Machine Learning Applied to Human Metagenomics

Serena Lakhiani, Arya Shukla
Northeastern University, Boston, MA
DS 5220: Supervised Machine Learning
Lakhiani.s@northeastern.edu, Shukla.ar@northeastern.edu

Abstract

Human metagenomics is a booming field of research focused on microbes and their interactions with their human hosts. Shotgun metagenomic sequencing provides data on microbial abundance and identification of specific markers of microbe strains. To build a classifier for gut microbe data on disease classification, the team used a compiled dataset of 14 studies that combined gut microbe metagenomic data with disease status in 1989 participants. With this, three models were implemented to build a binary classifier of diseased or healthy individuals: Decision Tree, Random Forest, and XGBoost. To evaluate these models, the team output the confusion matrix for each model and used recall as the primary metric. The decision tree classifier performed well as a baseline, with a recall of 0.89. The random forest classifier and xgboost classifier performed nearly identically with a recall of 0.92 for both. The team also began implementation of a multiclass classifier to identify predictive power of gut microbes for individual disease classes. With more time and resources, the team may expand on this classifier and/or implement a stacked model that includes demographic data. It can be extended further via additional hypertuning or feature-weighted linear stacking once a stacked model is implemented.

Introduction

Human metagenomics is a booming field of research that studies microbes and their unique DNA sequences. Microbes are organisms that inhabit the human body–taking shelter on the skin, nose, and gut. From these microbes, scientists are able to learn about various bodily phenomena, such as their connection to the immune system, an individual's emotions and mood, as well as certain diseases and sicknesses. The human gut is one of the most densely populated environments, in that trillions of microbes coinhabit the space; scientists have showed that studying the gene content of the human gut and the enzymes encoded by these genes can provide insight into the chemical capabilities of the microbial environment and a deeper understanding of the symbiotic relationship between these microbes and their human hosts (Pasolli, Truong, Malik 2016)..

Shotgun metagenomic sequencing provides a comprehensive understanding of genes in all organisms in a complex sample, and enables researchers to evaluate bacterial diversity and abundance of microbes. It allows for sequencing of thousands of organisms in parallel, making way for meaningful and interesting comparisons of combinations of microbes in relation to health and disease patterns (Illumina). Depicted below in Figure 1 is a summary of the workflow of shotgun metagenomic sequencing, beginning from sampling to analysis (Quince, Walker, Simpson 2017).

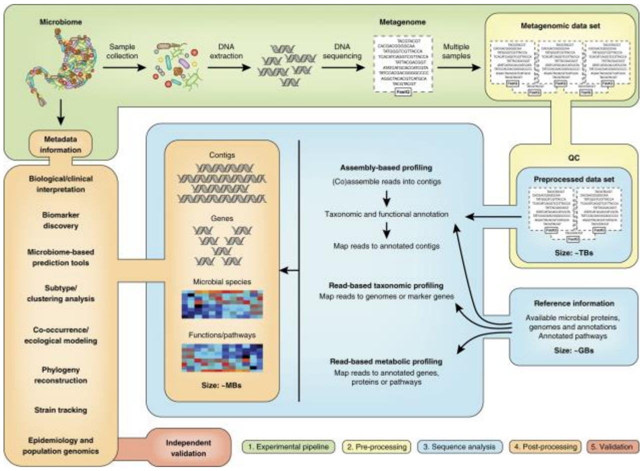


Figure 1: Workflow of shotgun metagenomic sequencing, from sampling to analysis

Using shotgun metagenomics sequencing and known participant disease status, scientists can build classification models to predict disease status in patients. There is currently little understanding of a "healthy" or "ideal" gut microbiome. Building deeper knowledge in this space can open a world of research and development and can aid in disease prediction, diagnosis, prevention, and treatment.

Background & Related Work

There have been major strides in studying the human microbiome to characterize the environments of healthy and diseased individuals. Advances in both sequencing technology and machine learning tools allow for concise representations of large datasets that can generalize a larger set of the human microbiome. Shotgun metagenomics sequencing can be used to detect the abundance of various microbes, as well as specific markers of a microbe strain (Illumina). This investigation will study the predictive power of gut microbe abundance level for human disease.

There are various machine learning tools that can be used in classifying complex datasets like this one. Some that are expected to perform well are Decision Trees, Random Forests, and XGBoost.

Decision Trees are capable of fitting complex datasets and provide easily interpretable results. Tree pruning is usually required to prevent over-generalization via overfitting.

Bagging, or bootstrap aggregating, is a method in statistics to introduce randomness when building predictive models. Bagging can reduce variance of base classifiers, thus improving generalization error. A specific method of bagging designed for decision trees is also known as a Random Forest. In a random forest, tree diversity is increased by sampling with replacement and then splitting each node with a random subset of features. Compared to decision trees, random forests are more robust to noise, but less interpretable and take longer to make predictions. Random forests can be optimized by tuning the number of trees and number of candidate features in each tree.

Boosting is a method in which base classifiers are trained sequentially, so that samples that are misclassified can hold a greater weight for future classifiers. XGBoost is an optimized library for gradient boosting that is more efficient and portable than gradient boosting. In general, gradient boosting takes longer to train than random forest but is often considered one of the best classifier for structured data.

Since the goal of these models is to predict whether an individual is diseased or not, recall would be an appropriate metric to evaluate them. Recall penalizes false negatives and is found by creating the confusion matrix and performing calculation using the true positives and false negatives. A confusion matrix, also referred to as the precision-recall framework, is depicted below in Table 1.

Table 1: Confusion Matrix

	Predicted	
Actual	True Negative	False Positive
	False Negative	True Positive

From the confusion matrix, recall would be calculated using Equation 1.

$$\text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad \text{Eq. 1}$$

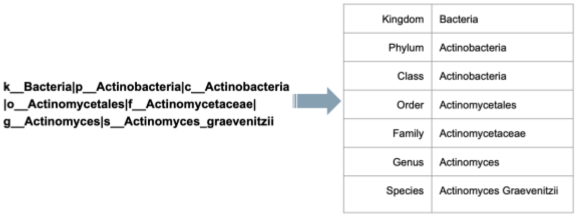
In researching classifiers used for disease prediction with metagenomic sequencing data, the team found some papers on the topic. One such example is Multimodal Variational Information Bottleneck (MVIB), which uses deep learning to build a model capable of learning joint representation of gut microbial species-relative abundances and strain-level markers. (Grazioli, Siarheyeu, Alqassem 2022). These techniques are advanced and require a larger dataset. The current investigation will focus on developing a strong binary classifier based on gut microbe species-relative abundances for healthy vs. diseased populations, and will begin to explore multi-classification of various diseased populations.

Data Collection & Description

The dataset included 1989 participants across 14 different studies. The participants that provided a stool sample across all studies were consolidated. There were 2128 total microbes measured. This dataset presents the abundance of each microbe; shotgun metagenomics sequencing can also be leveraged to derive marker profiles. The dataset also consisted of demographics data on each participant, which was excluded from the metagenomics models built at this stage (Kaggle).

The metagenomic data is described based on taxonomy. An example of a strain broken down by taxonomy is depicted below in Table 2.

Table 2: Example of taxonomic breakdown of metagenomic data



k__Bacteria|p__Actinobacteria|c__Actinobacteria
|o__Actinomycetales|f__Actinomycetaceae|
g__Actinomyces|s__Actinomyces_graevenitzii

Kingdom	Bacteria
Phylum	Actinobacteria
Class	Actinobacteria
Order	Actinomycetales
Family	Actinomycetaceae
Genus	Actinomyces
Species	Actinomyces Graevenitzii

The dataset was provided in .csv format and was imported into a pandas dataframe for easy manipulation. During the consolidation of studies, the number of disease classes was reduced from 18 unique values to 9: control/healthy, IBD, stec2-positive, cirrhosis, type-2 diabetes, impaired glucose tolerance, cancer, and adenoma. The participant count for each class is depicted below in Figure 2.

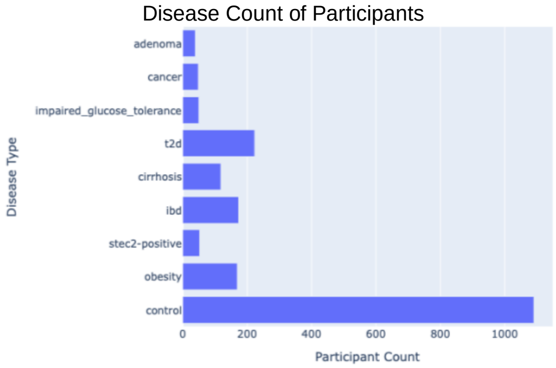


Figure 2: Disease count of participants across 14 studies

The disease classes are clearly imbalanced; the team combined all diseased participants into a single ‘diseased’ class. The count of diseased and healthy participants is visualized below in Figure 3.

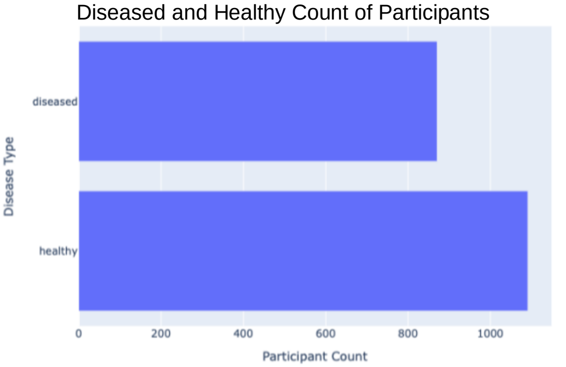


Figure 3: Diseased and Healthy Count of Participants across 14 studies

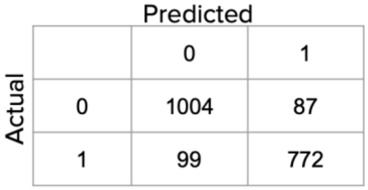
After combining disease groups into a single ‘diseased’ class, the classes are more balanced.

Model Implementation & Results

Binary Classifier

To implement the binary classifier, the diseased class was assigned the positive class. First, a Decision Tree classifier was generated using sklearn. The tree was optimized using pruning with max_depth to prevent overfitting to the training set. The confusion matrix was generated with sklearn and is depicted below in Table 3.

Table 3: Confusion matrix from Decision Tree Classifier

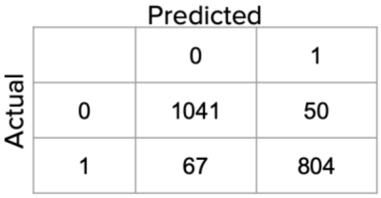


		Predicted	
		0	1
Actual	0	1004	87
	1	99	772

The recall calculated from this matrix is 0.89, as there is a relatively low number of false negatives. The decision tree classifier is a good baseline predictor.

Random Forest classification from sklearn was used next to introduce more randomness and robustness to the classifier. Without hypertuning, this model already outperformed the Decision Tree classifier; the model was optimized by reducing the number of classifiers. Again, the confusion matrix was generated with sklearn and is depicted below in Table 4.

Table 4: Confusion matrix from Random Forest Classifier

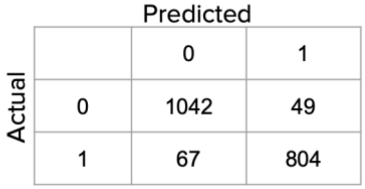


		Predicted	
		0	1
Actual	0	1041	50
	1	67	804

The recall calculated from this matrix is 0.92, as there is an even lower number of false negatives. The random forest classifier is a better predictor than decision tree.

Finally, XGBoost was used as it is often a high-performing model for structured data. XGBoost needed minimal optimization to perform well. The confusion matrix was generated and is depicted below in Table 5.

Table 5: Confusion matrix from XGBoost



		Predicted	
		0	1
Actual	0	1042	49
	1	67	804

The recall calculated from this matrix is 0.92. The XGBoost classifier performed nearly identically to the optimized random forest classifier for this dataset.

Both Random Forest and XGBoost classifiers can perform feature importance based on which features have the greatest impact during classification of samples. A sample of the top features (top 10) is pictured below in Figure 4 a-b.

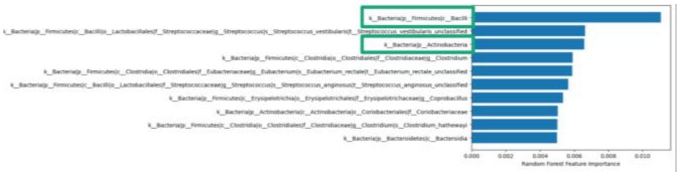


Figure 4(a): Top 10 features extracted from Random Forest Classifier

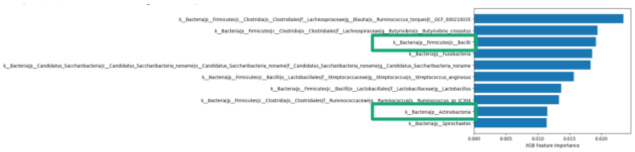


Figure 4(b): Top 10 features extracted from XGBoost

It is notable that specific taxonomy and strains are important features extracted from both classifiers, such as ‘k_Bacteria|p_Firmicutes|c_Bacilli’. These strains are highlighted in green.

Feature importance can be used to perform feature selection and re-model the data. The original dataset provides 2128 features of metagenomic data, which were all used in these classifiers.

Multiclass Classifier

The multiclass classifier was investigated using the same three classification techniques. During initial investigation of this model, the dataset was not balanced, and so these initial results are not necessarily a good representation of classification of this dataset. The data can be balanced using the imblearn library to oversample the minority classes with Synthetic Minority Oversampling Technique (SMOTE) before using the chosen classifiers.

The best of the three classifiers was XGBoost. The 9x9 confusion matrix was generated using sklearn, and the precision, recall, and F1 scores were calculated on an external calculator. The metrics for each class are depicted below in table 6.

Table 6: Mutliclass XGBoost Metrics

Class	Precision	Recall	F1 Score
Control	0.98	0.93	0.95
Obesity	0.83	0.92	0.87
Stec-2-positive	1.0	1.0	1.0
IBD	0.95	0.99	0.97
Cirrhosis	0.97	0.99	0.98
Type-2 Diabetes	0.85	0.95	0.90
Impaired Glucose Tolerance	0.92	0.92	0.92
Cancer	0.85	0.95	0.90
Adenoma	0.79	1.0	0.89

Although these metrics indicate high performance, this model needs to be closely evaluated due to the strongly imbalanced dataset. There is very limited data for some classes, making the training and test sets very small. In addition to SMOTE, some disease classes may need to be excluded from this classifier until more study data is generated.

Conclusions & Future Implementations

In this project, the team successfully created a binary classifier using XGBoost and Random Forest for predicting disease in participants based on shotgun metagenomic sequencing data. Due to the long training time of XGBoost, Random Forest is the recommended classifier for this dataset. The training and test accuracy were 100% and 70%, respectively, and the recall was 0.92, indicating high model performance. Overall, the precision-recall framework of all the models indicate good performance in predicting the positive and negative class.

Although the binary classifier performs well, there is a lot of room for improvement. One improvement could be to perform feature selection, as mentioned. The dataset provides 2128 features, which can lead to a lot of noise due to high dimensionality. Dimensionality reduction would be an important improvement for true implementation of this model. An additional improvement is to build a classifier with the demographics data, and then stack this model on the random forest classifier. These models could be individually hypertuned and then optimized using feature-weighted linear stacking. With more time and resources, the team could expand on this project with the multiclass classifier.

A useful application of this model is to focus on the key microbe strains of a healthy gut microbiome to identify an ‘ideal’ gut microbiome, as this is not well-understood. Realistically, prediction of diseased vs. healthy populations may not provide meaningful benefit to the health and genomics space, as there are better disease detection methods such as imaging. That said, the multiclass classifier could be investigated further to increase understanding around how specific diseases affect the gut microbiome and vice versa.

It is interesting that the binary classifier performed well with this dataset but more data and further investigation is needed to understand how meaningful specific strains are in disease prediction. Overall, the team built a useful disease prediction classifier, that could be expanded in many ways to better fit the needs of healthcare professionals with more time and resources.

References

Pasolli, Edoardo, Truong, Duy Tin, Malik, Faizan, Waldron, Levi, Segata, Nicola. 2016. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. https://doi.org/10.1371/journal.pcbi.1004977. Accessed: 2023-03-20.

Illumina. 2023. Shotgun Metagenomic Sequencing. https://www.illumina.com/areas-of-interest/microbiology/microbial-sequencing-methods/shotgun-metagenomic-sequencing.html. Accessed: 2023-03-11.

Quince, Christopher, Walker, Alan W, Simpson, Jared T. 2017. Shotgun Metagenomic Sequencing, from sampling to analysis. https://doi.org/10.1038/nbt.3935. Accessed: 2023-04-23.

Grazioli, Filippo, Siarheyeu, Raman, Alqassem, Israa. 2022. Microbiome-based disease prediction with multimodal variational information bottlenecks. 10.1371/journal.pcbi.1010050. Accessed: 2023-04-18.

Kaggle. 2021. Human Metagenomics. https://www.kaggle.com/datasets/antaresnyc/human-metagenomics. Accessed: 2023-03-09.