**Evaluation of binary classification pipelines and methods for 16S rRNA gene data**

Running title: Machine learning methods in microbiome studies

Begüm D. Topçuoğlu, Jenna Wiens, Patrick D. Schloss1

To whom correspondence should be addressed: [pschloss@umich.edu](mailto: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

## Abstract

As the microbiome field continues to grow, there is an ever-increasing demand for reproducible machine learning methods for identifying associations between members of the microbiome and phenotypes. Currently, the field’s use of machine learning lacks clarity and consistency. There is a need for guidance to implement good machine learning practices to generate reproducible and robust models. One application of machine learning to microbiome data has been to classify patients as having colorectal tumors based on microbiota-associated biomarkers. Colorectal cancer is one of the leading cause of death among cancers in the United States. Colonoscopy as a screening tool is effective, however it is invasive and have a low rate of patient adherence. Previous studies have shown that bacterial abundances in the stool can predict colorectal tumors in the colon and can be used as a non-invasive screening tool. However, the prediction performance of these models vary greatly, with areas under the receiver operating characteristic curve (AUC) of 0.7-0.9 (1–4). We used the fecal hemoglobin concentration and 16S rRNA gene sequences from stool samples to classify 490 patients as having advanced tumors or not. Modeling pipelines were established for L2-regularized Logistic Regression, L1 and L2 Linear Support Vector Machines (SVM), Radial Basis Function SVM, Decision Tree, Random Forest and Extreme Gradient Boosting (XGBoost) binary classification models. Tree-based methods, namely Decision Tree, Random Forest and XGBoost were less susceptible to overfitting and in general had higher sensitivity and specificity for advanced tumors. Aside from evaluating generalization and classification performance of each model, this study established standards for modeling pipelines of microbiome-associated machine learning models.

## Introduction

Colorectal cancer is the second most commonly occurring cancer in men and third in women, with 1.8 million new cases in 2018 (8). Early detection of screen-relevant neoplasias (SRN) which are cancerous or advanced precancerous lesions, is crucial for treatment of colorectal cancer. Patients whose lesions are detected at an early stage have a greater than 90 % chance of survival, yet a significant proportion of the US population remain unscreened (9). The high cost and invasiveness of screening methods such as colonoscopy and sigmoidoscopy lower patient adherence. There is a need for non-invasive screening tests that have high sensitivity and specificity for SRN.

Patients with colorectal cancer have different stool community of microbes than adults with normal colons (10–15). This difference however cannot be explained by a single species in the gut but by many of them in relation to one another. Therefore, machine learning emerged as a tool to identify the microbial biomarkers of CRC. Previous studies that used cohorts of 90 (30 SRN, 30 adenomas, 30 normal colons) (7) and 490 (229 SRN, 89 adenoma, 172 normal colons) patients (5) suggest that machine learning models can predict SRN. However, the prediction performance of previous models varies greatly. There is a need for implementing consistent and transparent machine learning practices to generate reproducible and replicable CRC biomarker models. This study aims to determine whether a reproducible and replicable classification model that incorporate the relative abundance of fecal bacterial populations can be developed to discriminate between individuals with and without, SRN.

Here, colonic disease status is defined as Normal or Screen Relevant Neoplasias (SRN). Stool bacterial population abundances and stool hemoglobin levels of 261 Normal and 229 SRN samples were used to learn binary classifiers and evaluate their performances. Modeling pipelines were established for L2-regularized Logistic Regression, L1 and L2 Linear Support Vector Machines (SVM), Radial Basis Function SVM, Decision Tree, Random Forest and XGBoost binary classification models. The mean AUCs of these models were 0.68 ± 0.04, 0.76 ± 0.05, 0.68 ± 0.05, 0.69 ± 0.05, 0.71 ± 0.04, 0.82 ± 0.04, and 0.76 ± 0.04, respectively.

## Results and Discussion

##### Results of modeling in text, tables and figures

##### Comparisons among modeling approaches

#### Interpretation of modeling results in terms of reproducibility, robustness, actionability, interpretibility and susceptibility

#### Consideration of possible weaknesses for each model

The interactions between the biomarkers may be nonlinear. Obviously, the linear models will not incorporate this because they are linear. Tools like linear models (e.g. metastats, lefse, wilcoxon, etc) are likely worthless.

#### Consideration of possible weaknesses for our approach and chosen dataset

#### Relationship of results to previous literature and broader implications of this work

#### Prospects of future progress

## Conclusions

## Materials and Methods

#### Data collection

The data used for this analysis are stool bacterial abundances, stool hemoglobin levels 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 (1). 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 disease status was defined 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 Fecal Immunological Tests (FIT) which measure human hemoglobin concentrations and 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 (5–7).

#### Data definitions and pre-processing

The colonic disease status is re-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. Colonic disease status is the label predicted with each classifier. The bacterial abundances and FIT results are the features used to predict colonic disease status. Bacterial abundances are discrete data in the form of Operational Taxonomic Unit (OTU) counts. There are 6920 OTUs for each sample. FIT levels are continuous data present for each sample. Because the data are in different scales, Python programming language v3.6.6, module scikit-learn v0.19.2 is used to transform features by scaling each feature to a [0-1] range (Table 1) (8).

#### Learning the Classifier

To train and validate our model, labeled data is randomly split 80/20 into a training set and testing set. Then, seven binary class classifiers, L2 logistic regression, L1 and L2 linear suppor vector machines (SVM), radial basis function SVM, decision tree, random forest and XGBoost, are learned. The training set is used for training purposes and validation of hyperparameter selection, and the test set is used for evaluation purposes. Hyperparameters are selected using 5-fold cross-validation with 100-repeats on the training set. Since the colonic disease status are not uniformly represented in the data, 5-fold splits are stratified to maintain the overall label distribution on the training set. Python programming language v3.6.6, module scikit-learn v0.19.2 functions are used to learn the seven classifiers (Table 1).

#### Classifier Performance

The classification performance of learned classifier is evaluated on the labeled held-out testing set. The optimal classifier with optimal hyperparameters selected in the cross-validation step is used to produce a prediction for the testing set. The performance of this prediction is measured in terms of the sensitivity and specificity, in addition to Area Under the Curve (AUC) metrics. This process of splitting the data, learning a classifier with cross-validation, and testing the classifier is repeated on 100 different splits. In the end cross-validation AUC and testing AUC averaged over the 100 different training/test splits are reported. Hyperparameter budget and performance for each split is also reported.

**Figure 1. Generalization and classification performance of modeling methods**  AUC values of all cross validation and testing performances. 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.

## References

1. **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](https://doi.org/10.1128/mBio.00630-18).

2. **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](https://doi.org/10.1186/s13073-016-0290-3).

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](https://doi.org/10.1186/s40168-016-0205-y).

4. **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](https://doi.org/10.1158/1940-6207.CAPR-14-0129).

5. **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.

6. **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](https://doi.org/10.1128/mSphereDirect.00073-17).

7. **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](https://doi.org/10.7717/peerj.2584).

8. **Pedregosa F**, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V**, **Vanderplas J**, **Passos A**, **Cournapeau D**, **Brucher M**, **Perrot M**, **Duchesnay E**. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**:2825–2830.