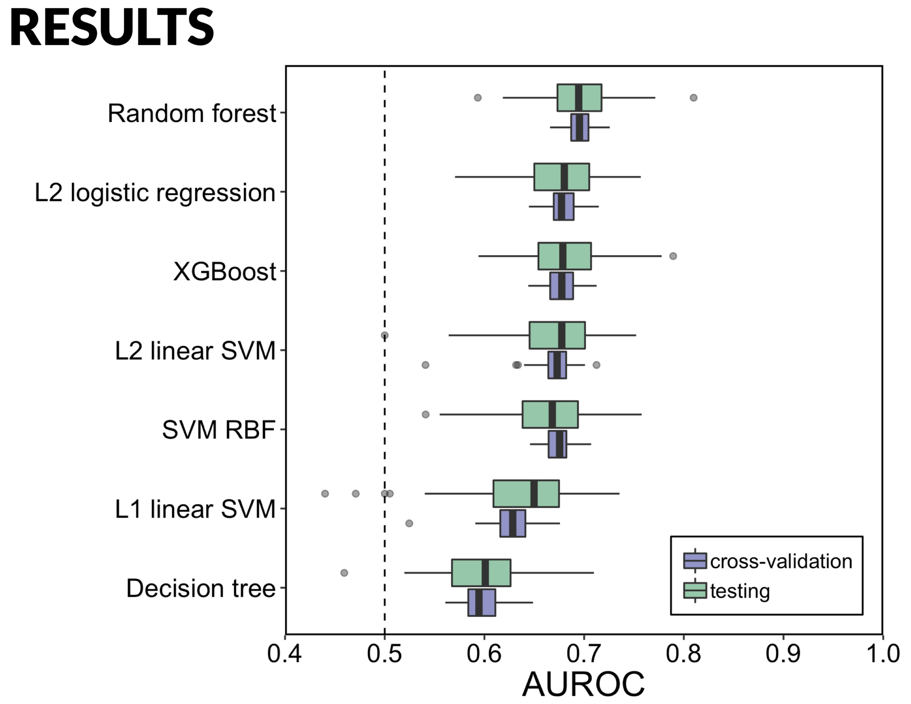

Figure 1 Generalization and classification performance of ML models using AUROC values of all cross validation and testing performances. The median AUROC for diagnosing individuals with CRC lesions using bacterial abundances was higher than chance (depicted by horizontal line at 0.50) for all the ML models. Discriminative performance of random forest model was higher than other ML models.

❖ For **microbiome data**, we need to select **machine learning models** that address the **goal of the study** because this selection will decide **model accuracy**, **complexity**, **interpretability** and **computational efficiency**.

❖ **Random forest** model was **best** at **detecting colorectal lesions** using 16S rRNA sequences but it was **slow and more complex**. Despite the **simplicity**, the **L_2-regularized logistic regression** followed random forest. It was also **fast** and **easy to interpret**.

Take a picture to go to my Github repository.

@Begum_Topcuoglu

Additional Information

Machine Learning Pipeline

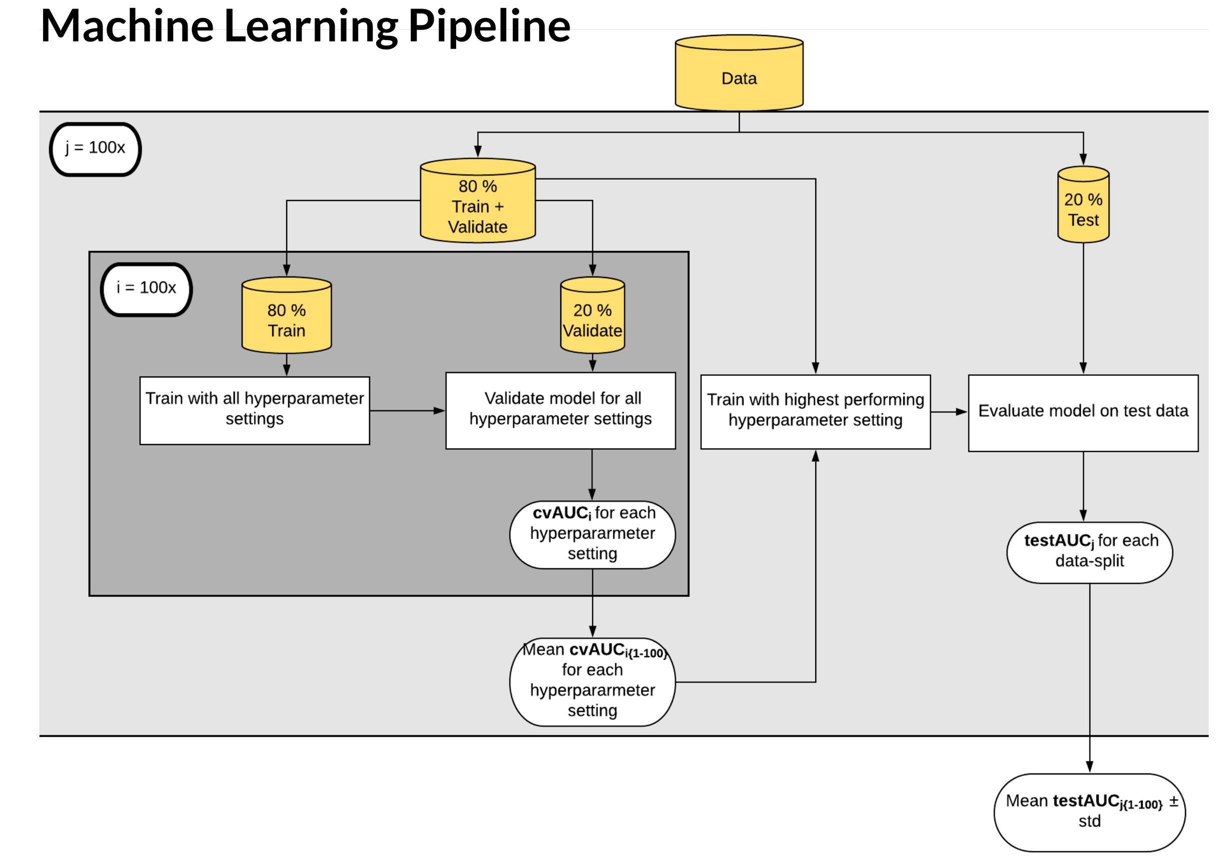

Figure 2. Machine learning pipeline showing predictive model training and evaluation flowchart. We split the data stratified to maintain the overall label distribution, performed five-fold cross-validation on the training data to select the best hyperparameter setting and then using these hyperparameters to train all of the training data. The model was evaluated on a held-out set of data.

Feature Importance

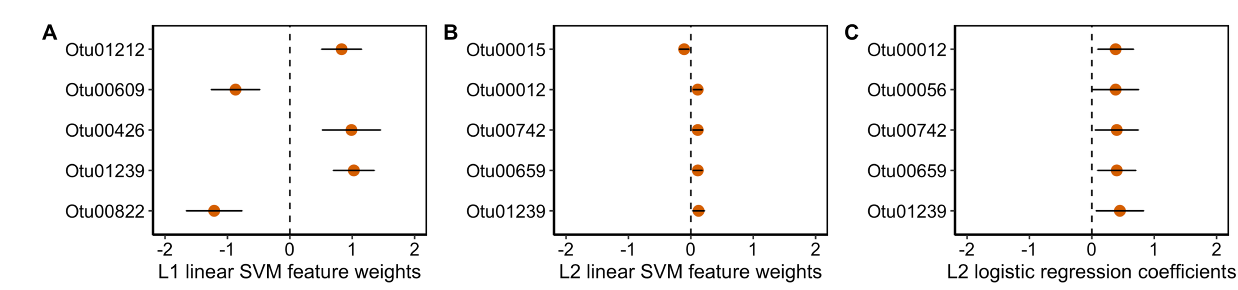

Figure 3. Interpretation of the linear ML models. (A) L2 logistic regression coefficients (B) L1 SVM with linear kernel feature weights (C) L2 SVM with linear kernel feature weights. The means weights and coefficients of the most important 5 OTUs are shown here with the standard deviation over 100 data-splits. Similar OTUs had the largest impact on the predictive performance.

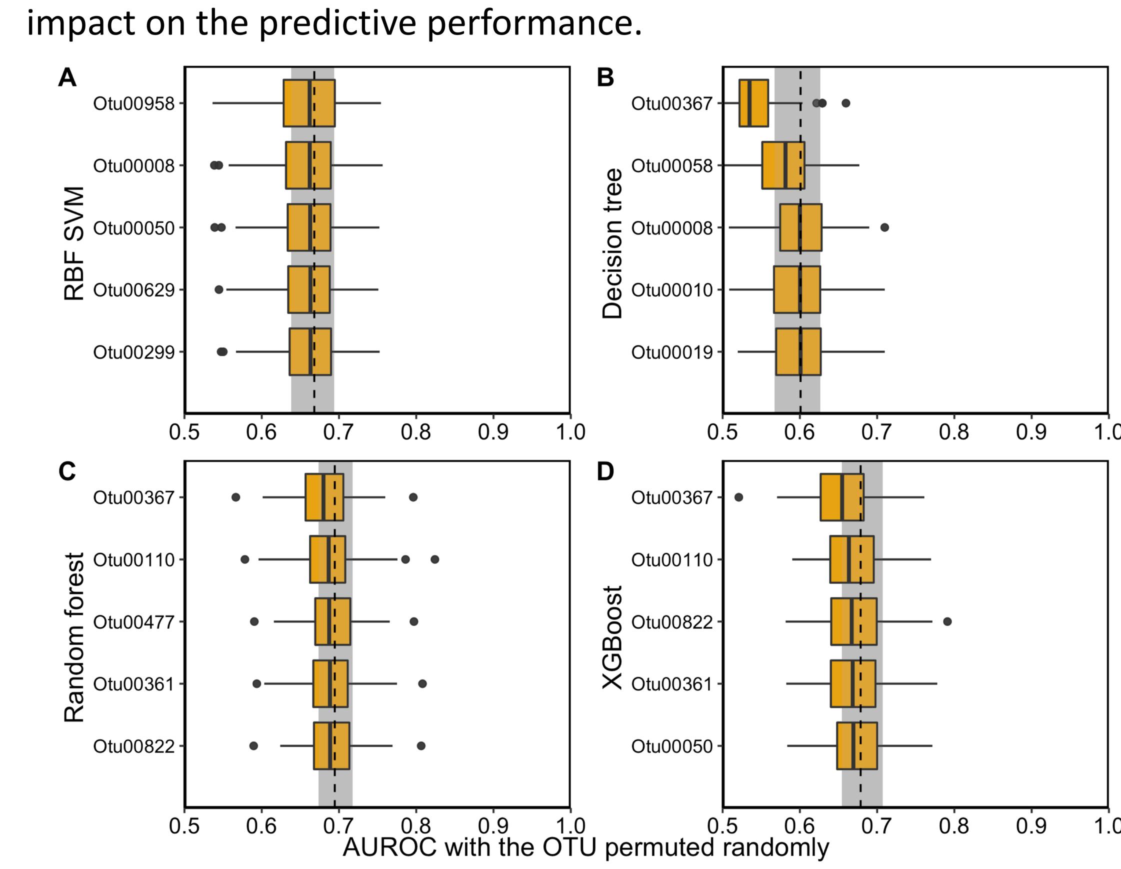

Figure 4. Explanation of the non-linear ML models. (A) SVM with radial basis kernel (B) decision tree (C) random forest (D) XGboost feaure importances were explained using permutation importance using held-out test set. The gray rectangle and the dashed line show the IQR range and median of the base testing AUROC without any permutation performed. A *Peptostreptococcus* species had the largest impact.

Computational Efficiency

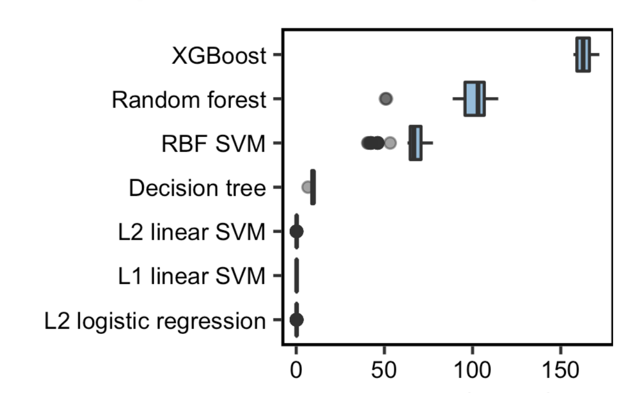

Figure 5. Computational efficiency of seven ML models. The wall-times for training and testing of each data-split showed the differences in computational efficiency of the seven models.