Machine Learning Manuscript Outline

## Introduction

##### General Context of the work

As the microbiome field continues to grow, there is an ever-increasing demand for reproducible methods for identifying associations between members of the microbiome and phenotypes. Currently, the field’s use of machine learning lacks clarity and consistency.

##### Narrower research area and statement of its importance

* 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. Early diagnosis increases the chance of survival. However the current diagnostic methods are expensive and invasive. As a less invasive tool, studies use relative abundances of the gut bacteria populations to predict disease progression.
* Most microbial communities are pretty patchy and the likelihood of a single feature that explains the differences in health is pretty small. It is likely that many biomarkers are needed to account for the patchiness as well as the context dependency of the features.

##### Identification of a gap or other need for research

* ML use in microbiome literature is a bit like the wild west with lack of clarity over methods, testing, validation, etc. There is a need for guidance on how to properly implement these different methods.
* The prediction performance of previous CRC microbiome models vary greatly, with areas under the receiver operating characteristic curve (AUC) of 0.7-0.9.

##### Summary of appraoch to answer the research question

* We establish a non-leaky pipeline.
* We perform L2-regularized logistic regression, Linear L1 and L2 SVM, Non-Linear SVM, Decision tree, Random forest, XGBoost.
* We evaluate the classification success of different machine learning methods and discuss the reproducibility, robustness, actionability, interpretibility and susceptibility to overfitting of each method.

##### Announcement of principal findings

* Prediction perfomance of each model.
* Is there a maximum threshold of prediction with all these methods?
* Does an increase in model complexity improve predictibility?
* Synthesis statement regarding modeling 16S microbiome data

## Methods

##### Brief explanation of study design/patient sampling and 16S rRNA gene sequencing/curation

##### Analysis of data

* What are the features and what are the labels? We used the fecal hemoglobin concentration and 16S rRNA gene sequences from stool samples to classify 490 patients as having advanced tumors (advanced adenoma or carinoma) or not (non-advanced adenoma or normal colon).
* What is the data (temporal or not)? Assumptions we make when we use a dataset. What will the future data look like?
* Pre-proccessing of the data
* Machine Learning pipeline backbone. How do I split, train, validate and test? A diagram to explain the modeling pipeline.
* Which methods are linear/non-linear? Talk about interpretibility.
* Cross-validation and hyper-parameter tuning methods for each modeling method
* Programming languages and packages/modules utilized
* Statistical methods for comparison of model performance

## Results

##### AUC results of 7 modeling approaches. (FIT + OTUs)

##### Comparisons among the 7 modeling approaches. (FIT + OTUs)

* Compare prediction performance, generalization performance and susceptibility to overfitting.

##### Hyper-parameter tuning budgets and corresponding AUC values.

* This will show that we have used the right budgets to allow the model to pick the right hyper-parameter.

##### AUC results of models with just (FIT)

* This could be in supplemental. We want to make sure that using just FIT as a feature does worse than FIT+OTUs.

## Discussion

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

* What are the metrics to talk about these concepts?

##### 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