

NAME OF THIS STUDY

Running title: INSERT RUNNING TITLE HERE

Begüm D. Topçuoğlu¹, Jenna Wiens², 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 of Michigan, Ann Arbor, MI 49109

1 Abstract

2 Introduction

3 As gut microbiome field continues to grow, there will be an ever-increasing demand for reproducible
4 machine learning methods to analyze microbiome sequence read count data and to determine
5 association with a continuous or categorical phenotype of interest.

6 Colorectal cancer is one of the leading cause of death among cancers in the United States. Early
7 diagnosis increases the chance of survival. However the current diagnostic methods are expensive
8 and invasive. As a less invasive tool, numerous studies use relative abundances of the gut bacteria
9 populations to predict disease progression. Most microbial communities are pretty patchy and the
10 likelihood of a single feature that explains the differences in health is pretty small. It is likely that
11 many biomarkers are needed to account for the patchiness as well as the context dependency of
12 the features.

13 ML use in microbiome literature is a bit like the wild west with lack of clarity over methods,
14 testing, validation, etc. There is a need for guidance on how to properly implement these different
15 methods. We need to emphasize good machine learning practices and pipelines and discuss the
16 reproducibility, robustness and actionability of models.

17 We established a non-leaky pipeline. We performed L1 and L2-regularized logistic regression,
18 Linear SVM, Non-Linear SVM, Decision tree, Random forest, XGBoost and Feed Forward Neural
19 Net classification models. We evaluated the classification performance of different machine learning
20 methods. We also want to discuss the reproducibility, robustness, actionability, interpretability and
21 susceptibility to overfitting of each method.

22 Generalisation Performance of each model. Is there a maximum threshold of prediction with all
23 these methods? Does an increase in model complexity improve predictability? Synthesis statement
24 regarding modeling 16S microbiome data

25 **Results and Discussion**

26 **Conclusions**

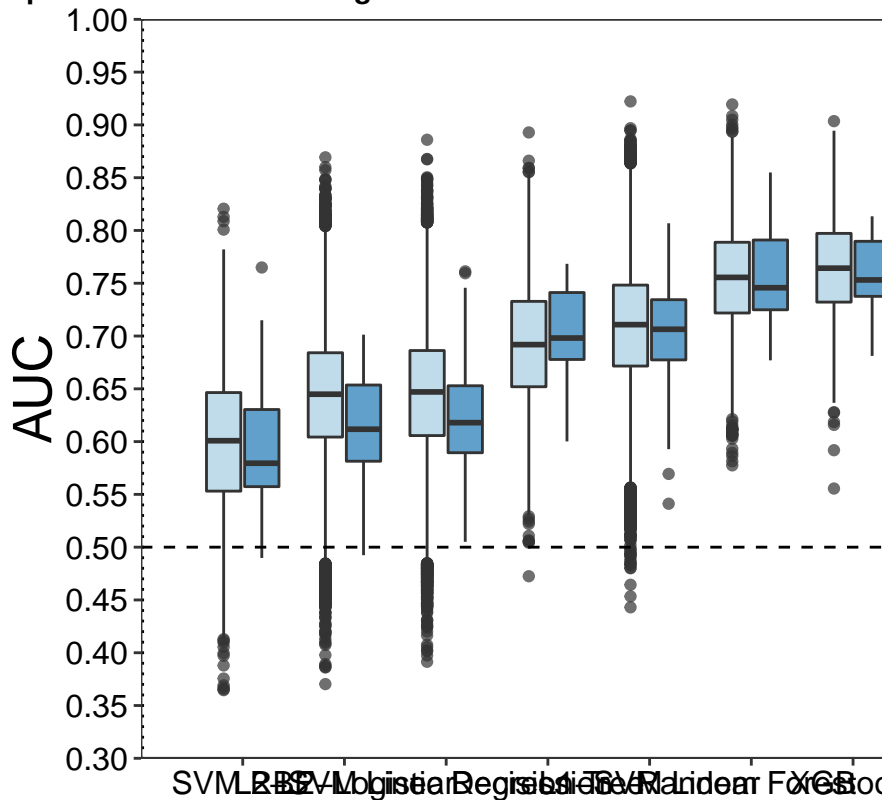
27 **Materials and Methods**

Table 1: Optimized hyper-parameters, pre-processing and cross-validation methods and software implementation of the classification algorithms

| Method | Parameter | Cross Validation | Epoch | Scaler | Sklearn Function |
|----------------------|--|---------------------|-------|----------|------------------------|
| Logistic Regression | C | 5-fold, 100-repeats | 100 | MinMax | LogisticRegression |
| L1 SVM Linear Kernel | C | 5-fold, 100-repeats | 100 | Standard | LinearSVC |
| L2 SVM Linear Kernel | C | 5-fold, 100-repeats | 100 | Standard | LinearSVC |
| SVM RBF Kernel | C, gamma | 5-fold, 100-repeats | 100 | Standard | SVC |
| Decision Tree | max_depth, min_samples_split | 5-fold, 100-repeats | 100 | MinMax | DecisionTreeClassifier |
| Random Forest | n_estimators, max_features | 5-fold, 100-repeats | 100 | MinMax | RandomForestClassifier |
| XGBoost | n_estimators, colsample_bytree, learning_rate, subsample, max_depth, min_child_weight | 5-fold, 100-repeats | 100 | MinMax | XGBClassifier |

Insert figure legends with the first sentence in bold, for example:

Figure 1. Generalization and classification performance of modeling methods AUC values of



all cross validation and testing performances.

