Evaluation of machine learning methods for 16S rRNA gene data

Running title: Machine learning methods in microbiome studies

Begüm D. Topçuoğlu[1], Jenna Wiens[2], Mack Ruffin[3], Patrick D. Schloss[1†]

† To whom correspondence should be addressed: pschloss@umich.edu

1. Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109

2. Department of Computer Science and Engineering, University or Michigan, Ann Arbor, MI 49109

3. Department of Family Medicine and Community Medicine, Penn State Hershey Medical Center, Hershey, PA

Abstract

Introduction

Advances in sequencing technology and decreasing costs of generating 16S rRNA gene sequences have allowed rapid exploration of human associated microbiome and its health implications. Currently, the human microbiome field is growing at an unprecedented rate and as a result, there is an increasing demand for methods that identify associations between members of the microbiome and human health. However, this is difficult as human associated microbial communities are remarkably complex and uneven. It is unlikely that a single species can explain a disease. Instead, subsets of those communities, in relation to one another and to their host, account for the differences in health outcomes. Machine learning (ML) methods are effective at recognizing and highlighting patterns in complex microbial datasets. Therefore, researchers have started to explore the utility of ML models that use microbiota associated biomarkers to predict human health and to understand the microbial ecology of diseases such as liver cirrhosis, colorectal cancer, inflammatory bowel diseases (IBD), obesity, and type 2 diabetes (1–11). However, currently the field's use of ML lacks clarity and consistency on which methods are used and how these methods are implemented. Most notably, there are misleading practices of using "leaky" ML pipelines where models are tested on training data or reporting the best outcome of different iterations of cross-validation. Moreover, there is a lack of discussion on why a particular ML model is utilized. Recently, there is a trend towards using more complex ML models such as random forest, extreme gradient boosting and neural networks without a discussion on if and how much model interpretibility is necessary for the study (11–14). The lack of transparency on modeling methodology and model selection negatively impact model reproducibility and reliability. We need to strive toward better machine learning practices by (1) implementing consistent and reliable machine learning pipelines; (2) selecting ML models that reflect the goal of the study as it will inform our expectations of model accuracy, complexity, interpretibility and computational efficiency.

To showcase reliable ML pipeline and to shed light on how much ML model selection can affect modeling results, we performed an empirical analysis comparing several different ML models using the same dataset and with a robust ML pipeline. We used a previously published colorectal cancer (CRC) study (3) which had fecal 16S rRNA gene sequences from 490 patients. We built ML models

using fecal 16S rRNA gene sequences to predict patients with normal colons or patients with colonic tumors which are called screen relevant neoplasias (SRN). The study had 261 normal and 229 SRN samples. We established modeling pipelines for L2-regularized logistic regression, L1 and L2 support vector machines (SVM) with linear and radial basis function kernels, a decision tree, random forest and XGBoost. Our ML pipeline utilized held-out test data to evaluate generalization and prediction performance of each ML model. The median test AUROC varied from 0.601 to 0.695. Random forest had the highest median AUROC for detecting SRN. Despite the simplicity, the L2-regularized logistic regression followed random forest in performance. In terms of computational efficiency, L2 logistic regression trained the fastest (0.202 hours, std ± 0.019), while XGBoost took the longest (162.843 hours, std ± 3.986). We found that median cross-validation and testing AUROC varied only 0.021, which highlights the importance of a seperate held-out test set and consistent preprocessing of the data prior to evaluation. Aside from evaluating generalization and classification performances for each of these models, this study established standards for modeling pipelines of microbiome-associated machine learning models.

Results

Model selection and pipeline construction

We used a cohort of 490 patients with 261 cases of SRN. For each patient, we had 6920 features (fecal bacterial abundances) and a two-class label that defines their colonic health (SRN or normal). All the cases were independently labeled through colonoscopies. We established modeling pipelines for a 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 XGBoost to emphasize the differences in model accuracy, complexity, interpretibility and computational efficiency due to model selection. We randomly split the data into training/validation and test sets so that the training/validation set consisted of 80% of the full dataset while the test set was composed of the remaining data [Figure 1]. Since the cases are not uniformly represented in the data, the initial data-split was stratified to maintain the overall label distribution on both the training/validation and test sets. Training/validation set consisted of 393 patients (209 SRN), while

the test set was composed of 97 patients (52 SRN). The training/validation data was used for training purposes and validation of hyperparameter selection, and the test set was used for evaluation purposes. Validation of hyperparameter selection was performed using 100 randomizations of five-fold cross-validation on the training/validation set [Figure 1]. Similar to the initial data-split, five-fold cross-validation was also stratified to maintain the overall label distribution on the training and validation sets. We validated the cross-validation performances of each hyperparameter setting over the 100 randomizations and selected the best performing hyper-parameter setting to train the full training/validation dataset [Figures S1 and S2]. We then used the held-out test set to evaluate the prediction performance of each ML model. The data-split, hyperparameter selection, training and testing steps were repeated 100 times to get a reliable and robust reading of model prediction performance [Figure 1].

Random forest has the highest prediction performance when applied to the held-out test data compared the other six ML models .

We evaluated the prediction performances of seven binary classification models when applied to held-out test data using the area under the receiver operating characteristic curve (AUROC) as the discriminative performance metric. Random forest had significantly higher test AUROC values than the other models for detecting SRNs. The median AUROC of random forest was 0.695 (IQR 0.044). L2 logistic regression, XGBoost, L2 SVM with linear and radial basis function kernel AUROC values were not significantly different from one another. They had median AUROC values of 0.68 (IQR 0.055), 0.679 (IQR 0.053), 0.678 (IQR 0.056) and 0.668 (IQR 0.056) respectively. L1 SVM with linear kernel and decision tree had significantly lower AUROC values than the other ML models with median AUROC of 0.65 (IQR 0.066) and 0.601 (IQR 0.059), respectively [Figure 2].

Predictive performances of the seven ML models during cross-validation and when applied to the held-out test data do not show large differences.

We evaluated the generalization performance of each ML model by comparing their median cross-validation AUROC and median testing AUROC. The largest difference between the two was 0.021 in L1 SVM with linear kernel followed by SVM with radial basis function kernel and decision tree with a difference of 0.007 and 0.006, respectively.

The complexity and interpretibility of each ML model.

The ML models we built using L2-regularized logistic regression, L1 and L2 support vector machines (SVM) with linear and radial basis function kernels, a decision tree, random forest and XGBoost increase in complexity as they decrase in interpretibility. We interpreted L1 and L2 SVM with linear kernel using the feature weights and L2 logistic regression using regression coefficients of the trained models. We calculated the mean weights and coefficients of all the features over the 100 data-splits. In the three linear models, OTUs that had the largest mean weights and drove the detection of SRNs belong to family *Lachnospiraceae*, and *Ruminococcaceae* (OTU01239, OTU00659, OTU00742, OTU00012, OTU00050, OTU00015, OTU00768, OTU00822, OTU00609, OTU01212, OTU00629) and genera *Gamella* (OTU00426) [Figure 3]. We explained the feature importances in non-linear models using permutation importance on the held-out test data. For the tree-based models, permuting *Peptostreptococcus* (OTU00367) abundances randomly, dropped the predictive performances the most. Decision tree, random forest and XGBoost models' predictive performance dropped from 0.6 base testing AUROC median to 0.52, from 0.69 to 0.68 and from 0.68 to 0.65, respectively [Figure 4]. The negative predictive impact of *Peptostreptococcus* in the decision tree model was followed by a *Lachnospiraceae* species (OTU00058) (0.6 base testing AUROC median to 0.58) [Figure 4B]. Other OTUs had none to minimal effect on the predictive performance.

The computational efficiency of each ML model.

Linear models trained faster than non-linear models. L2 logistic Rregression and L1 and L2 SVM with linear kernel had training times of 0.2 hours, (std ± 0.02), 0.2 hours, (std ± 0.03), and 0.2 hours, (std ± 0.03), respectively. Whereas, SVM with radial basis function kernel, decision tree, random forest and xgboost had training times of 9.4 hours, (std ± 0.8), 64.7 hours, (std ± 9.9), 101.3 hours, (std ± 10) and 162.8 hours, (std ± 4), respectively [Figure 4].

Discussion

Materials and Methods

Data collection and study population

The data used for this analysis are stool bacterial abundances and clinical information of the patients recruited by Great Lakes-New England Early Detection Research Network study. These data were obtained from Sze et al (15). The stool samples were provided by recruited adult participants who were undergoing scheduled screening or surveillance colonoscopy. Colonoscopies were performed and fecal samples were collected from participants in four locations: Toronto (ON, Canada), Boston (MA, USA), Houston (TX, USA), and Ann Arbor (MI, USA). Patients' colonic health was labeled by colonoscopy with adequate preparation and tissue histopathology of all resected lesions. Patients with an adenoma greater than 1 cm, more than three adenomas of any size, or an adenoma with villous histology were classified as advanced adenoma. Study had 172 patients with normal colonoscopies, 198 with adenomas and 120 with carcinomas. Of the 198 adenomas, 109 were identified as advanced adenomas. Stool provided by the patients was used for 16S rRNA gene sequencing to measure bacterial population abundances. The bacterial abundance data was generated by Sze et al, by processing 16S rRNA sequences in Mothur (v1.39.3) using the default quality filtering methods, identifying and removing chimeric sequences using VSEARCH and assigning to OTUs at 97% similarity using the OptiClust algorithm (16–18).

Data definitions and pre-processing

The colonic health of the patient was defined as two encompassing classes; Normal or Screen Relevant Neoplasias (SRNs). Normal class includes patients with non-advanced adenomas or normal colons whereas SRN class includes patients with advanced adenomas or carcinomas. The bacterial abundances are the features used to predict colonic health of the patients. Bacterial abundances are discrete data in the form of Operational Taxonomic Unit (OTU) counts. OTU counts were set to the size of our smallest sample and were subsampled at the same distances. They

were then transformd by scaling to a [0-1] range.

Model training and evaluation

For L2-regularized logistic regression, L1 and L2 support vector machines (SVM) with linear and radial basis function kernels we tuned the **cost** hyperparameter which determines the regularization strength where smaller values specify stronger regularization. For SVM with radial basis function kernel we also tuned **sigma** hyperparameter which determines the reach of a single training instance where for a high value of sigma, the SVM decision boundary will be dependent on the points that are closest to the decision boundary. For the decision tree model, we tuned the **depth of the tree** where deeper the tree, the more splits it has. For random forest, we tuned the **number of features** to consider when looking for the best tree split. For xgboost, we tuned for **learning rate** and the **fraction of samples** to be used for fitting the individual base learners. Models were trained using the machine learning wrapper caret package (v.6.0.81) in R (v.3.5.0).

Statistical analysis workflow. Data summaries, statistical analysis, and data visualizations were performed using R (v.3.5.0) with the tidyverse package (v.1.2.1). We compared the AUROC values of the seven ML models by Wilcoxon rank sum tests to determine the best discriminative performance.

Code availability. The code for all sequence curation and analysis steps including an Rmarkdown version of this manuscript is available at https://github.com/SchlossLab/Topcuoglu_ML_XXXX_2019/.

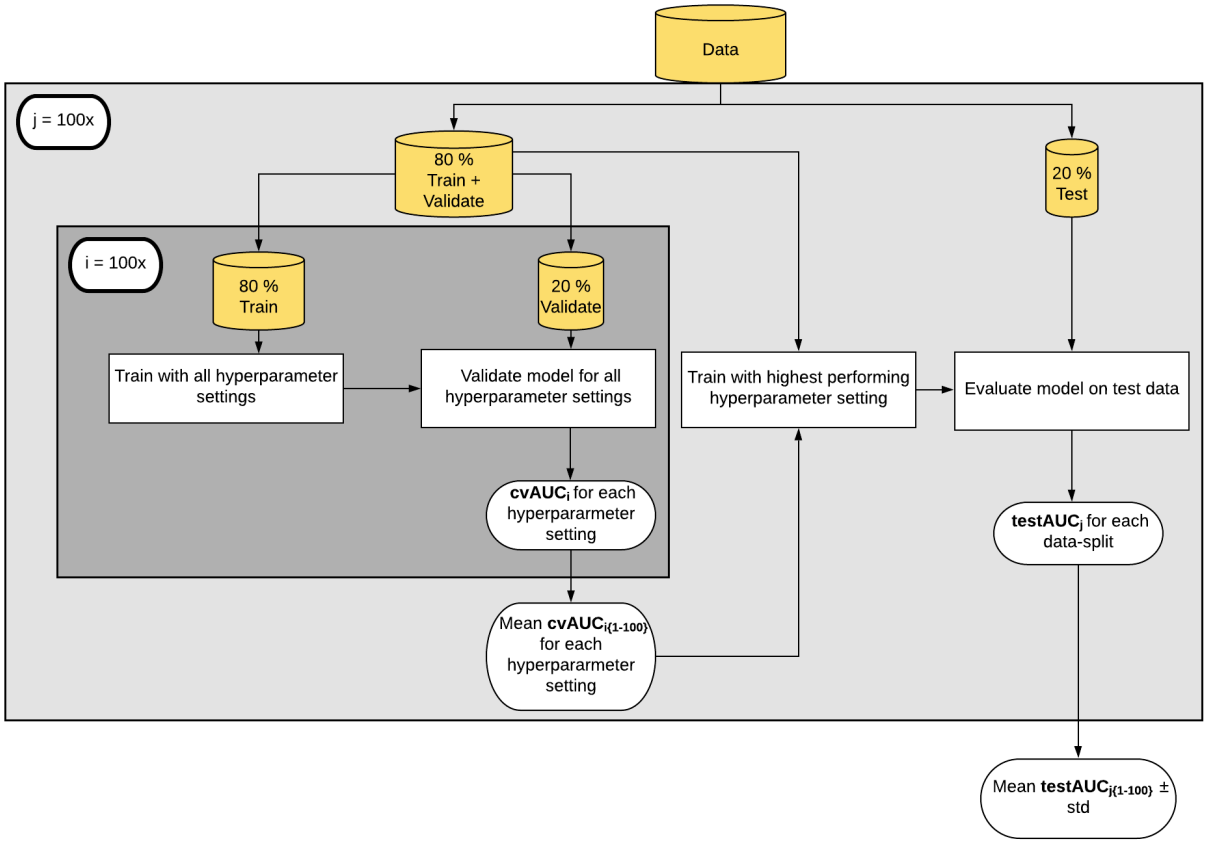

Figure 1. Machine learning pipeline showing predictive model training and evaluation flowchart. We split the data 80%/20% stratified to maintain the overall label distribution, performed five-fold cross-validation on the training data to select the best hyperparameter setting and then using these hyperparameters to train all of the training data. The model was evaluated on a held-out set of data (not used in selecting the model). Abbreviations: AUROC, area under the receiver operating characteristic curve

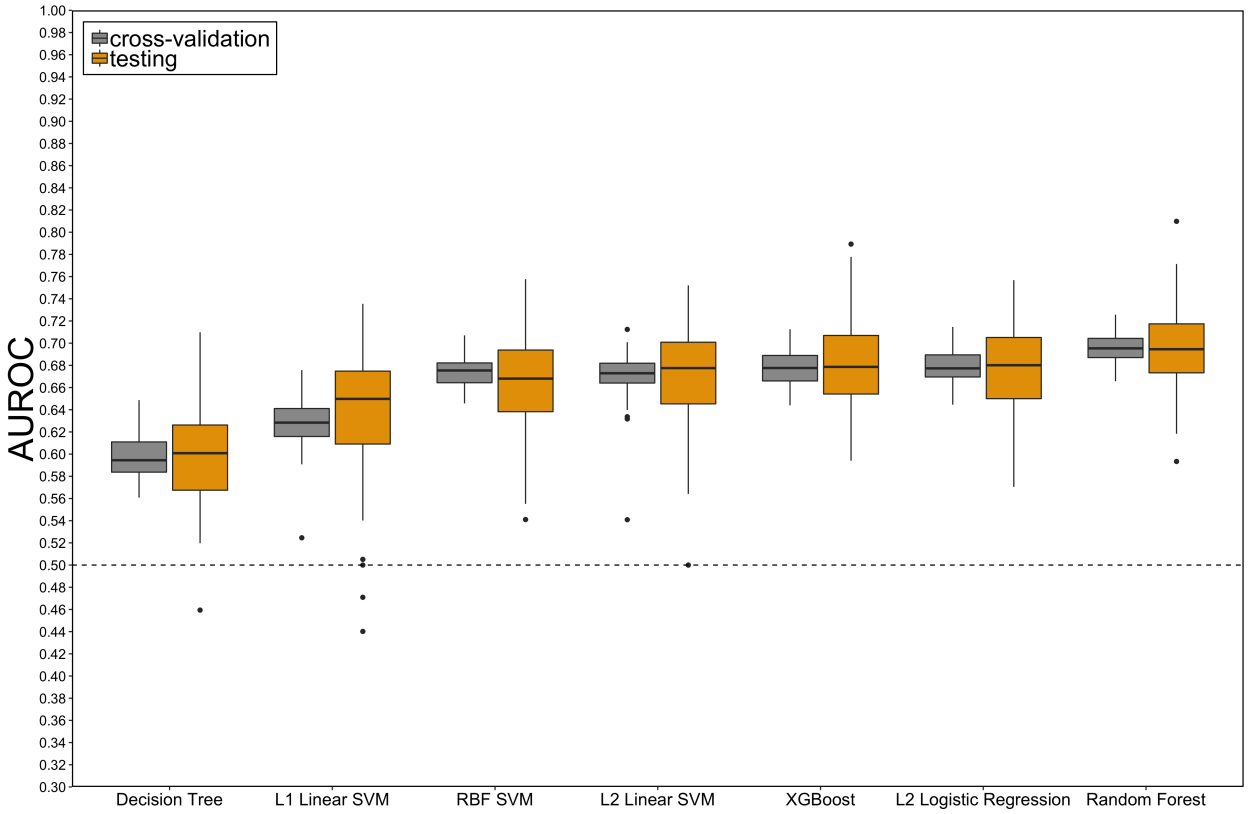

Figure 2. Generalization and classification performance of ML models using AUROC values of all cross validation and testing performances. The median AUROC for diagnosing individuals with SRN using bacterial abundances was higher than chance (depicted by horizontal line at 0.50) for all the ML models. Discriminative performance of random forest model was higher than other ML models. The boxplot shows quartiles at the box ends and the statistical median as the horizontal line in the box. The whiskers show the farthest points that are not outliers. Outliers are data points that are not within 3/2 times the interquartile ranges. Abbreviations: SRN, screen-relevant neoplasias; AUROC, area under the receiver operating characteristic curve; SVM, support vector machine; XGBoost, extreme gradient boosting

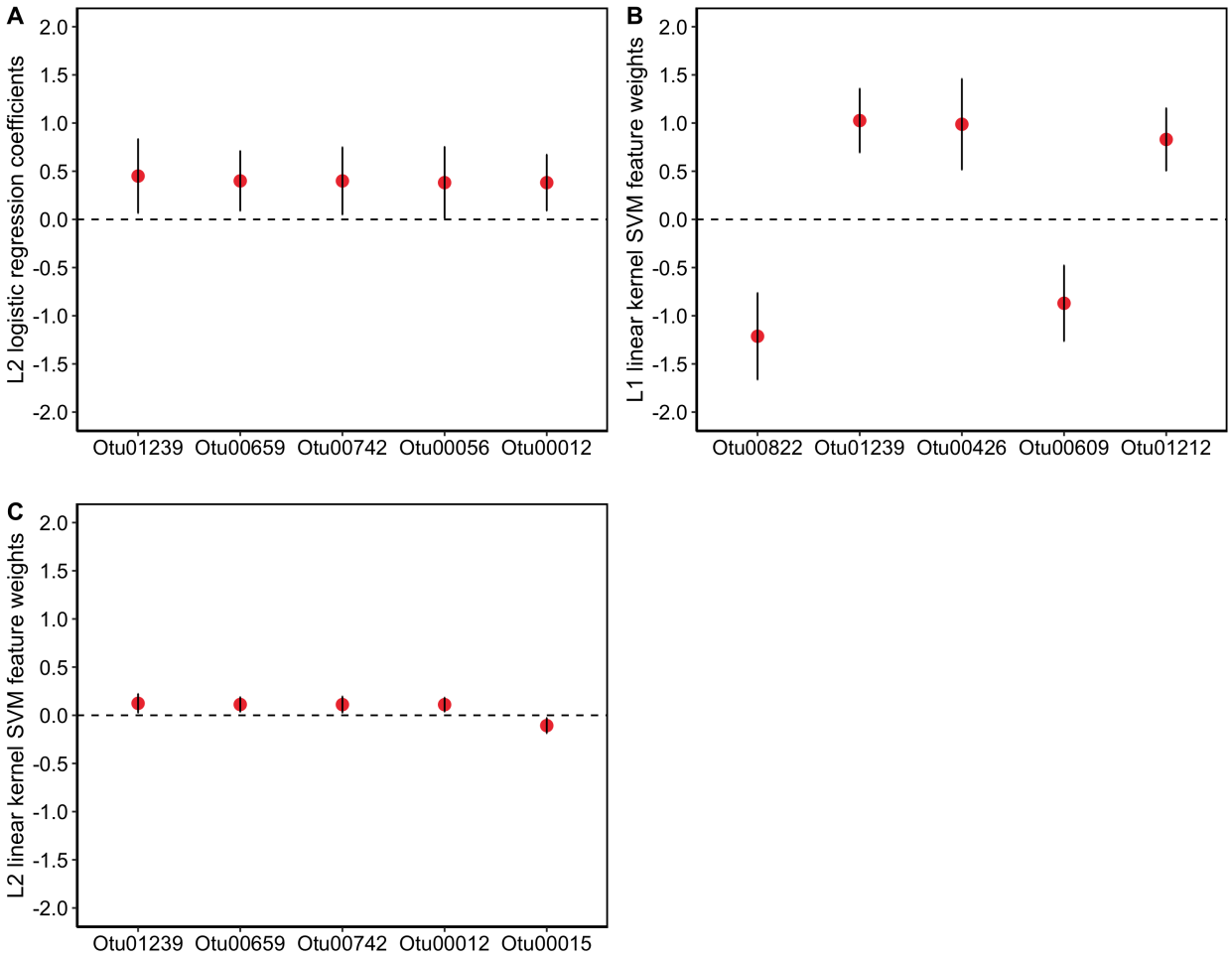

Figure 3. Interpretation of the linear ML models. (A) L2 logistic regression coefficients (B) L1 SVM with linear kernel feature weights (C) L2 SVM with linear kernel feature weights. The means weights and coefficients of the most important five OTUs for each model are shown here with the standard deviation over 100 data-splits. Similar OTUs had the largest impact on the predictive performance of L2 logistic regression and L2 SVM with linear kernel. Abbreviations: SVM, support vector machine; OTU, Operational Taxonomic Unit.

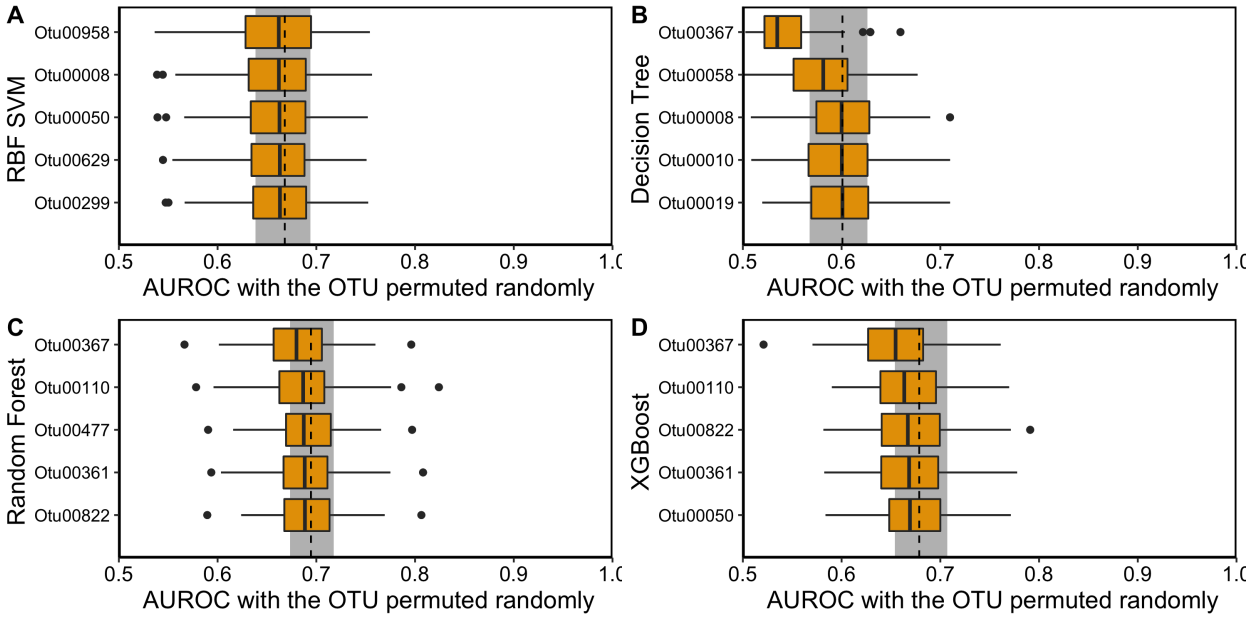

Figure 4. Explanation of the non-linear ML models. (A) SVM with radial basis kernel (B) decision tree (C) random forest (D) XGboost feaure importances were explanied using permutation importance using held-out test set. The gray rectangle and the dashed line show the IQR range and median of the base testing AUROC without any permutation performed. For all the tree-based models, a *Peptostreptococcus* species (OTU00367) had the largest impact on predictive performance of the model. Abbreviations: SVM, support vector machine; OTU, Operational Taxonomic Unit; RBF, radial basis kernel; OTU, Operational Taxonomic Unit.

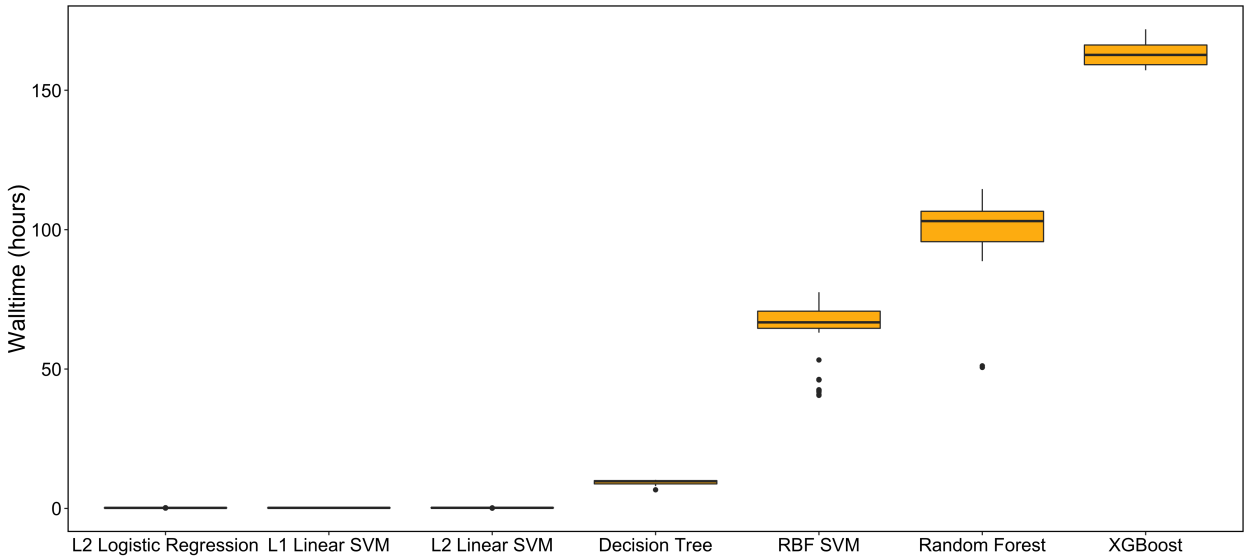

Figure 5. Computational efficiency of seven ML models. The walltimes for training and testing of each data-split showed the differences in computational efficieny of the seven models. The median walltime in hours was the highest for XGBoost and shortest for L2 logistic regression. The boxplot shows quartiles at the box ends and the statistical median as the horizontal line in the box. The whiskers show the farthest points that are not outliers. Outliers are data points that are not within 3/2 times the interquartile ranges. Abbreviations: AUROC, area under the receiver operating characteristic curve; SVM, support vector machine; XGBoost, extreme gradient boosting.

References

1. **Zeller G**, **Tap J**, **Voigt AY**, **Sunagawa S**, **Kultima JR**, **Costea PI**, **Amiot A**, **Böhm J**, **Brunetti F**, **Habermann N**, **Hercog R**, **Koch M**, **Luciani A**, **Mende DR**, **Schneider MA**, **Schrotz-King P**, **Tournigand C**, **Tran Van Nhieu J**, **Yamada T**, **Zimmermann J**, **Benes V**, **Kloor M**, **Ulrich CM**, **Knebel Doeberitz M von**, **Sobhani I**, **Bork P**. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Mol Syst Biol **10**. doi:10.15252/msb.20145645.

2. **Zackular JP**, **Rogers MAM**, **Ruffin MT**, **Schloss PD**. 2014. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res **7**:1112–1121. doi:10.1158/1940-6207.CAPR-14-0129.

3. **Baxter NT**, **Koumpouras CC**, **Rogers MAM**, **Ruffin MT**, **Schloss PD**. 2016. DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model. Microbiome **4**. doi:10.1186/s40168-016-0205-y.

4. **Baxter NT**, **Ruffin MT**, **Rogers MAM**, **Schloss PD**. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. Genome Medicine **8**:37. doi:10.1186/s13073-016-0290-3.

5. **Hale VL**, **Chen J**, **Johnson S**, **Harrington SC**, **Yab TC**, **Smyrk TC**, **Nelson H**, **Boardman LA**, **Druliner BR**, **Levin TR**, **Rex DK**, **Ahnen DJ**, **Lance P**, **Ahlquist DA**, **Chia N**. 2017. Shifts in the fecal microbiota associated with adenomatous polyps. Cancer Epidemiol Biomarkers Prev **26**:85–94. doi:10.1158/1055-9965.EPI-16-0337.

6. **Pasolli E**, **Truong DT**, **Malik F**, **Waldron L**, **Segata N**. 2016. Machine learning meta-analysis of large metagenomic datasets: Tools and biological insights. PLoS Comput Biol **12**. doi:10.1371/journal.pcbi.1004977.

7. **Sze MA**, **Schloss PD**. 2016. Looking for a signal in the noise: Revisiting obesity and the microbiome. mBio **7**. doi:10.1128/mBio.01018-16.

8. **Walters WA**, **Xu Z**, **Knight R**. 2014. Meta-analyses of human gut microbes associated with

obesity and IBD. FEBS Lett **588**:4223–4233. doi:10.1016/j.febslet.2014.09.039.

9. **Vázquez-Baeza Y**, **Gonzalez A**, **Xu ZZ**, **Washburne A**, **Herfarth HH**, **Sartor RB**, **Knight R**. 2018. Guiding longitudinal sampling in IBD cohorts. Gut **67**:1743–1745. doi:10.1136/gutjnl-2017-315352.

10. **Qin N**, **Yang F**, **Li A**, **Prifti E**, **Chen Y**, **Shao L**, **Guo J**, **Le Chatelier E**, **Yao J**, **Wu L**, **Zhou J**, **Ni S**, **Liu L**, **Pons N**, **Batto JM**, **Kennedy SP**, **Leonard P**, **Yuan C**, **Ding W**, **Chen Y**, **Hu X**, **Zheng B**, **Qian G**, **Xu W**, **Ehrlich SD**, **Zheng S**, **Li L**. 2014. Alterations of the human gut microbiome in liver cirrhosis. Nature **513**:59–64. doi:10.1038/nature13568.

11. **Geman O**, **Chiuchisan I**, **Covasa M**, **Doloc C**, **Milici M-R**, **Milici L-D**. 2018. Deep learning tools for human microbiome big data, pp. 265–275. *In* Balas, VE, Jain, LC, Balas, MM (eds.), Soft computing applications. Springer International Publishing.

12. **Galkin F**, **Aliper A**, **Putin E**, **Kuznetsov I**, **Gladyshev VN**, **Zhavoronkov A**. 2018. Human microbiome aging clocks based on deep learning and tandem of permutation feature importance and accumulated local effects. bioRxiv. doi:10.1101/507780.

13. **Reiman D**, **Metwally A**, **Dai Y**. 2017. Using convolutional neural networks to explore the microbiome, pp. 4269–4272. *In* 2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC).

14. **Fioravanti D**, **Giarratano Y**, **Maggio V**, **Agostinelli C**, **Chierici M**, **Jurman G**, **Furlanello C**. 2017. Phylogenetic convolutional neural networks in metagenomics. arXiv:170902268 [cs, q-bio].

15. **Sze MA**, **Schloss PD**. 2018. Leveraging existing 16S rRNA gene surveys to identify reproducible biomarkers in individuals with colorectal tumors. mBio **9**:e00630–18. doi:10.1128/mBio.00630-18.

16. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Van Horn DJ**, **Weber CF**. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. ApplEnvironMicrobiol

75:7537–7541.

17. **Westcott SL**, **Schloss PD**. 2017. OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. mSphere **2**. doi:10.1128/mSphereDirect.00073-1

18. **Rognes T**, **Flouri T**, **Nichols B**, **Quince C**, **Mahé F**. 2016. VSEARCH: A versatile open source tool for metagenomics. PeerJ **4**:e2584. doi:10.7717/peerj.2584.