Best practices for applying machine learning to bacterial 16S rRNA gene sequencing data

Running title: Machine learning methods in microbiome studies

Begüm D. Topçuoğlu[1], Nicholas A. Lesniak[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

Machine learning (ML) modeling of the human microbiome has the potential to identify the microbial biomarkers and aid in diagnosis of many chronic diseases such as inflammatory bowel disease, diabetes and colorectal cancer. Progress has been made towards developing ML models that predict health outcomes from bacterial abundances, but rigourous ML models are scarce due to the flawed methods that call the validity of developed ML models into question. Furthermore, the use of black box ML models has hindered the validation of microbial biomarkers. To overcome these challenges, we benchmarked seven different ML models that use fecal 16S rRNA sequences to predict colorectal cancer (CRC) lesions (n=490 patients, 261 controls and 229 cases). To show the effect of model selection, we assessed the predictive performance, interpretability, and computational efficiency of the following models: L2-regularized logistic regression, L1 and L2-regularized support vector machines (SVM) with linear and radial basis function kernels, a decision tree, random forest, and extreme gradient boosting (XGBoost). The random forest model was best at detecting CRC lesions with an AUROC of 0.695 but it was slow to train (83.2 h) and hard to interpret. Despite its simplicity, L2-regularized logistic regression followed random forest in predictive performance with an AUROC of 0.680, and it trained much faster (12 min). In this study, we established standards for the development of modeling pipelines for microbiome-associated ML models. Additionally, we showed that ML models should be chosen based on expectations of predictive performance, interpretability and available computational resources.

Importance (needs work)

Prediction of health outcomes using ML is rapidly being adopted by human microbiome studies. However, the developed ML models so far are overoptimistic in terms of validity and predictive performance. Without rigorous ML pipelines, we cannot trust ML models. Before we can speed up progress, we need to slow down, define and implement good ML practices.

Background

Advances in sequencing technology and decreases in the costs of generating 16S rRNA gene sequences have allowed rapid exploration of the human microbiome and its health implications. Currently, the human microbiome field is growing at an unprecedented rate and there is an increasing demand for methods that identify associations between microbiome members and human health. However, this is difficult as human associated microbial communities are remarkably complex, high-dimensional and uneven within and between individuals with the same disease. 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 differences in health outcomes.

Machine learning (ML) methods are effective at recognizing and highlighting patterns in complex microbial datasets. They learn from existing data to predict the outcomes of new data and allow us to infer on the reasons underlying that prediction. Therefore, researchers have started to explore the utility of ML models that use microbiota associated biomarkers to predict human health and to understand the ecological basis of diseases such as liver cirrhosis, colorectal cancer, inflammatory bowel diseases (IBD), obesity, and type 2 diabetes (1–11). The high-stake task of predicting the correct diagnosis with high confidence relies on a ML model that is built with rigorous methods. However, the field's use of ML lacks clarity and consistency in which methods are used and how these methods are implemented (12, 13). More notably, we commonly see flawed ML practices such as using ML pipelines where there is no seperate held-out test dataset to evaluate predictive performance or reporting few or only the best outcomes of cross-validation. Even when there are seperate testing sets to evaluate predictive performance, there are large differences between cross-validation and testing performances and large confidence intervals in testing performances (4, 14–20). These practices prevent the development of generalizable models, where the model makes accurate predictions with newly acquired data as well as it does with the training data.

Moreover, there is a lack of discussion on why a particular ML model is utilized and a lack of emphasis on the strengths and weaknesses of it. Recently, there is a trend towards using more complex ML models such as random forest, extreme gradient boosting (XGBoost) and neural networks without a discussion on if and how much model interpretibility is necessary for the study (11, 21–23). These models are also called black box ML models because they are not inherently interpretable and require posthoc explanations to determine the feature importances in making a prediction. These explanations can be misleading and at times unreliable when making high-stake decisions about patient health (24). The models we develop for healthcare, both to predict disease and to understand underlying reasons behind that prediction, should be transparent and accountable (25).

The lack of transparency on model selection and interpretation as well as flawed modeling methods negatively impact model validity and reproducibility. We need to strive toward better machine learning practices by (1) implementing rigorous machine learning pipelines and (2) selecting ML models that reflect the goal of the study as it will inform our expectations of predictive performance, complexity, interpretability and computational efficiency. To showcase a rigorous ML pipeline and to shed light on how ML model selection can affect modeling results, we performed an empirical analysis comparing several different ML models using the same dataset and the same ML pipeline. We used a previously published colorectal cancer (CRC) study (3) which had fecal 16S rRNA gene sequences of CRC patients. The human microbiome is hypothesized to directly contribute to the development of CRC and fecal 16S rRNA gene sequences have been used to detect it (1, 3, 4, 26). We built seven ML models using fecal 16S rRNA gene sequences to predict healthy patients versus patients with colorectal lesions that were identified by colonoscopy as screen relevant neoplasias (SRN). We established modeling pipelines for three linear models with different forms of regularization; L2-regularized logistic regression, L1 and L2-regularized support vector machines (SVM) with linear kernel. We also developed four non-linear models; SVM with radial basis function kernel, a decision tree, random forest and XGBoost. We compared the predictive performance, interpretability and computational efficiency of the seven models to highlight the importance of model selection. We established standards for ML pipeline construction, predictive performance evaluation and model interpretation for microbiome-associated ML models that have a binary prediction task.

Results

Model selection and pipeline construction

We first determined the dataset and the ML models to use in our study. We used a previously published study on a CRC 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 colorectal health (normal or SRN colorectal lesions as defined by 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 extreme gradient boosted decision tree (XGBoost).

We used ML models with different classification algorithms and regularization parameters. For regularized logistic regression and SVM with linear kernel, we used L2 regularization to keep all potentially important features. For comparison, we also trained an L1 regularized SVM model with linear kernel. L1-regularization on microbiome data lead to a sparser solution (i.e., force many coefficients to zero). Finally, to explore the potential for non-linear relationships among features and the outcome of interest, we trained tree based models, decision tree, random forest and XGBoost, as well as an SVM with non-linear kernel.

We established a ML pipeline where we train and validate each of the seven models [Figure 1]. 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 in 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.

Hyperparameters are the rules that are learned from the training data in a classification algorithm. When they are tuned over a full grid search and selected by validation, they make the ML model better are predicting. We selected hyperparameter settings using repeated five-fold cross-validation on the training/validation set [Figure 1]. We chose the best hyperparameter setting for each model based on its predictive performance on the validation set using the area under the receiver operating characteristic curve (AUROC) metric [Figure S1 and S2]. The AUROC ranges from 1.0, where the model perfectly distinguishes between cases and controls, to 0.50, where the model's predictions are no different from random chance. Similar to the initial data-split, five-fold cross-validation for hyperparameter selection was also stratified to maintain the overall label distribution on the training and validation sets. The cross-validation of each hyperparameter setting was repeated over 100 randomizations to get a robust reading of predictive performance.

We then trained the full training/validation dataset with the selected hyperparameters. We used the held-out test set to evaluate the predictive 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 performance [Figure 1].

Predictive performance and generalizability of the seven models.

We evaluated the predictive performances of seven binary classification models when applied to held-out test data using the AUROC metric [Figure 2]. Random forest had significantly higher test AUROC values than the other models for detecting SRNs when AUROC values were compared to the other six by Wilcoxon rank sum test ($p < 0.01$). The median AUROC of the random forest model was 0.695 (IQR 0.044). L2-regularized logistic regression, XGBoost, L2-regularized 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 0.052), 0.678 (IQR 0.056) and 0.668 (IQR 0.056) respectively. L1-regularized 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]. Random forest had the highest median AUROC for detecting SRN. Despite its simplicity, the L2-regularized logistic regression was second best in predictive performance.

To evaluate the generalizability of each model, we compared the median cross-validation AUROC to the median testing AUROC. The difference between the two should be low to suggest the model is not overfitting despite the large number of features. The largest difference between the two was

0.021 in L1-regularized 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 [Figure 2]. We also reported the testing AUROC values over 100 randomizations of the initial data-split. The testing AUROC values within each model varied 0.23 on average across the seven models. For instance, the lowest AUROC value of the random forest model was 0.59 whereas the highest was 0.81. These results showed that depending on the data-split, the testing AUROC values showed great variability [Figure 2].

Interpretation of each ML model.

Interpretability is the degree to which humans can understand the reasons behind a model prediction (27). Because we often use ML models not just to predict a health outcome but also to learn the ecology behind a disease, model interpretation becomes crucial for microbiome studies. 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 decrease in interpretability as they increase in complexity. In this study we highlighted two methods to interpret models with varying complexity.

We interpreted linear models (L1 and L2-regularized SVM with linear kernel and L2-regularized logistic regression) using the absolute feature weights of the trained models. We ranked the absolute weights of all the OTUs for each data-split [Figure 3]. We calculated the median ranks of these features over the 100 data-splits. In the three linear models, OTUs that had the largest median ranks and drove the detection of SRNs belonged to families *Lachnospiraceae*, and *Ruminococcaceae* (OTU01239, OTU00659, OTU00742, OTU00012, OTU00015, OTU00768, OTU00822, OTU00609), genera *Gamella* (OTU00426) and genera *Peptostreptococcus* (OTU00367) [Figure 3]. Some of the OTUs with the highest ranks were shared among the linear models.

We explained the feature importances in non-linear models; SVM with radial basis kernel, decision tree, random forest and XGBoost, using a method called permutation importance on the held-out test data. Permutation importance analysis is a posthoc explanation of the model where we randomly permute non-correlated features individually and groups of highly correlated features together. We then calculate how much the predictive performance of the model (i.e AUROC values) decrease when each OTU or group of OTUs is permuted randomly. We ranked the OTUs based on how much they decreased the median testing AUROC; the OTU with the largest decrease ranking highest. The top 5 OTUs with the largest negative impact on testing AUROC overlapped in tree-based models [Figure 4]. Specifically, permuting *Peptostreptococcus* (OTU00367) abundances randomly, dropped the predictive performances the most in all tree-based methods [Figure 4]. Decision tree, random forest and XGBoost models' predictive performance dropped from 0.6 median AUROC to 0.52, from 0.69 to 0.68 and from 0.68 to 0.65, respectively [Figure 4].

To highlight the differences between the two interpretation methods, we used permutation importance to interpret linear models as well [Figure S3]. L1-regularized SVM with linear kernel picked out some of the same OTUs (OTU00822, OTU01239, OTU00609) as important in feature rankings based on weights [Figure 3] and permutation importance [Figure S3]. Similarly, L2-regularized SVM and L2-regularized logistic regression picked out some of the same OTUs in both interpretation methods, OTU00659 and OTU00012, respectively. However, for all the linear models, the rankings of these features were different due to the collinearity in microbial communities.

The computational efficiency of each ML model.

We compared the training times of the seven ML models. As the complexity of a ML model [Table S1] and the number of tuned hyperparameter settings increased [Figures S1-S2], its training times increased as well [Figure 5]. Linear models trained faster than non-linear models. L1 and L2 SVM with linear kernel and L2-regularized logistic regression had the shortest training times with 0.2 hours, (std ± 0.03), 0.2 hours, (std ± 0.02), and 0.2 hours, (std ± 0.02), respectively. Whereas, a decision tree, SVM with radial basis function kernel, random forest and XGBoost had training times of 4.4 hours, (std ± 0.3), 59.6 hours, (std ± 8.8), 83.2 hours, (std ± 11.3) and 155.1 hours, (std ± 1), respectively [Figure 5].

Discussion

In this study we established a rigorous ML pipeline to use 16S rRNA sequence counts to predict a binary health outcome. We built on others' work in the microbiome field to set-up standards for developing and evaluating ML models for microbiome data (28–30). First, we used a held-out test set to illustrate the difference between cross-validation and testing AUROC values. When the difference between cross-validation and test performance is low, this suggest the models are not overfit and that they will perform similar with similar data. In all seven models, the difference between median cross-validation and testing AUROC values did not exceed 0.021 which suggests that these models are generalizable and can be used to test similar new data. Second, we performed the initial 80%-20% random datasplit 100 times in our ML pipeline. The randomization of the initial data-split to create a held-out test set is a crucial step in the ML pipeline to develop robust ML models and to report reliable performance metrics. Depending on how the data is split, there is the chance of being overoptimistic about the predictive performance of a model. In our study, we showed that there was variability in AUROC values between different random data-splits in each of the models we tested. Our results showed that the testing AUROC values varied 0.23 on average between different data-splits. Third, we used the AUROC metric in our study to evaluate the predictive performance of the ML models. AUROC is always random at the value 0.5 and is a robust metric when a dataset is imbalanced. We also performed a full grid search for hyperparameter settings when building a ML model. Default hyperparameter settings in previously developed ML packages in R, Python, and Matlab programming languages are inadequate for effective application of classification algorithms and need to be optimized for each new dataset used to generate a model. In the example of L1-regularized SVM with linear kernel [Figure S1], the model showed large variability between different regularization coefficients (C) and was susceptible to performing poorly if the wrong regularization coefficient was assigned to the model by default.

In this study, we benchmarked seven ML models with different classification algorithms to show that we should use ML models based on the goal of the study and our expectations of predictive performance, interpretability and computational burden. Microbiome studies use ML models with a classification task to learn the training data to assign labels, such as healthy or not to new data, but also to learn which features are important to discriminate between labels (2–11). Our results show that if the goal of a study is to learn the ecology behind a disease and to identify microbial biomarkers, we could create ML models that are inherently interpretable without losing predictive power. In terms of predictive performance, random forest model had the best testing AUROC values compared to the other 6 models. However, the second best model was L2-regularized logistic regression with a median AUROC difference of only 0.015 compared to random forest. In terms of interpretability, random forest was a complex ML model and could only be explained using methods such as permutation importance. On the other hand, L2-regularized logistic regression was easier to interpret (i.e. ranking absolute regression coefficients of the trained model).

Even with interpretable models such as L2-regularized logistic regression, we need to be careful with how we treat the information we get from the models. In this study we used two different methods to interpret our linear models; ranking each OTU by (1) their absolute weights in the trained models and (2) their impact on the predictive performance based on permutation importance. We observed differences in the OTU rankings between the two interpretation methods due to collinearity in the dataset. Collinearity in a microbial dataset is when one OTU is highly correlated with another OTU which causes feature weights to not be unique. The feature weights of correlated OTUs are influenced by one another which makes it difficult to interpret the ML model. To avoid misinterpreting the models we should use the highly ranked correlated OTUs to generate hypotheses about the ecology of the disease and test them with follow-up experiments.

There are other criteria when choosing ML models such as the computational burden of developing it and the sample size. In terms of computational burden, random forest model trained each data-split in 83.2 hours whereas L2-regularized logistic regression trained in 12 minutes. The generalization performance of ML models depends on sample size. The more complex the model, the more data it will need. The dataset we used for our study had 490 samples, however microbiome studies that have smaller sample sizes would benefit from using less complex models such as L2-regularized logistic regression.

This study highlights the need to make educated choices at every step of developing a ML model with microbiome data. Model selection should be done with a solid understanding of model complexity and interpretability, rigorous ML pipelines should be built with cross-validation for hyperparameter tuning and with a held-out test set for evaluating predictive performance and models should be interpreted while considering collinearity in datasets. The right methods will help us achieve the level of validity and accountability we want from models built for patient health.

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 (31). 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 (32–34).

Data definitions and pre-processing.

The colorectal 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 study had 261 normal and 229 SRN samples. 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 transformed by scaling to a [0-1] range.

Model training and evaluation.

Models were trained using the machine learning wrapper caret package (v.6.0.81) in R (v.3.5.0). Within the caret package, we have made modifications to L2-regularized SVM with linear kernel function **svmLinear3** and developed a L1-regularized SVM with linear kernel function **svmLinear4** to calculate decision values instead of predicted probabilities. These changes are available at https://github.com/SchlossLab/Topcuoglu_ML_XXXX_2019/.

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.For hyperparameter selection, we started with a granular grid search. Then we narrowed and fine-tuned the range of each hyperparameter. The range of the grid depends on the ML task and ML model. A full grid search needs to be performed to avoid variability in testing performance. We can use hyper-band to help us with our hyperparameter selection (35).

The computational burden during model training due to model complexity was reduced by parallelizing segments of the ML pipeline. In this study we have parallelized each data-split which allowed 100 data-splits to be processed through the ML pipeline at the same time for each model. We can further parallelize the cross-validation step for each hyperparameter setting.

Permutation importance workflow. We created a Spearman's rank-order correlation matrix, corrected for multiple pairwise comparisons. We then defined correlated OTUs as having perfect correlation (correlation coef=1 and p<0.01). Non-correlated OTUs were permuted individually whereas correlated ones were grouped together and permuted at the same time.

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 predictive 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/.

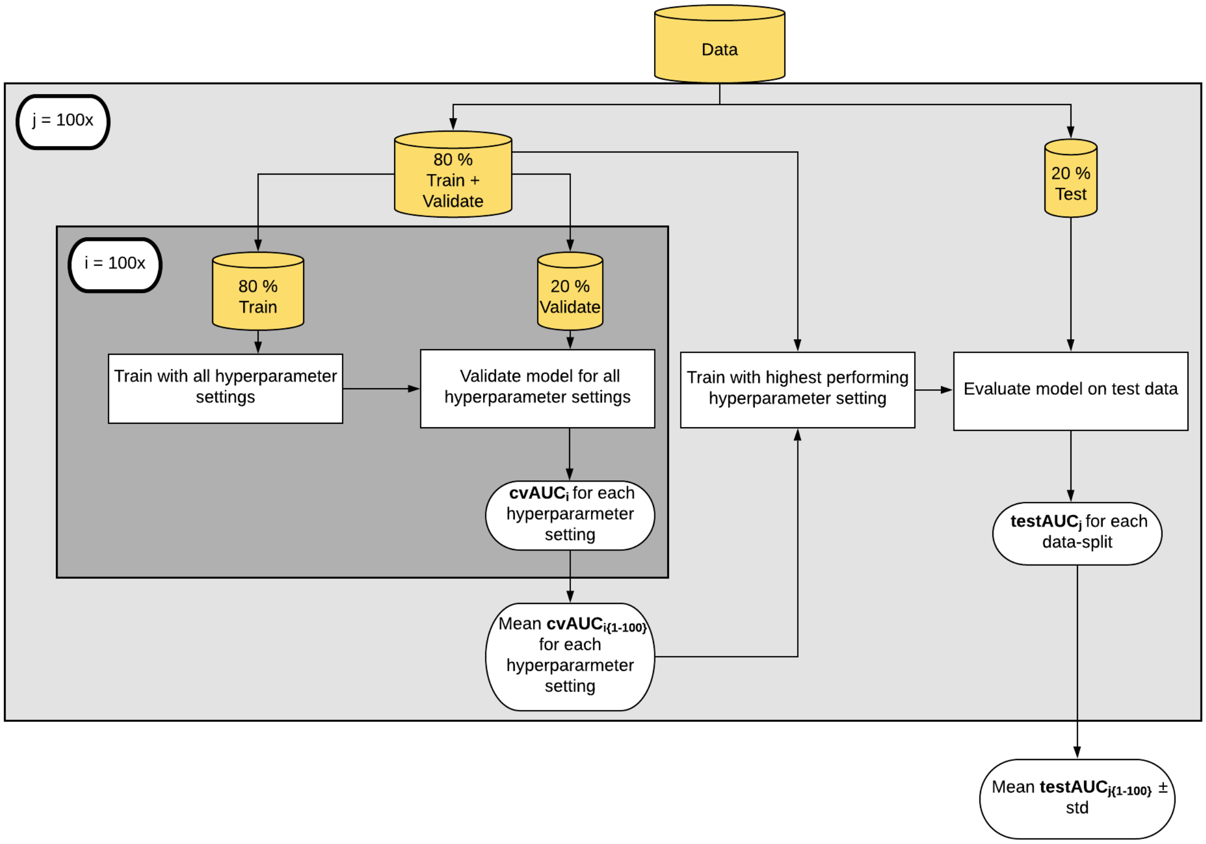
Data
j = 100x
80 % Train + Validate
20 % Test
i = 100x
80 % Train
20 % Validate
Train with all hyperparameter settings
Validate model for all hyperparameter settings
$cvAUC_i$ for each hyperpararmeter setting
Mean **$cvAUC_{i\{1\text{-}100\}}$** for each hyperpararmeter setting
Train with highest performing hyperparameter setting
Evaluate model on test data
$testAUC_j$ for each data-split
Mean **$testAUC_{j\{1\text{-}100\}}$** ± std

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: cvAUROC, cross-validation area under the receiver operating characteristic curve

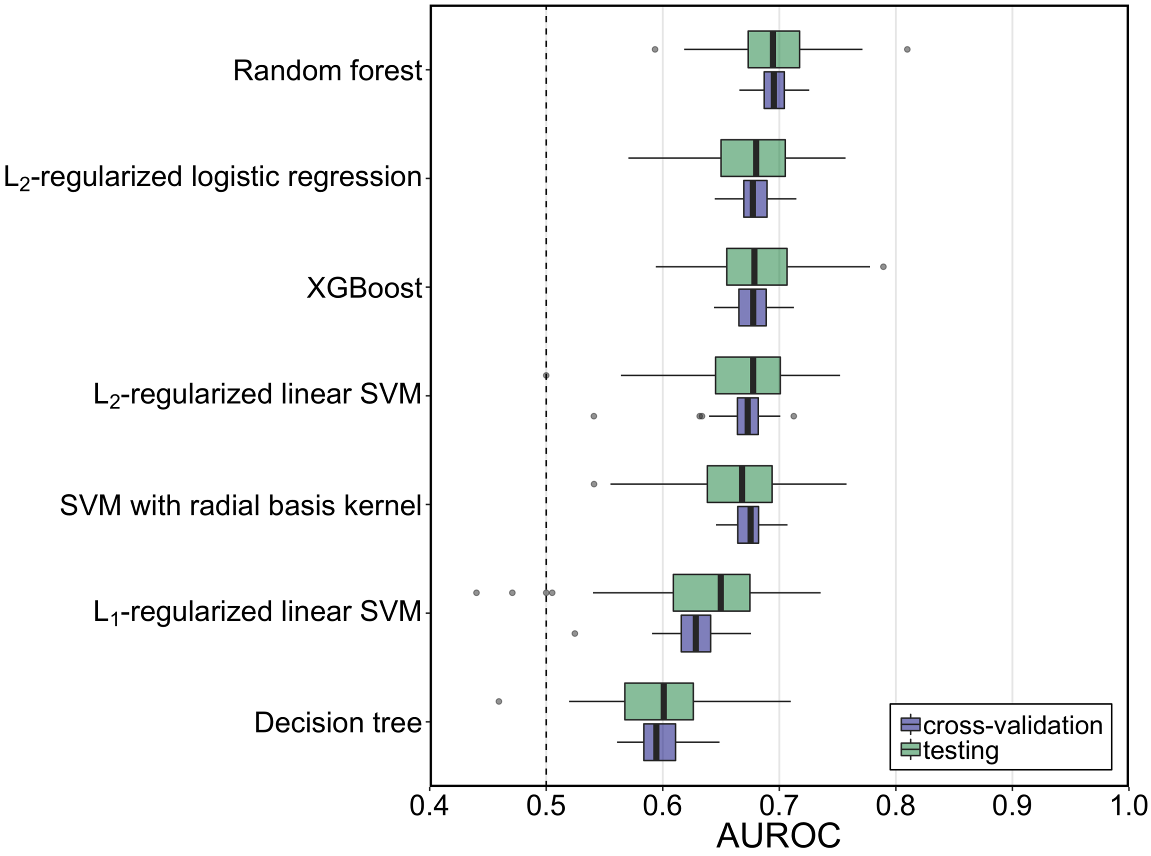

Figure 2. Generalization and classification performance of ML models using AUROC values of all cross validation and testing performances. The median AUROC for diagnosing individuals with SRN using bacterial abundances was higher than chance (depicted by horizontal line at 0.50) for all the ML models. Predictive 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

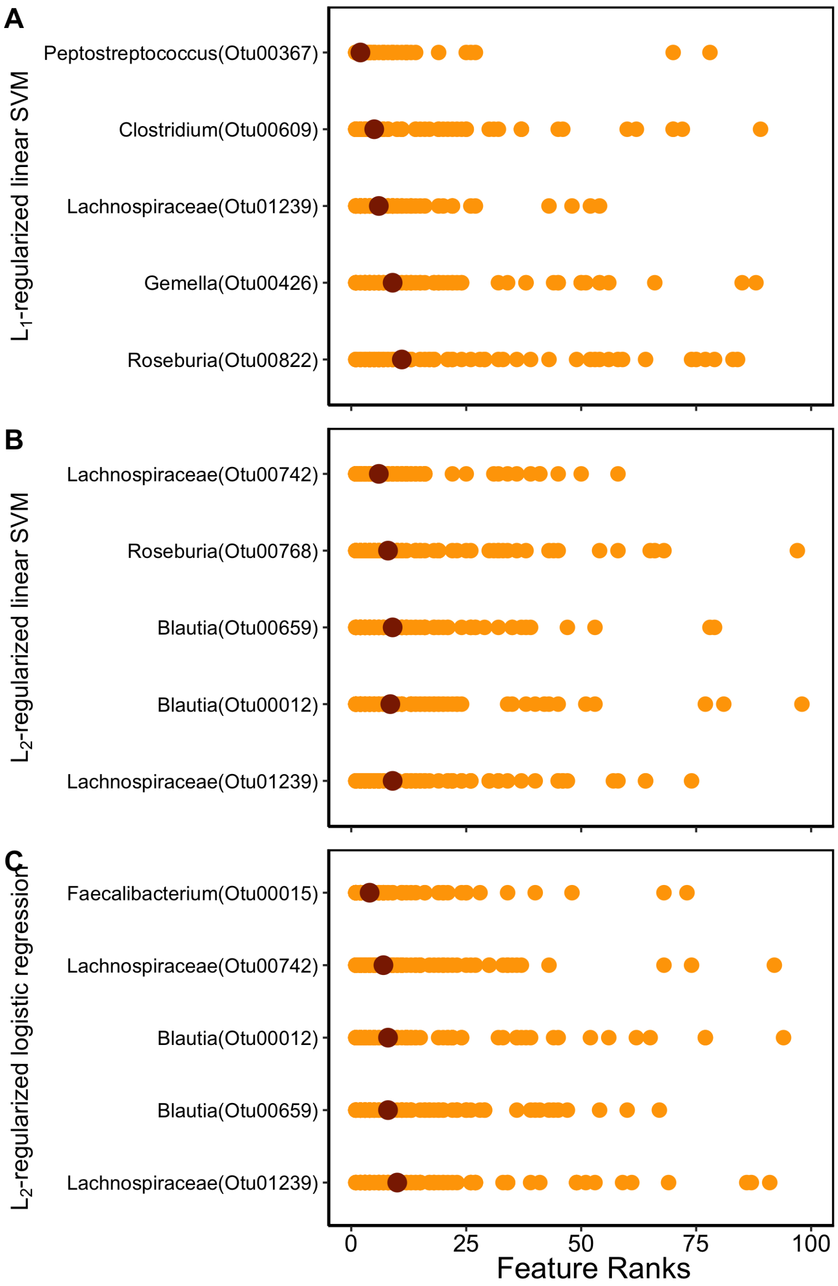

Figure 3. Interpretation of the linear ML models. The absolute feature weights of (A) L2 logistic regression coefficients (B) L1 SVM with linear kernel (C) L2 SVM with linear kernel were ranked from highest rank 1 to 100 for each data-split. The feature ranks of the highest ranked five OTUs based on their median ranks are shown here. 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.

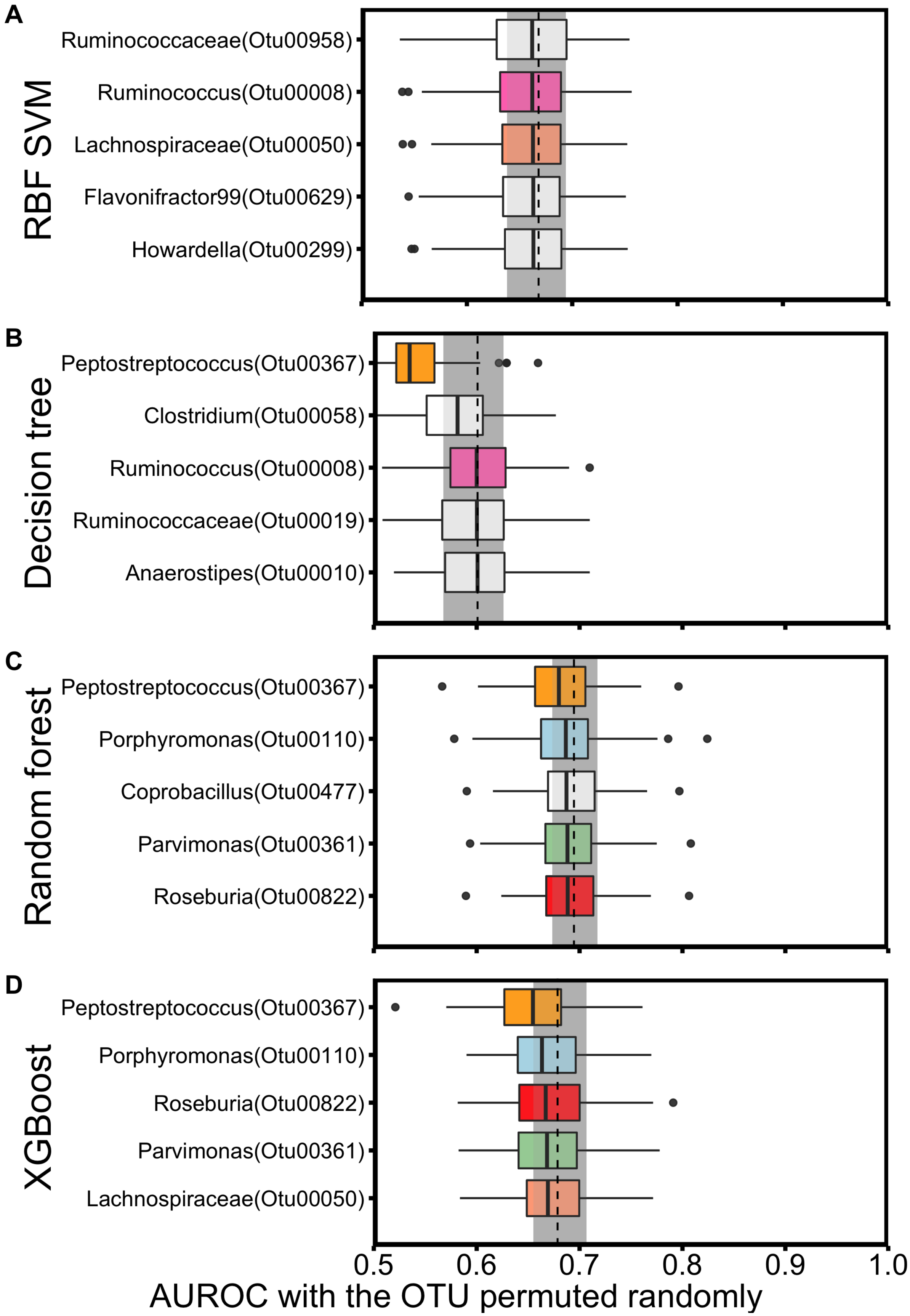
A
RBF SVM
Ruminococcaceae(Otu00958)
Ruminococcus(Otu00008)
Lachnospiraceae(Otu00050)
Flavonifractor99(Otu00629)
Howardella(Otu00299)
B
Decision tree
Peptostreptococcus(Otu00367)
Clostridium(Otu00058)
Ruminococcus(Otu00008)
Ruminococcaceae(Otu00019)
Anaerostipes(Otu00010)
C
Random forest
Peptostreptococcus(Otu00367)
Porphyromonas(Otu00110)
Coprobacillus(Otu00477)
Parvimonas(Otu00361)
Roseburia(Otu00822)
D
XGBoost
Peptostreptococcus(Otu00367)
Porphyromonas(Otu00110)
Roseburia(Otu00822)
Parvimonas(Otu00361)
Lachnospiraceae(Otu00050)
0.5
0.6
0.7
0.8
0.9
1.0
AUROC with the OTU permuted randomly

Figure 4. Interpretation of the non-linear ML models. (A) SVM with radial basis kernel (B) decision tree (C) random forest (D) XGBoost feature 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. The colors of the box plots stand for the unique OTUs that are shared among the different models; pink for OTU0008, salmon for OTU0050, yellow for OTU00367, blue for OTU00110, green for OTU00361 and red for OTU00882. 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.

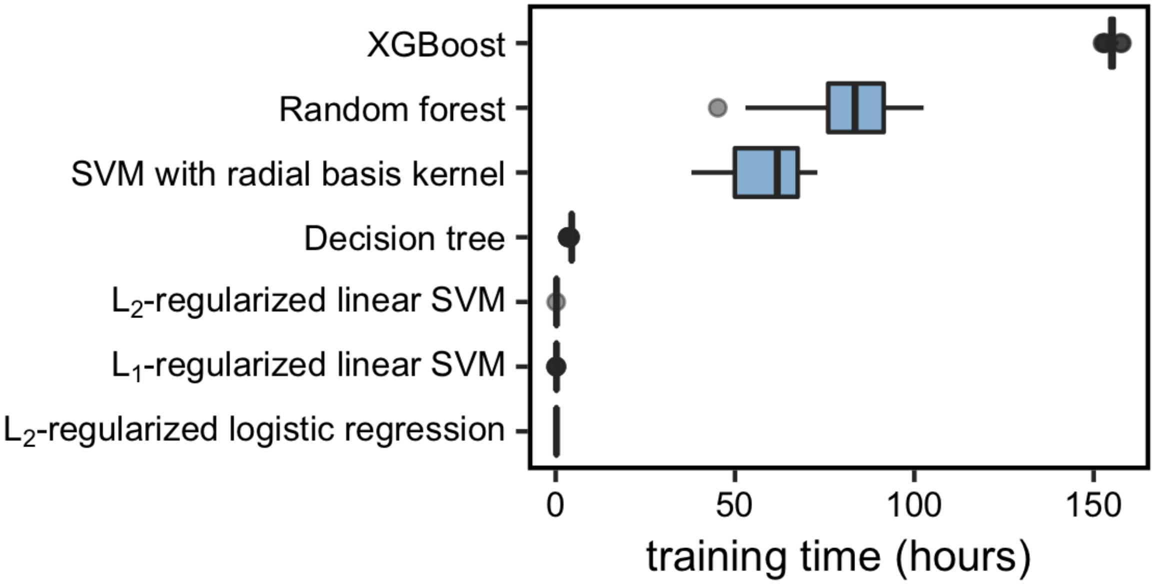

Figure 5. Computational efficiency of seven ML models. The training times for of each data-split showed the differences in computational efficiency of the seven models. The median training time in hours was the highest for XGBoost and shortest for L1-regularized SVM with linear kernel. The boxplot shows 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.

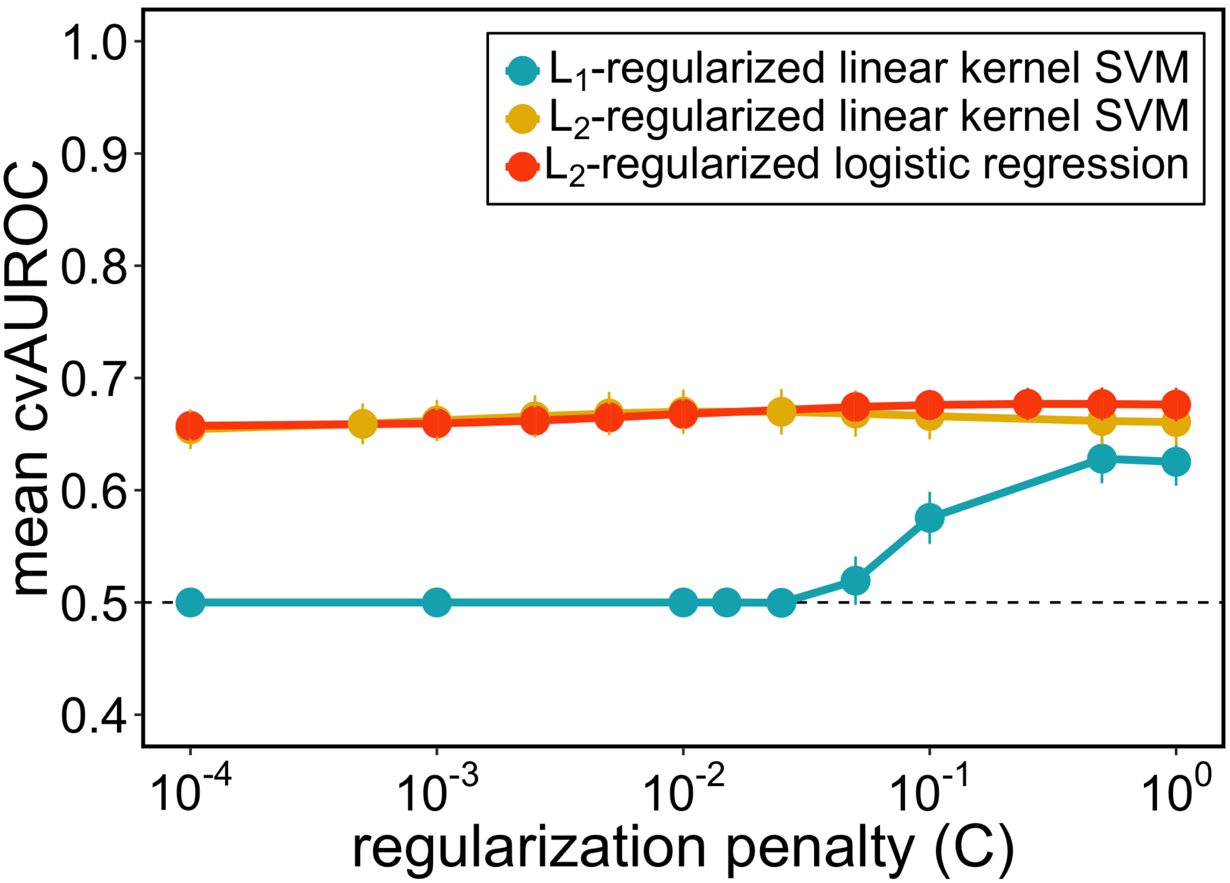

Figure S1. Hyperparameter setting performances for linear models. (A) L2 logistic regression (B) L1 SVM with linear kernel (C) L2 SVM with linear kernel mean cross-validation AUROC values when different hyperparameters are used in training the model. The differences in AUROC values when hyperparameters change show that hyperparameter tuning is a crucial step in building a ML model.

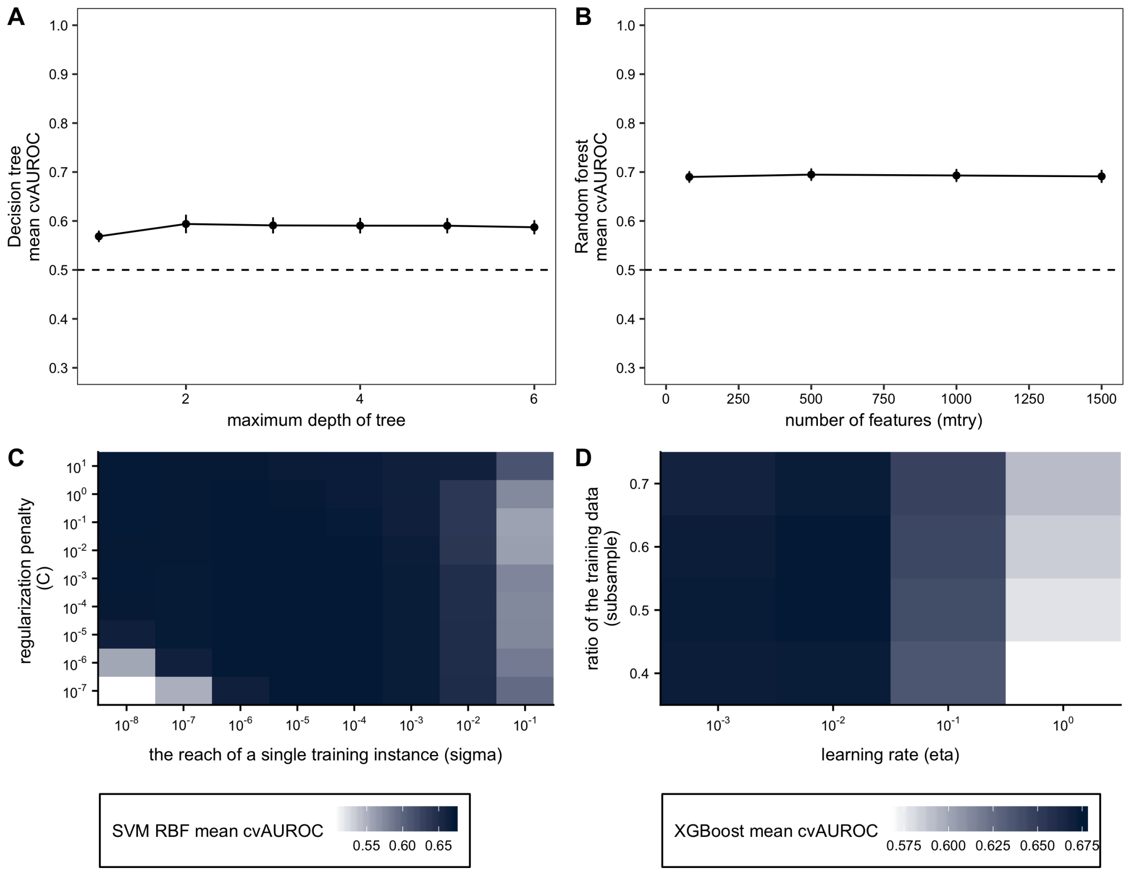

Figure S2. Hyperparameter setting performances for non-linear models. (A) Decision tree (B) Random forest (C) SVM with radial basis kernel (D) XGBoost mean cross-validation AUROC values when different hyperparameters are used in training the model. The differences in AUROC values when hyperparameters change show that hyperparameter tuning is a crucial step in building a ML model.

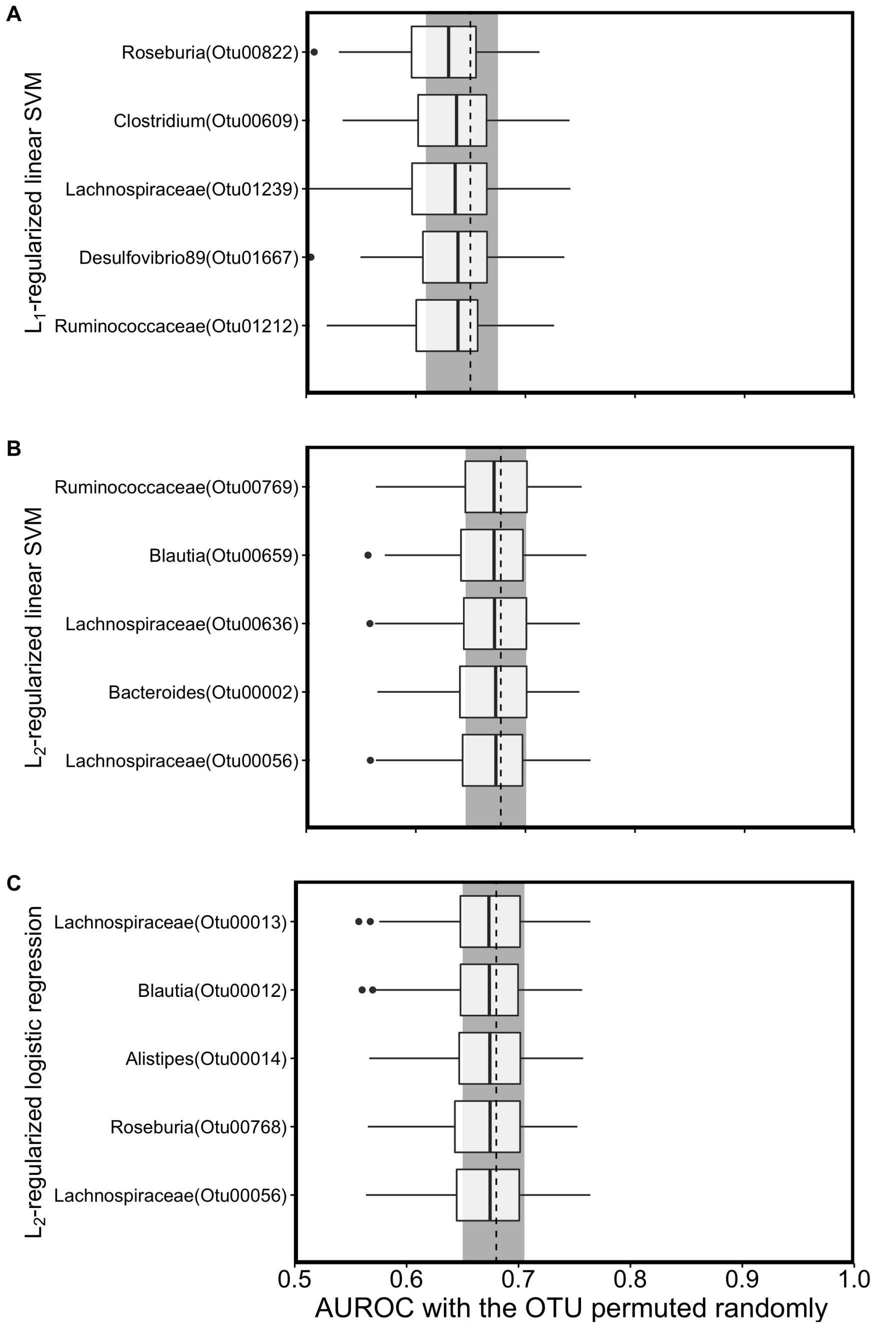

Figure S3. Interpretation of the linear ML models with permutation importance. (A) L1-regularized SVM with linear kernel (B) L2-regularized SVM with linear kernel and (C) L2-regularized logistic regression were interpreted using permutation importance using held-out test set. The gray rectangle and the dashed line show the IQR range and median of the base testing AUROC without any permutation performed. Abbreviations: SVM, support vector machine; OTU, Operational Taxonomic Unit; RBF, radial basis kernel; OTU, Operational Taxonomic Unit.

Table 1: Characteristics of the machine learning models in our comparative study.

Model	Description	Linearity	Interpretability	Refs.
Logistic regression	A predictive regression analysis when the dependent variable is binary.	Linear	Interpretable	36
SVM with linear kernel	A classifier that is defined by an optimal linear seperating hyperplane that discriminates between labels.	Linear	Interpratable	37
SVM with radial basis kernel	A classifier that is defined by an optimal gaussian seperating hyperplane that discriminates between labels.	Non-linear	Explainable*	38
Decision tree	A classifier that sorts samples down from the root to the leaf node where an attribute is tested to discriminate between labels	Non-linear	Interpretable	39
Random forest	A classifier that is a decision tree ensemble that grow randomly with subsampled data.	Non-linear	Explainable*	40−41
XGBoost	A classifier that is a decision tree ensemble that grow with additive training.	Non-linear	Explainable*	42−43

*Explainable models are not inherently interpretable but can be explained with post-hoc analyses.

Table 2: An aspirational rubric for evaluating the rigor of ML practices.

Practice	Good	Better	Best
Problem definition	Have we clearly stated the ML task? Do we have a priori hypotheses? Do we know the predictions a domain expert would make manually?	Do we know the motivation for solving the problem? How much interpretability does the problem need?	Do we know our data? Do we know the confounding variables?
Model selection	Do we know the candidate algorithms for the ML problem?	Do we know our computational resources to fully train each model?	How much interpretibility does the problem need? How much each candidate algorithm can provide?
ML pipeline preparation	Do we have an held-out test dataset?	Have we tested our model on many different held-out datasets?	Have we tuned our model hyperparameters in cross-validation?
Hyperparameter selection	Do we know the different hyperparameters each model can use and why?	Did we use historically effective hyperparameters?	Did we search the full grid space and optimized our model?
Model evaluation	Have we chosen an appropriate metric to evaluate predictive performance?	Have we reported the predictive performance on a held-out test data?	Have we provided an average predictive performance of many model runs?
Model interpretation	Do we know if our model is interpretable?	If the model is not interpretable, do we know how to explain it? Have we checked for the effect of confounding variables?	Have we generated new hypotheses based on model interpretation to test model results?

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. **Thaiss CA**, **Itav S**, **Rothschild D**, **Meijer MT**, **Levy M**, **Moresi C**, **Dohnalová L**, **Braverman S**, **Rozin S**, **Malitsky S**, **Dori-Bachash M**, **Kuperman Y**, **Biton I**, **Gertler A**, **Harmelin A**, **Shapiro H**, **Halpern Z**, **Aharoni A**, **Segal E**, **Elinav E**. 2016. Persistent microbiome alterations modulate the rate of post-dieting weight regain. Nature **540**:544–551. doi:10.1038/nature20796.

13. **Dadkhah E**, **Sikaroodi M**, **Korman L**, **Hardi R**, **Baybick J**, **Hanzel D**, **Kuehn G**, **Kuehn T**, **Gillevet PM**. 2019. Gut microbiome identifies risk for colorectal polyps. BMJ Open Gastroenterology **6**:e000297. doi:10.1136/bmjgast-2019-000297.

14. **Flemer B**, **Warren RD**, **Barrett MP**, **Cisek K**, **Das A**, **Jeffery IB**, **Hurley E**, **O'Riordain M**, **Shanahan F**, **O'Toole PW**. 2018. The oral microbiota in colorectal cancer is distinctive and predictive. Gut **67**:1454–1463. doi:10.1136/gutjnl-2017-314814.

15. **Dai Z**, **Coker OO**, **Nakatsu G**, **Wu WKK**, **Zhao L**, **Chen Z**, **Chan FKL**, **Kristiansen K**, **Sung JJY**, **Wong SH**, **Yu J**. 2018. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. Microbiome **6**:70. doi:10.1186/s40168-018-0451-2.

16. **Montassier E**, **Al-Ghalith GA**, **Ward T**, **Corvec S**, **Gastinne T**, **Potel G**, **Moreau P**, **Cochetiere MF de la**, **Batard E**, **Knights D**. 2016. Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. Genome Medicine **8**:49. doi:10.1186/s13073-016-0301-4.

17. **Papa E**, **Docktor M**, **Smillie C**, **Weber S**, **Preheim SP**, **Gevers D**, **Giannoukos G**, **Ciulla D**, **Tabbaa D**, **Ingram J**, **Schauer DB**, **Ward DV**, **Korzenik JR**, **Xavier RJ**, **Bousvaros A**, **Alm EJ**. 2012. Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. PLOS ONE **7**:e39242. doi:10.1371/journal.pone.0039242.

18. **Mossotto E**, **Ashton JJ**, **Coelho T**, **Beattie RM**, **MacArthur BD**, **Ennis S**. 2017. Classification of paediatric inflammatory bowel disease using machine learning. Scientific Reports **7**. doi:10.1038/s41598-017-02606-2.

19. **Ai L**, **Tian H**, **Chen Z**, **Chen H**, **Xu J**, **Fang J-Y**. 2017. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. Oncotarget **8**:9546–9556. doi:10.18632/oncotarget.14488.

20. **Wong SH**, **Kwong TNY**, **Chow T-C**, **Luk AKC**, **Dai RZW**, **Nakatsu G**, **Lam TYT**, **Zhang L**, **Wu JCY**, **Chan FKL**, **Ng SSM**, **Wong MCS**, **Ng SC**, **Wu WKK**, **Yu J**, **Sung JJY**. 2017. Quantitation of faecal fusobacterium improves faecal immunochemical test in detecting advanced colorectal neoplasia. Gut **66**:1441–1448. doi:10.1136/gutjnl-2016-312766.

21. **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.

22. **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).

23. **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].

24. **Rudin C**. 2018. Please stop explaining black box models for high stakes decisions. arXiv:181110154 [cs, stat].

25. **Rudin C**, **Ustun B**. 2018. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. Interfaces **48**:449–466. doi:10.1287/inte.2018.0957.

26. **Knights D**, **Parfrey LW**, **Zaneveld J**, **Lozupone C**, **Knight R**. 2011. Human-associated microbial signatures: Examining their predictive value. Cell Host Microbe **10**:292–296. doi:10.1016/j.chom.2011.09.003.

27. **Miller T**. 2017. Explanation in artificial intelligence: Insights from the social sciences. arXiv:170607269 [cs].

28. **Statnikov A**, **Henaff M**, **Narendra V**, **Konganti K**, **Li Z**, **Yang L**, **Pei Z**, **Blaser MJ**, **Aliferis CF**, **Alekseyenko AV**. 2013. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome **1**:11. doi:10.1186/2049-2618-1-11.

29. **Knights D**, **Costello EK**, **Knight R**. 2011. Supervised classification of human microbiota. FEMS Microbiology Reviews **35**:343–359. doi:10.1111/j.1574-6976.2010.00251.x.

30. **Wirbel J**, **Pyl PT**, **Kartal E**, **Zych K**, **Kashani A**, **Milanese A**, **Fleck JS**, **Voigt AY**, **Palleja A**, **Ponnudurai R**, **Sunagawa S**, **Coelho LP**, **Schrotz-King P**, **Vogtmann E**, **Habermann N**, **Niméus E**, **Thomas AM**, **Manghi P**, **Gandini S**, **Serrano D**, **Mizutani S**, **Shiroma H**, **Shiba S**, **Shibata T**, **Yachida S**, **Yamada T**, **Waldron L**, **Naccarati A**, **Segata N**, **Sinha R**, **Ulrich CM**, **Brenner H**, **Arumugam M**, **Bork P**, **Zeller G**. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nature Medicine **25**:679. doi:10.1038/s41591-019-0406-6.

31. **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.

32. **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.

33. **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.00073-17.

34. **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.

35. **Li L**, **Jamieson K**, **DeSalvo G**, **Rostamizadeh A**, **Talwalkar A**. 2016. Hyperband: A novel bandit-based approach to hyperparameter optimization. arXiv:160306560 [cs, stat].