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 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

As the number of people represented in human microbiome datasets grow, there is an increasing desire to use microbiome data to diagnose diseases. However, the structure of the human microbiome is remarkably variable between individuals to the point where it is often difficult to identify the bacterial populations that are associated with diseases using traditional statistical models. This variation is likely due to the ability of many bacterial populations to fill the same niche such that different populations cause the same disease in different individuals. Furthermore, a growing number of studies have shown that it is rare for a single bacterial species to be associated with a disease. Instead, subsets of the microbiome account for differences in health. Traditional statistical approaches do not adequately account for the variation in the human microbiome and typically consider the protective or risk effects of each bacterial population individually. Recently, machine learning models have grown in popularity among microbiome researchers because of the large amount of data that can now be generated and because the models are effective at accounting for the interpersonal microbiome variation and the ecology of the disease.

ML models are useful for understanding the variation in the structure of existing data and to apply that knowledge to make predictions about new data. Researchers have used ML models to diagnose and understand the ecological basis of diseases such as liver cirrhosis, colorectal cancer, inflammatory bowel diseases (IBD), obesity, and type 2 diabetes (1–16, 16–18). The task of diagnosing an individual with high confidence relies on a ML model that is built with rigorous methods. However, there are common methodological problems across many of these studies that need to be addressed as the field progresses. These include a lack of transparency in which methods are used and how these methods are implemented; developing and evaluating models without a separate held-out test data; large variation between the predictive performance on different folds of cross-validation; and large variation between cross-validation and testing performances. Nevertheless, the microbiome field is making progress to avoid some of these pitfalls including validating their models on independent datasets (7, 18, 19) and introducing analysis frameworks to better use ML tools (20–23). More work is needed to further improve reproducibility and minimize over-optimism for model performance.

Among microbiome researchers the lack of transparency in justifying a modelling approach has been due to an implicit assumption that more complex models are better because they are more complex. This has resulted in a trend towards using models such as random forest and neural networks (2, 11, 24–26) over simpler models such as logistic regression or other linear models (18, 22, 27). Although the more complex models may be better at incorporating non-linear relationships or yield better predictions, they are considered to be black box models because they are not inherently interpretable. These models require post hoc explanations to quantify the importance of each feature in making a prediction and they do not show the structure of how the features are used. Depending on the application of the model, researchers may choose to use different modeling approaches. For example, researchers trying to identify the populations causing a disease would likely want a more interpretable model whereas clinicians may emphasize performance. Although one may feel that they are sacrificing interpretability for performance, that tradeoff may be minimal (28, 29). Regardless, it is important for researchers to articulate why they have selected a specific modelling approach or even compare multiple approaches in the same study.

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 7 modeling approaches with the same dataset and pipeline. We built three linear models with different forms of regularization: L2-regularized logistic regression and L1 and L2-regularized support vector machines (SVM) with a linear kernel. We also built 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. To demonstrate the performance of these modeling approaches and our pipeline, we used data from a previously published study that sought to classifiy individuals as having normal colons or colonic lesions based on the 16S rRNA gene sequences collected from fecal samples (3). This dataset was selected because it is a relatively large collection of individuals (N=490) connected to a clinically significant disease where there is ample evidence that the disease is driven by variation in the microbiome (1, 3, 4, 30). With this dataset we developed a framework that implements a ML pipeline that can be used for any modeling approach, evaluates predictive performance, and demonstrates how to interpret these models. This framework can be easily applied to other host-associated and environmental microbiome datasets.

Results

Model selection and pipeline construction We established a ML pipeline where we trained and validated each of the seven models using a common approach that is based on standard methods within the ML community (REFS)[Figure 1].

First, we randomly split the data into training and test sets so that the training set consisted of 80% of the full dataset while the test set was composed of the remaining 20% of the data [Figure 1]. To maintain the distribution of controls and cases that was found with the full dataset, we performed stratified splits. For example, our full dataset included 490 individuals. Of these, 261 had normal colons (53%) and 229 had a screen relevant neoplasia (SRN; 46.7%). A training set included 393 individuals, of which 209 had an SRN (53%), while the test set was composed of 97 individuals of which 52 had an SRN (54%). The training data was used to build the models and the test set was used for evaluating predictive performance.

Second, we trained seven different models using the training data [Table 1]. We selected models with different classification algorithms and regularization methods. Regularization is a technique that discourages overfitting by penalizing the model for learning the training data too well. 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 led to a sparser solution (i.e., force many coefficients to zero). To explore the potential for non-linear relationships among features to improve classification, we trained tree-based models including decision tree, random forest, and XGBoost and we trained an SVM model with non-linear kernel.

Third, fitting of these models require selecting appropriate hyperparameters. Hyperparameters are the rules that are learned from the training set in a classification algorithm. For example, in the linear models the regularization term (C) is a hyperparameter that indicates the penalty for overfitting. Similar to regularization term C, all hyperparameters are tuned to find the best model. We selected hyperparameters by performing 100 five-fold cross-validation (CV) repeats on the training set [Figure 1]. The five-fold CV was also stratified to maintain the overall case and control distribution. We chose the best hyperparameter values for each model based on its CV predictive performance 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. To select the best performing hyperparameter, we performed a full grid search for hyperparameter settings when training our models. 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.

Finally, we trained the full training dataset with the selected hyperparameter values and applied the model to the held-out data to evaluate the testing predictive performance of each 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 performance of seven models to classify individuals as having normal colons or SRNs [Figure 2]. The random forest model had significantly higher test AUROC values than the other models for detecting SRNs (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 and 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]. Interestingly, these results demonstrate that the most complex model (XGBoost) did not have the best performance and that the most interpretable models (L2-regularlized logistic regression and linear SVM) performed nearly as well as random forest.

To evaluate the generalizability of each model, we compared the median cross-validation AUROC to the median testing AUROC. If the difference between the cross-validation and testing AUROCs was large, then that would indicate that the models were overfit. The difference in median AUROCs was NA in L1-regularized SVM with linear kernel, followed by SVM with radial basis function kernel and decision tree with a difference of NA and NA, respectively [Figure 2]; however, these differences are relatively small and would not indicate a problem with overfitting.

To evaluate the risk for over-optimism of each model, we calculated the range of AUROC values for each model using 100 splits. The range among the testing AUROC values within each model varied by 0.23 on average across the seven models. If we had only done a single split, then there is a risk that we could gotten lucky or unlucky with the performance of the model. For instance, the lowest AUROC value of the random forest model was 0.593 whereas the highest was 0.81. These results showed that depending on the data-split, the testing AUROC values showed great variability [Figure 2]. Therefore, it is important to employ the hierarchical data splits that were included in our pipeline to minimize the risk of over-optimism.

To show the effect of sample size on model generalizability, we compared cross-validation AUROC values of L2-regularized logistic regression and random forest models when we subsetted our original study design with 490 subjects to 15, 30, 60, 120, and 245 subjects [Figure S3]. The range among the cross-validation AUROC values within both models at lower sample sizes were much larger than when the full collection of samples was used to train and validate the models. These results showed that because the microbiome data had many features (6920 OTUs), it was important to train the models using appropriate sample sizes to avoid problems with generalizability and over-optimism. Furthermore, it was encouraging that even for a small number of samples, the interquartile range included the median AUROC values for the larger subsetted datasets.

Interpretation of each ML model. Interpretability is the degree to which humans can understand the reasons behind a model prediction (31). 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. ML models decrease in interpretability as they increase in complexity. In this study we used two methods to help interpret our models.

First, we interpreted the feature importance of the linear models (L1 and L2-regularized SVM with linear kernel and L2-regularized logistic regression) using the median rank of absolute feature weights for each OTU [Figure 3]. We also reviewed the signs of feature weights to determine whether an OTU is associated with classifying a subject as being healthy or having an SRN - negative sign indicated being healthy and positive sign indicated having an SRN. It was encouraging that many of the highest ranked OTUs were shared across these three models, (e.g. OTU 50, 426, 609, 822, 1239). The benefit of this approach was that the results of the analysis were based on the trained model parameters and provided information regarding the sign and magnitude of the impact of each OTU. However, this approach was only possible with linear models.

Second, to analyze non-linear models we interpreted the feature importance using permutation importance. Whereas the absolute feature weights were determined from the trained models, here we measured importance using the held-out test data. Permutation importance analysis is a posthoc explanation of the model where we randomly permuted non-correlated features individually and groups of perfectly correlated features across the two groups in the held-out test data. We then calculated how much the predictive performance of the model (i.e testing AUROC values) decreased when each OTU or group of OTUs was randomly permuted. We ranked the OTUs based on how much the median testing AUROC decreased when it was permuted; the OTU with the largest decrease ranked highest [Figure 4]. Among the twenty OTUs with the largest impact for each of these models, there was only one OTU (OTU 822) that was shared among all of the models; however, we found three OTUs (OTU 58, 110, 367) that were important in each of the tree-based models. Similarly, the random forest and XGBoost models, shared four of the most important OTUs (OTU 2, 12, 361, 477). Permutation analysis results also revealed that with the exception of the decision tree model, removal of any individual OTU had a minimal impact on model performance. For example, if OTU 367 were permuted across the diagnoses from the decision tree model, the median AUROC dropped from 0.601 to 0.525. In contrast, if we permuted the same OTU from the random forest model, the AUROC only dropped from 0.695 to 0.68. Effectively, the complexity of the communities was more fully represented in the better performing models (22, 32). At least in this case, it was not possible to distinguish between health and disease using a single OTU. While permutation analysis allowed us to gauge the importance of an OTU, the analysis was post-hoc (i.e. done using the test data) and these results did not allow us to directly interrogate the models to know whether an OTU is associated with classifying a subject as being healthy or having an SRN.

To further highlight the differences between the two interpretation methods, we used permutation importance to interpret the linear models [Figure S4]. When we analyzed the L1-regularized SVM with linear kernel model using feature rankings based on weights [Figure 3] and permutation importance [Figure S4], 17 out of the 20 top OTUs (e.g. OTU 609, 822, 1239) were deemed important by both approaches. Similarly, for the L2-regularized SVM and L2-regularized logistic regression, 9 and 12 OTUs, respectively, were shared among the two approaches. These results indicate that both approaches are consistent in selecting the most important OTUs.

The computational efficiency of each ML model. We compared the training times of the seven ML models. As expected, the training times increased with the complexity of the model and the number of tuned hyperparameter settings. Also, the linear models trained faster than non-linear models [Figures S1-S2; Figure 5]. When we subsetted the size of the training dataset, we observed a linear relationship between the size of the dataset and the training times for L2-regularized logistic regression and random forest models [Figure S5].

Discussion

There is a growing awareness that many human diseases and environmental processes are not driven by a single organism but are the product of multiple bacterial populations. Traditional statistical approaches are useful for identifying those cases where a single organism is associated with a process. In contrast, ML methods offer the ability to incorporate the structure of the microbial communities as a whole to classify them into different categories such as coming from a patient who is healthy or has SRNs. If it is possible to classify communities reliably, then ML methods also offer the ability to identify those microbial populations within the communities that are responsible for the classification. However, the application of ML in microbiome studies is still in its infancy and the field still needs to develop a better understanding of different ML methods, their strengths and weaknesses, and how to implement them.

To address these needs, we developed a framework to train rigorous, transparent, and reproducible models. We benchmarked seven ML models and showed that we can create models that are inherently interpretable and easily trained without losing predictive performance. In terms of predictive performance, the random forest model had the best testing AUROC values compared to the other six models. However, the second-best model was L2-regularized logistic regression with a median AUROC difference of only 0.015 compared to random forest. While random forest took 83.2 hours to train, L2-regularized logistic regression trained in 12 minutes. In terms of interpretability, random forest was a complex ML model and could only be explained using post-hoc methods such as permutation importance. On the other hand, L2-regularized logistic regression was easier to interpret by ranking the OTUs based on their feature weights and reviewing the signs of these weights. Comparing many different models showed us that the most complex model is not necessarily the best model for our ML task.

As we set out to select the best model, we established a pipeline that can be generalized to any modeling method that uses 16S rRNA sequence counts to predict a binary health outcome. We performed a random datasplit to create a training set (80% of the data) and a held-out test set (20% of the data), which we used to evaluate predictive performance. We repeated this datasplit 100 times to measure predictive performance. During the training, we tuned the model hyperparameters with a repeated five-fold cross-validation. Despite the high number of features microbiome datasets typically have, the models we built with this pipeline were generalizable as shown by the similar AUROC values from the cross-validation and testing.

We highlighted the importance of model interpretation to gain greater biological insights into microbiota-associated diseases. In this study we showcased two different interpretation methods: ranking each OTU by (i) their absolute weights in the trained models and (ii) their impact on the predictive performance based on permutation importance. Human-associated microbial communities have complex correlation structures which create collinearity in the datasets we work with. This can hinder our ability to reliably interpret models because the feature weights of correlated OTUs are influenced by one another (33). For example if one OTU is highly correlated with another OTU, only one of these OTUs would have a large feature weight thus hiding the importance of its correlated OTU. To capture all important features, once we identify highly ranked OTUs, we should review their relationships with other OTUs. These relationships will help us generate new hypotheses about the ecology of the disease and test them with follow-up experiments. When we used permutation importance, we took collinearity into consideration by grouping correlated OTUs to determine their impact as a group. We grouped OTUs that had a perfect correlation with each other however, we can reduce the correlation threshold to further investigate the relationships among features. It is important to know the correlation strutures of the data to avoid missinterpreting the models. This is likely to be a particular problem with shotgun metagenomic datasets where collinearity will be more pronunced due to many genes being correlated with one another because they come from the same chromosome. Therefore, to identify the true underlying microbial factors of a disease, it is crucial to do correlation analyses and further experimentation for biological validation.

In this study, we did not consider all possible modeling approaches. However, the principles highlighted throughout this study apply to all ML modeling tasks with microbiome data. For example, we did not evaluate multicategory classification methods to predict non-binary outcomes. For example, we could have trained models to differentiate between people with normal colons and those with adenomas or carcinomas (k=3 categories). We did not perform this analysis because the clinically relevant diagnosis grouping was between patients with normal colons and those with SRNs. Furthermore, as number of categories to classify increase, more samples are required for each category to train a model. We also did not use regression-based analyses to predict a non-categorical outcome. We have previously used such an approach to train random forest models to predict fecal short-chain fatty acid concentrations based on microbiome data (34). Our analysis was also limited to shallow learning methods and did not explore deep learning methods such as neural networks. Deep learning methods hold great promise (11, 35, 36) but microbiome datasets often suffer from having many features and small sample sizes, which makes the deep learning models prone to overfitting. These methods are even more complex than random forest and XGBoost and are considered uninterpretable. There is great potential for applying ML approaches to microbiome data.

Our framework gives structure to investigators wanting to train, evaluate, and interpret their own ML models to identify OTUs that might be biologically relevant. However, deploying microbiome-based models to make clinical diagnoses or predictions is a significantly harder and distinct undertaking. For example, we currently lack standardized methods to collect patient samples, generate sequence data, and report clinical data. We are also challenged by the practical constraints of OTU-based approaches. The de novo algorithms commonly in use are slow, require considerable memory, and result in different OTU assignments as new data are added. Finally, we also need independent validation cohorts to test the performance of a diagnostic model. To realize the potential for using ML approaches with microbiome data, it is necessary we direct our efforts to overcome these challenges.

This study highlighted the need to make educated choices at every step of developing a ML model with microbiome data. We created an aspirational rubric that researchers can use to identify potential pitfalls when using ML in microbiome studies and ways to avoid them [Table 2]. We have highlighted the tradeoffs between model complexity and interpretability, the need for cross-validation to tune hyperparameters, the utility of held-out test sets for evaluating predictive performance, and the importance of considering correlation structures in datasets for reliable interpretation. Furthermore, we underscored the importance of proper experimental design and methods to 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 (32). 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 (37–39).

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 colorectal health of the patients. 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 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 (40).

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/.

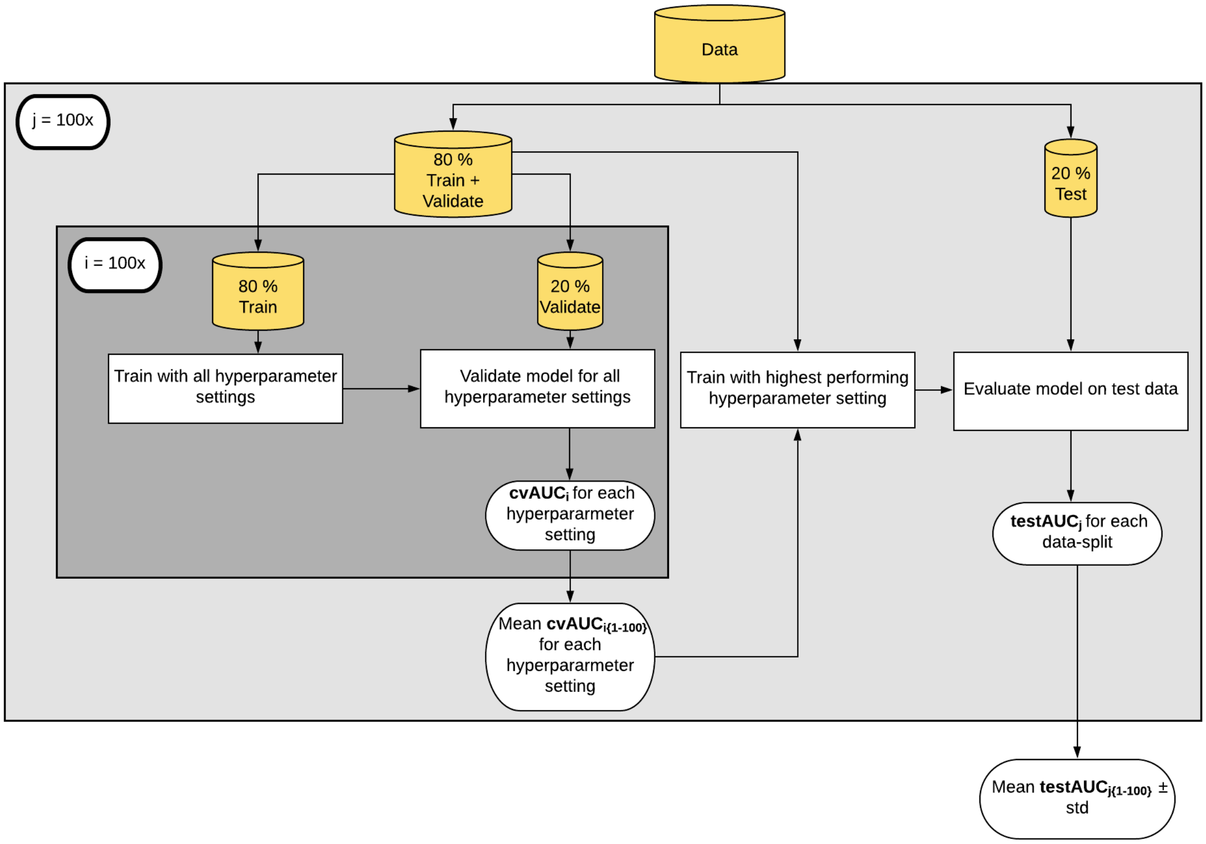
Data
j = 100x
80 %
Train +
Validate
20 %
Test
i = 100x
80 %
Train
20 %
Validate
Train with all hyperparameter settings
Validate model for all hyperparameter settings
cvAUCi for each hyperpararmeter setting
Mean cvAUCi{1-100} for each hyperpararmeter setting
Train with highest performing hyperparameter setting
Evaluate model on test data
testAUCj for each data-split
Mean testAUCj{1-100} ± 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

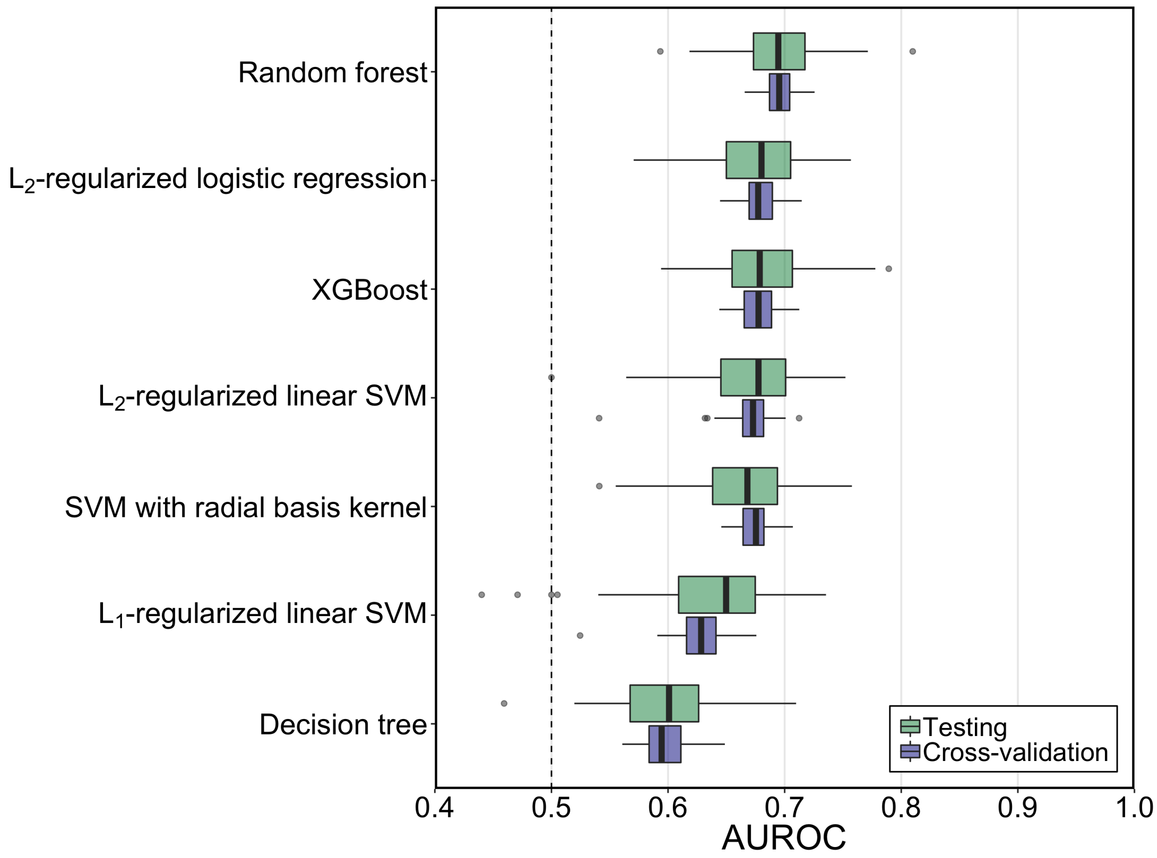

Figure 2. Generalization and classification performance of ML models using AUROC values of all cross validation and testing performances. The median AUROC for diagnosing individuals with SRN using bacterial abundances was higher than chance (depicted by horizontal line at 0.50) for all the ML models.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

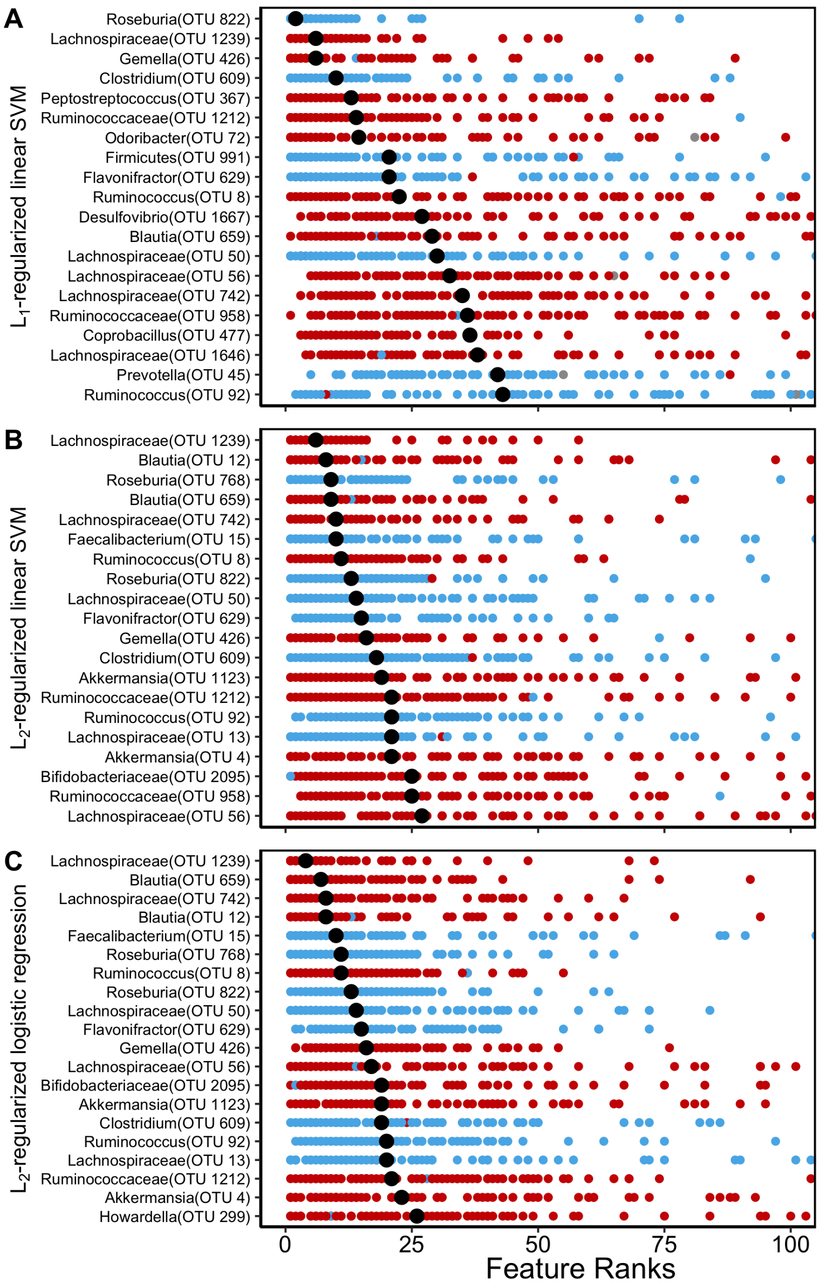

Figure 3. Interpretation of the linear ML models. The absolute feature weights of (A) L2 logistic regression coefficients (B) L1 SVM with linear kernel (C) L2 SVM with linear kernel were ranked from highest rank 1 to 100 for each data-split. The feature ranks of the highest ranked five OTUs based on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

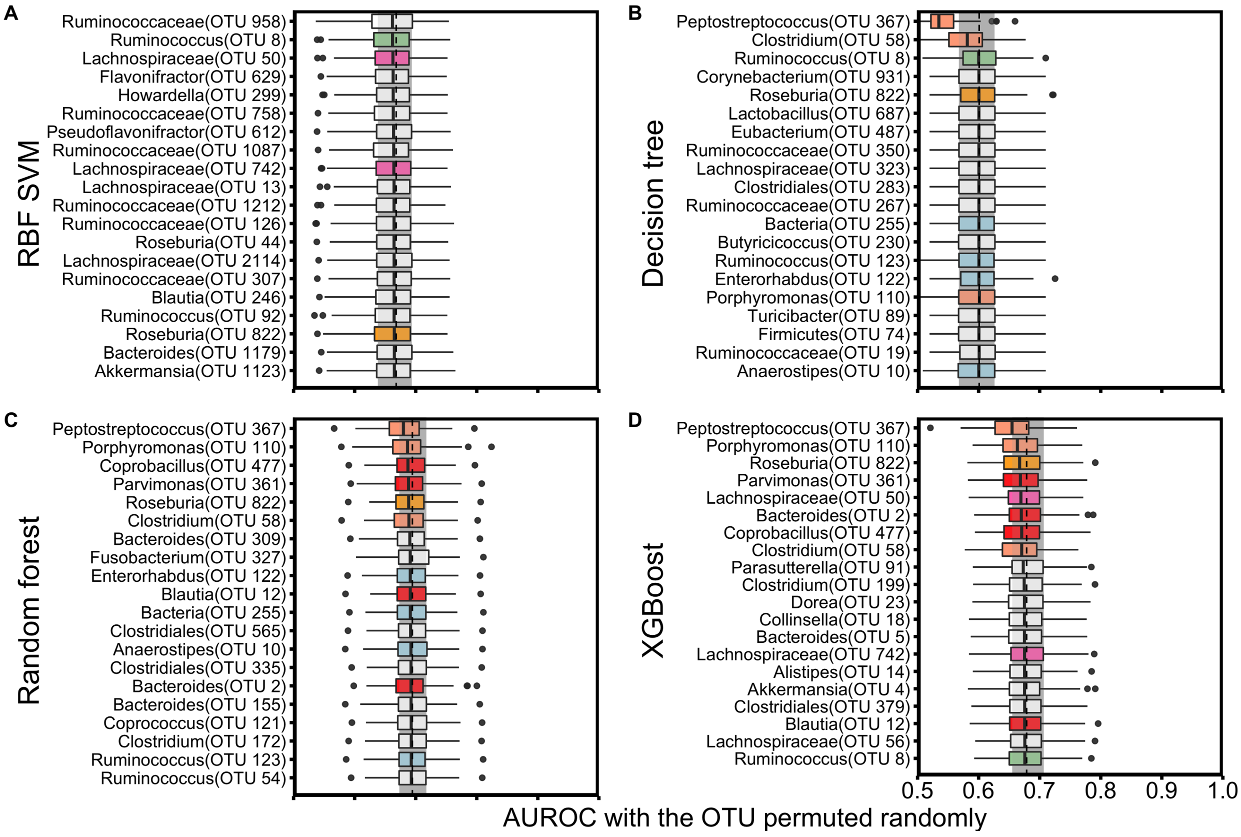

Figure 4. Interpretation of the non-linear ML models. (A) SVM with radial basis kernel (B) decision tree (C) random forest (D) XGBoost feature importances were explained using permutation importance using held-out test set. The gray rectangle and the dashed line show the IQR range and median of the base 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.

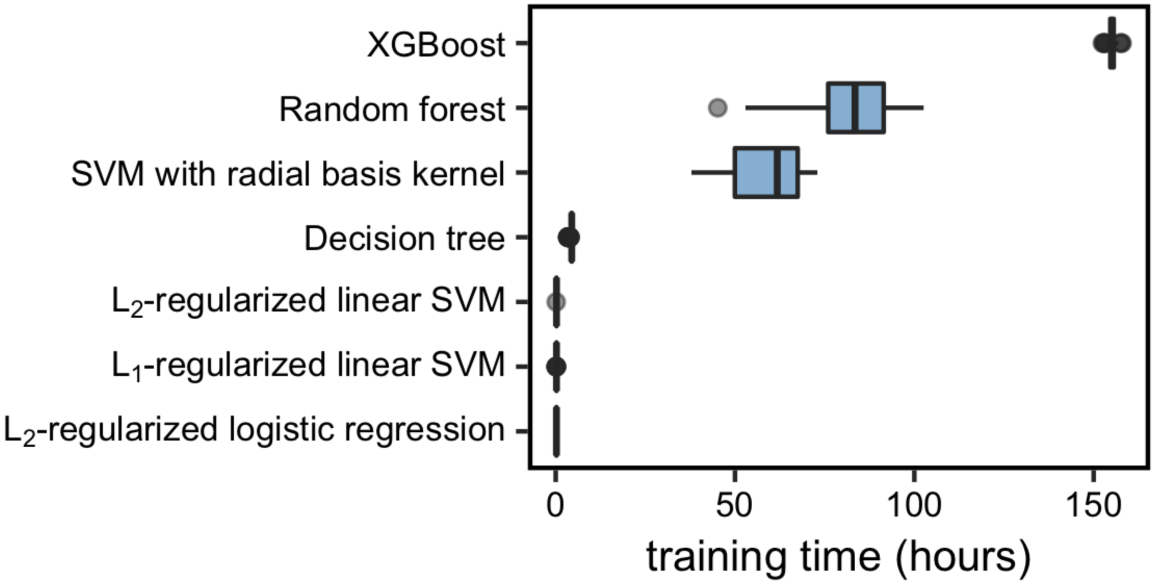

Figure 5. Computational efficiency of seven ML models. The training times for of each data-split showed the differences in computational efficiency of the seven models. The median training time in hours was the highest for XGBoost and shortest for L1-regularized SVM with linear kernel. The boxplot shows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

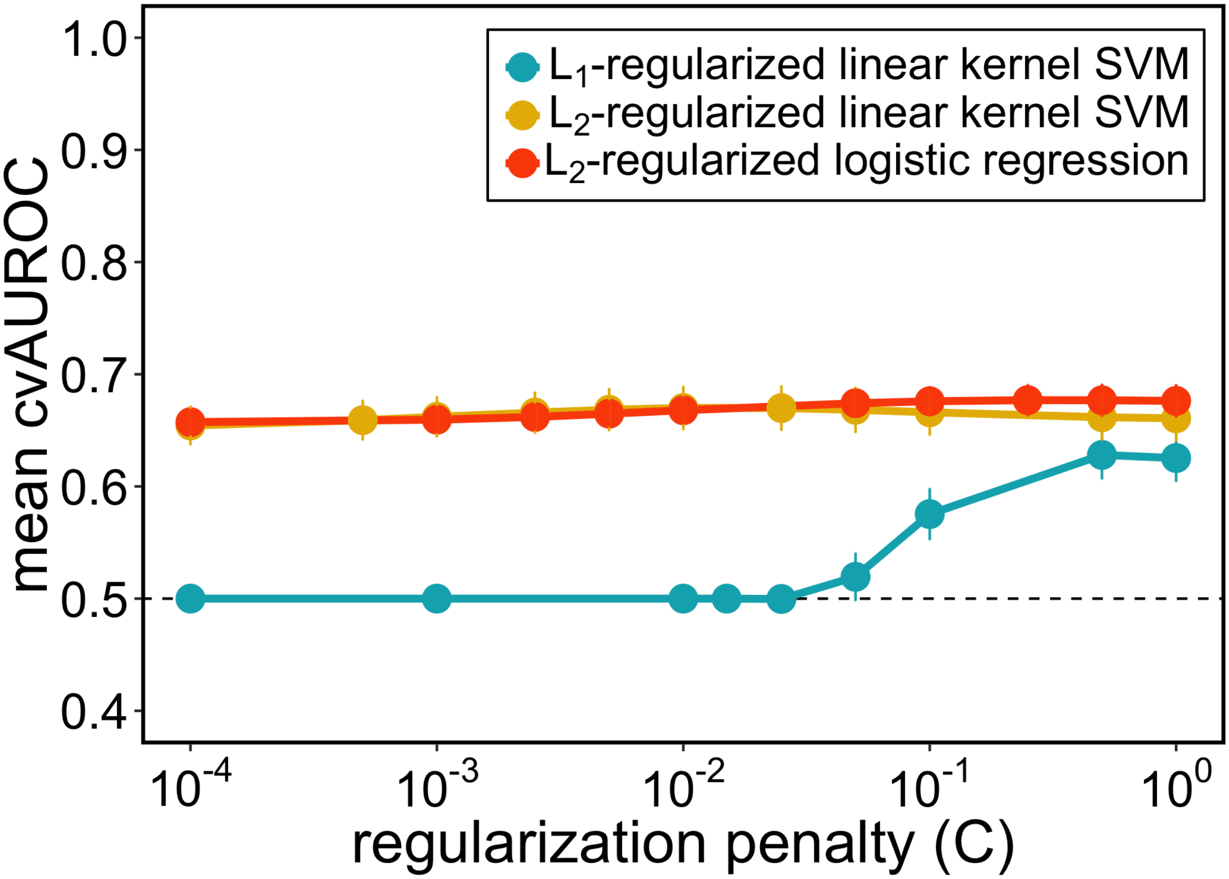

Figure S1. Hyperparameter setting performances for linear models. (A) L2 logistic regression (B) L1 SVM with linear kernel (C) L2 SVM with linear kernel mean cross-validation AUROC values when different hyperparameters are used in training the model. The differences in AUROC values when hyperparameters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

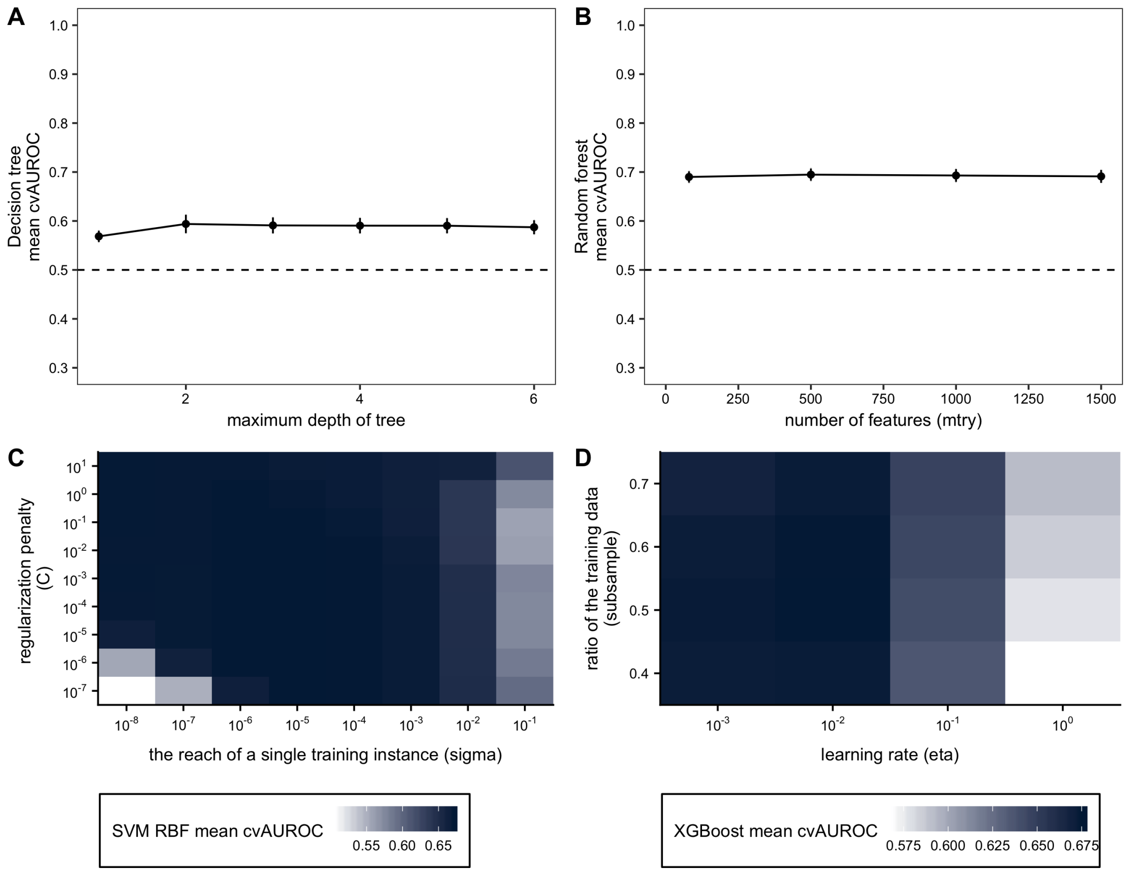

Figure S2. Hyperparameter setting performances for non-linear models. (A) Decision tree (B) Random forest (C) SVM with radial basis kernel (D) XGBoost mean cross-validation AUROC values when different hyperparameters are used in training the model. The differences in AUROC values when hyperparameters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

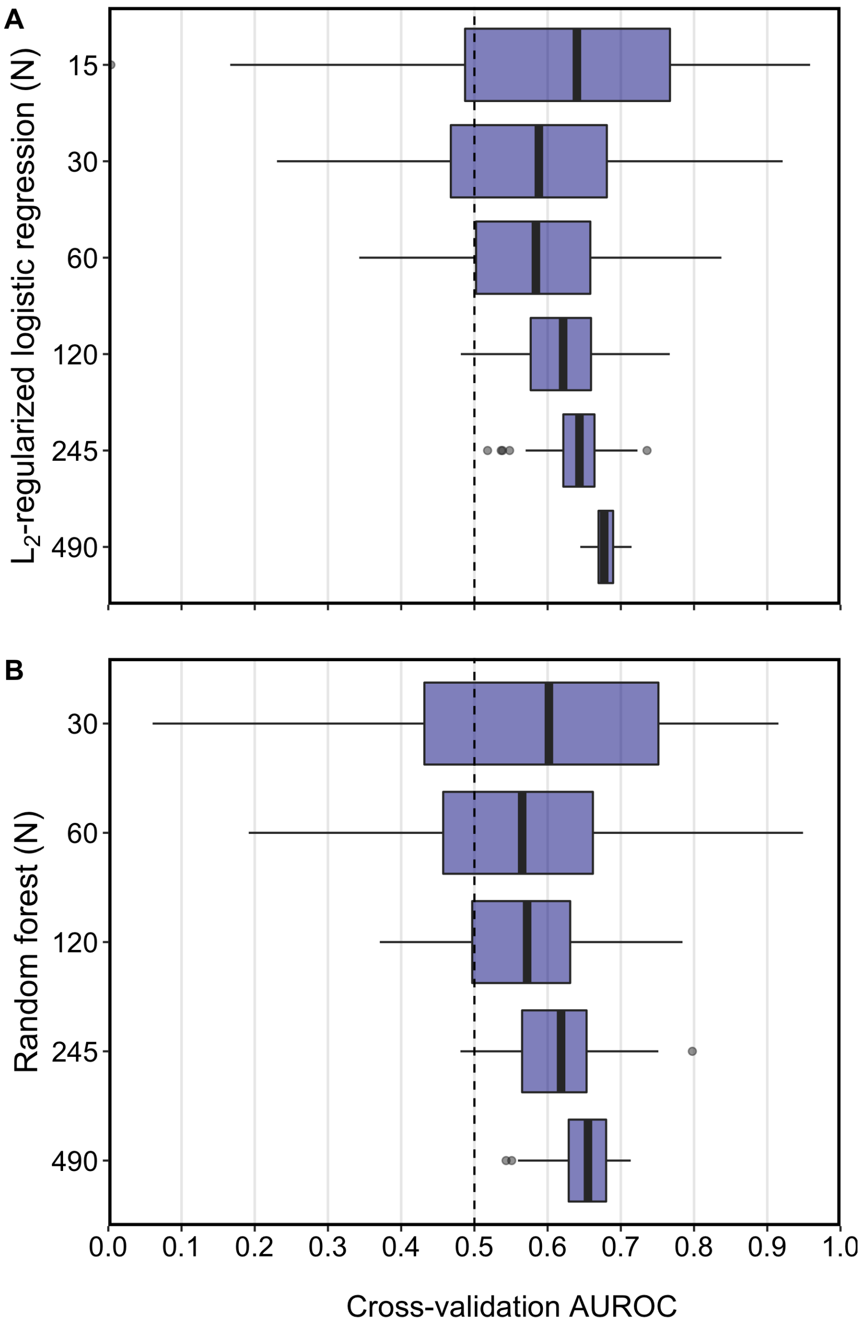

Figure S3. Interpretation of the linear ML models with permutation importance. (A) L1-regularized SVM with linear kernel (B) L2-regularized SVM with linear kernel and (C) L2-regularized logistic regression were interpreted using permutation importance using held-out test set. The gray rectangle and the 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 separating hyperplane that discriminates between labels.	Linear	Interpretable	37
SVM with radial basis kernel	A classifier that is defined by an optimal Gaussian separating 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 interpretability 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. **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.

16. **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.

17. **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.

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. **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.

20. **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.

21. **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.

22. **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.

23. **Vangay P**, **Hillmann BM**, **Knights D**. 2019. Microbiome learning repo (ML repo): A public repository of microbiome regression and classification tasks. Gigascience **8**. doi:10.1093/gigascience/giz042.

24. **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.

25. **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).

26. **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].

27. **Thomas AM**, **Manghi P**, **Asnicar F**, **Pasolli E**, **Armanini F**, **Zolfo M**, **Beghini F**, **Manara S**, **Karcher N**, **Pozzi C**, **Gandini S**, **Serrano D**, **Tarallo S**, **Francavilla A**, **Gallo G**, **Trompetto M**, **Ferrero G**, **Mizutani S**, **Shiroma H**, **Shiba S**, **Shibata T**, **Yachida S**, **Yamada T**, **Wirbel J**, **Schrotz-King P**, **Ulrich CM**, **Brenner H**, **Arumugam M**, **Bork P**, **Zeller G**, **Cordero F**, **Dias-Neto E**, **Setubal JC**, **Tett A**, **Pardini B**, **Rescigno M**, **Waldron L**, **Naccarati A**, **Segata N**. 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nature Medicine **25**:667. doi:10.1038/s41591-019-0405-7.

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

29. **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.

30. **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.

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

32. **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.

33. **Dormann CF**, **Elith J**, **Bacher S**, **Buchmann C**, **Carl G**, **Carré G**, **Marquéz JRG**, **Gruber B**, **Lafourcade B**, **Leitão PJ**, **Münkemüller T**, **McClean C**, **Osborne PE**, **Reineking B**, **Schröder B**, **Skidmore AK**, **Zurell D**, **Lautenbach S**. 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography **36**:27–46. doi:10.1111/j.1600-0587.2012.07348.x.

34. **Sze MA**, **Topçuoğlu BD**, **Lesniak NA**, **Ruffin MT**, **Schloss PD**. 2019. Fecal short-chain fatty acids are not predictive of colonic tumor status and cannot be predicted based on bacterial community structure. mBio **10**:e01454–19. doi:10.1128/mBio.01454-19.

35. **Kocheturov A**, **Pardalos PM**, **Karakitsiou A**. 2019. Massive datasets and machine learning for computational biomedicine: Trends and challenges. Ann Oper Res **276**:5–34. doi:10.1007/s10479-018-2891-2.

36. **Kim M**, **Oh I**, **Ahn J**. 2018. An improved method for prediction of cancer prognosis by network learning. Genes **9**:478. doi:10.3390/genes9100478.

37. **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.

38. **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.

39. **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.

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