Examining Disease through Microbiome Data Analysis

By

Brett Van Tassel B.Sc Bioinformatics

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science

in Data Science

December 2023

Committee Members:
Debbie Yuster, Advisor
Ashley Stuart, Reader
Keehoon Lee, Reader

Table of Contents

Acknowledgements

Thank you to my Advisor Debbie Yuster, my readers Keehoon Lee and Ashley Stuart, my coworkers, and mentors for their help throughout this project. I appreciate the time and expertise of everyone who has graciously helped me on this project. Thank you to my family for being there for me always. I have found continuous kindness from these people throughout the duration of this project and I am grateful each and every one of them in my life.

List Of Figures

Abstract

The objective of this project is to examine the relationship between gut microbiomes of human subjects having different disease statuses by examining microbial diversity shifts. Read analysis and data cleaning is recorded from beginning to end so that the unfiltered and unfettered data can be reanalyzed and processed. Here we strive to create a tool that works for well curated data. Data is gathered from the database QIITA and the read data and metadata are queried via the tool redbiom. The initial exploratory analysis involved an examination of metadata attributes. A heat map of correlating attributes of the metadata using Cramer's V algorithm allows visual correlation examination. Next, we train random forests based on metadata of interest. Due to the large quantity of attributes, many random forests are trained, and their respective significance values and Receiver Operating Characteristic curves (ROC) are generated. ROC curves are used to isolate optimal correlations. This process is built into a pipeline, ultimately allowing the efficient, automated analysis and assignment of disease susceptibility. Alpha and beta diversity metrics are generated and plotted for visual interpretation using QIIME2, a microbial analysis software platform. CLOUD, a tool for finding microbiome outliers, is used to identify markers of dysbiosis and contamination, and to measure rates of successful identification. CLOUD was found to identify positive diagnoses where Random Forests did not when examining positive samples and their predicted diagnosis status. SMOTE was found to perform similarly or slightly poorer compared to random sampling as a data balancing technique.

Introduction

The goal of this project is to identify disease status through microbial composition. Through the examination of several different metrics, disease status is examined and compared. Using microbial feature counts, machine learning classifiers are trained. Beta diversity is examined with CLOUD to identify high dimensional groupings of samples and for outlier identification. The future goals of this project are to create a tool that provides new patients with a predictive test suite that relays information on high profile diseases and the confidence at which the predictions are made. The random forest classification technique is used to generate high accuracy classifiers and CLOUD is implemented as an interpretable method for disease identification. The project is created with reproducibility in mind, stored in scripts and using QIIME2, a microbiome software suite, plugins to allow interpretability of results. This project seeks to lay the framework needed to standardize the analysis of patients sequenced gut microbiome data, with the future goal of utilizing providing confidence scores in an interpretable way for diagnosis statuses in a human interpretable way.

Background

In this project, the human gut microbiome is examined as it is a noninvasive method of disease identification. In each human gut is a diverse community (microbiome) of microorganisms, including bacteria, viruses, and fungi that inhabit the gastrointestinal tract. The complex ecosystem of the microbiome is involved in various physiological functions. The microbiome is implicated in age related inflammation [1] and in chronic disease [2]. Changes in gut microbiome composition have been observed to change when a person transitions from healthy status to diseased status. The microbial shift can be examined through the sequencing of DNA in the gut microbiome.

This examination is enabled by the increased power of recent sequencing methods. Whereas sequencing the DNA of a single organism was once a herculean task, we can now easily perform highly multiplexed microbiome analyses by gathering sequence data from thousands of bacteria and converting that into high resolution data.

For this analysis, the highly conserved 16S region of the bacterial genome was examined. Compared to exhaustive metagenomic studies which sequence the entire genome of each microbe, focusing on a single region requires fewer computing resources such as sequencing time, sequencing materials, data storage, man hours, and consumption of computational power in data analysis. The function of the 16S region remains stable evolutionarily because it encodes integral structural components of ribosomes. Additionally, transcription of the products encoded by the 16S region stops at RNA, and these are not translated into protein, which sidesteps the involvement of redundant codons and retains the region's original specificity with high fidelity. The 16S region therefore undergoes relatively few mutations over time compared to many parts of the genome. However, there are enough species-specific mutations in the 16S hypervariable regions that bacterial species can be identified without additional analysis of the rest of the genome. 16S sequencing is not without its weaknesses however. The V4 region of the 16S gene is not different enough to distinguish some species level classifications. For example, *Staphylococcus aureus* (pathogenic bacteria) and *Staphylococcus epidermidis* (commensal bacteria) are indistinguishable. The database we use contains sequences targeting the V4 region, as this is the most commonly sequenced target.

Using these 16S sequences, researchers can assign taxonomy by aligning to databases: in the instance of this study the database Greengenes [3], which is specifically focused on the study of bacterial and archaeal 16S rRNA. Different regions of the 16S have different levels of discriminatory power, for this project we use the V4 region which loses discriminatory power at the genus and species level [4]. Additionally, the accuracy of any large-scale microbiome classification is dependent on the quality of the database used. New species that are found or have evolved since database creation are not classifiable to the same taxonomic level. To future-proof this issue, we employ the purpose-built platform QIIME2, which enables simple implementation of updated reference databases into the same pipeline.

After the extraction of DNA and sequencing, short strings of sequences called reads are generated and stored into read files. Many samples are sequenced at once, generating a large quantity of short strings, typically between 75bp and 500bp long. This process multiplexes a large quantity of sample data into a single output stream. Individual reads are demultiplexed from the stream into their source samples using their attached barcodes to separate reads into each respective file.

After generating demultiplexed read files, these reads are “denoised” through bioinformatics tools implemented as QIIME2 plugins, such as Deblur [5] and DADA2 [6]. This is a process that eliminates errant reads and poor-quality reads, improving downstream results. Additionally, reads are trimmed and truncated to remove base pairs of low quality towards the starts and ends of the reads, respectively. These tools truncate the reads enforcing a static read length preventing potential overfitting of differently sequenced data and consistent downstream analysis results. However, ideally reads are sequenced with the same process to prevent biases in the classification model.

Methods

Data Acquisition and Preprocessing

The entirety of this thesis project is stored into pipelines to allow for the simple reanalysis of data. The first step queries data using the tool redbiom [7]. Redbiom allows us to draw data from QIITA [8], an open-source microbial study management platform. We use QIITA for this analysis over other databases like NCBI because the querying methods are simplified and the metadata for data in QIITA is necessarily consistent. Additionally, QIITA stores its data into QIIME2 artifacts which validates data types ensuring the data exactly matches common bioinformatic data type formats, such as FASTA files, which is a text file containing a list of sequences. and tracks artifact processing. When querying from QIITA, we ensure the data are all processed in the same way to eliminate biases from our downstream analysis. The most common strategy, used here, combines a specific sequencing method, denoising method, and read trimming length available in the QIITA database. This involves: sequencing on an Illumina instrument, targeting the V4 region of the 16S gene; denoising with Deblur 2021.09; and trimming the reads at exactly 90 nucleotides. Additionally, for the analysis described here, we selected only data from the American Gut Project (AGP) [9] to ensure a focus on the gut microbiota. We have thus narrowed down our source data to a total of 33,590 samples/patients, resulting in a file containing counts of unique reads generated by sequencing.

Data Exploration and Filtering

We begin exploratory data analysis after querying to examine the spread of attributes for our dataset, with respect to having Irritable Bowel Disorder.

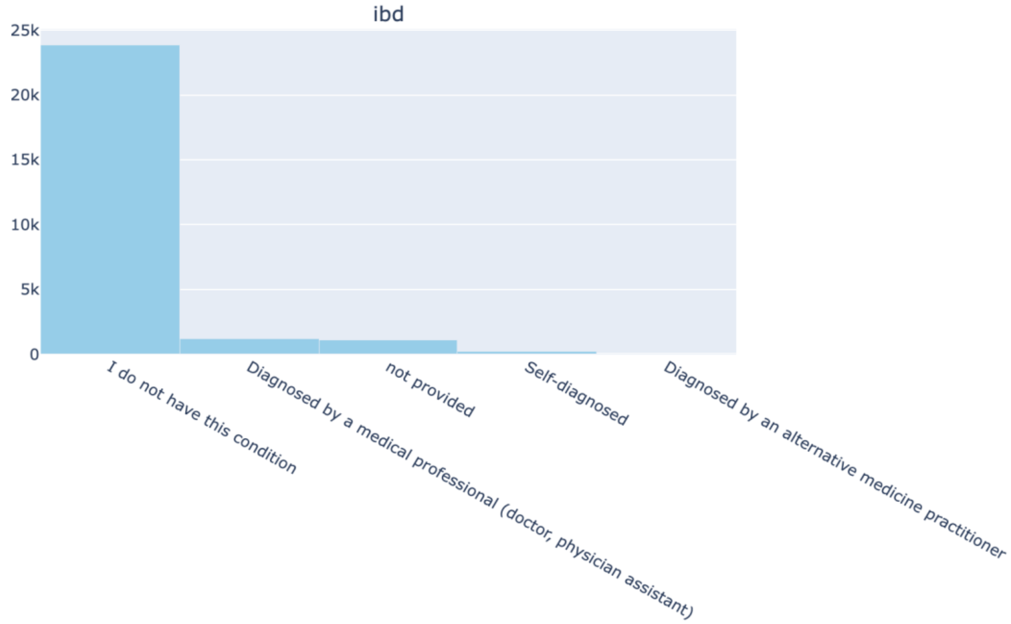


Figure 1: Representative Population of American Gut Project Irritable Bowel Disorder Status. This plot depicts the disease status for patients tested for Irritable Bowel Disorder within our queried dataset. The majority of samples have the status, “I do not have this condition”, with just under 25,000 samples with this status. The “Diagnosed by a medical professional (doctor, physician assistant)” status encompasses fewer than 2,500 samples.

Other diseases follow a similar trend, with their diseased class containing a quantity of samples several magnitudes lower than that of their diagnosed counterparts. This is expected, but Figure 1 exemplifies why we must mitigate this imbalance before we train our random forest classifiers. We shall implement sampling methods to balance our datasets. These methods are discussed in detail in the Sampling subsection on page 12.

To further investigate the spread of our data, we then generated and examined Kernel Density Estimation (KDE) hist plots [10] for three different metadata attributes; age, height, and weight. The data-smoothing function of KDE plots prevents noisy and sparse data from negatively affecting the interquartile range from which we use to select samples for further analysis. Figure 2 examines the distributions for age, height, and weight of our queried data.

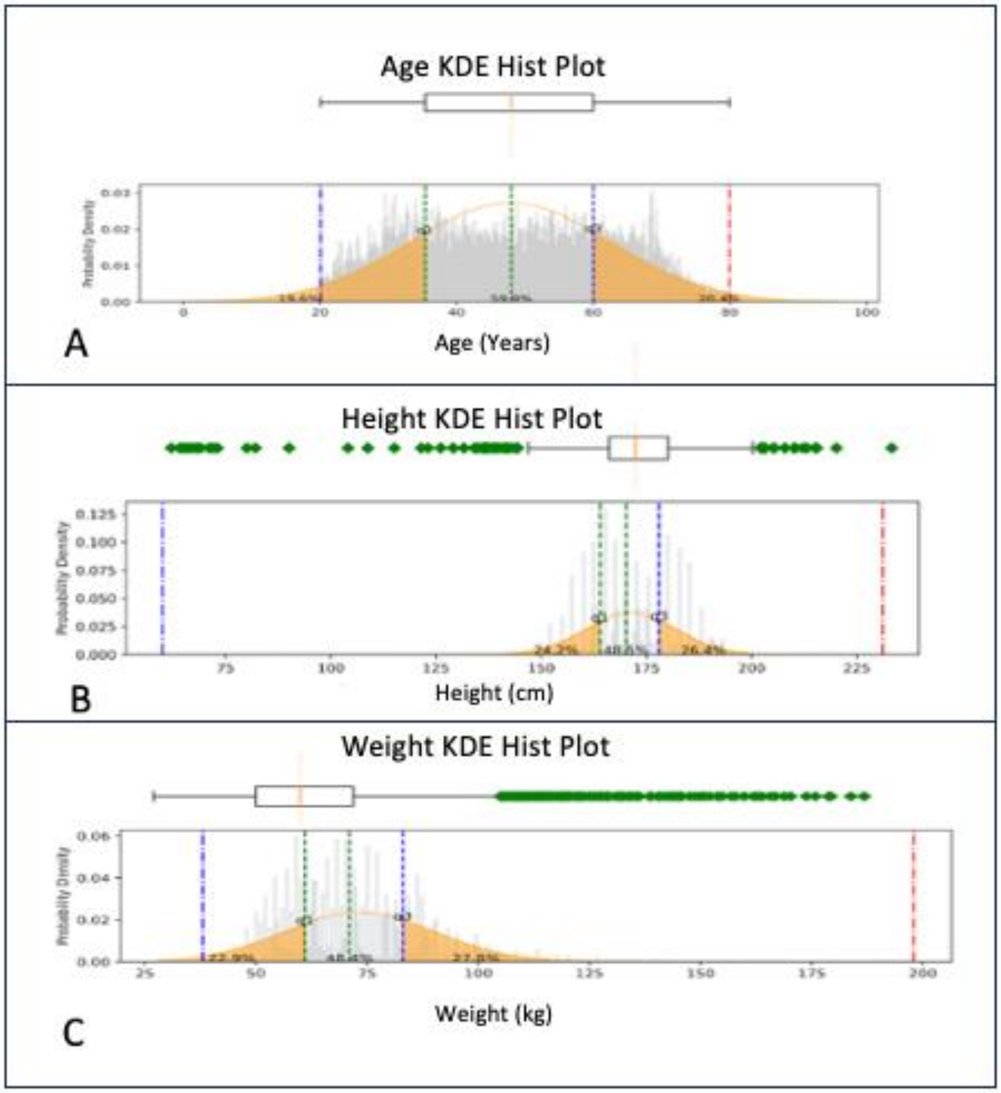


Figure 2: KDE Hist Plots of Attribute Distributions. A, 59.9% of the ages included fall within the interquartile range of the smoothed function between 47.9 years and 60.0 years. **B,** Height distribution KDE Hist Plot. 48.5% of heights included are in the interquartile range of the smoothed function, ranging from 170.18cm to 178.0cm. **C,** Weight distribution KDE Hist Plot. Interquartile range of the smooth function comprises 48.5% of the samples; weights range from 71.0kg to 83.0kg.

The given distributions are generated by placing an individual kernel (smoothing function) at each data point and summing up these kernel functions to create a smooth curve. The interquartile values of the smoothed function are retained, filtering our data of extreme values and examining relatively equivalent samples for age, weight, and height. The interquartile range was selected because it is a robust measure of spread that is less influenced by extreme values than the range or standard deviation [11].

When sequencing samples in the lab, it is possible that some samples may fail due to improper preparation techniques, and that small amounts of contamination may enter the samples. To mitigate the influence of contamination and failed sample sequencing on our data, we additionally filtered our input data to exclude samples with frequencies below 1,000 reads, and filtered features to exclude those with fewer than 1,000 reads. These given filtering options are not robustly examined, and the possibility of filtering this data with different thresholds could improve performance of later classifiers. In the future, we would like to examine the influence of different filtering methods on the taxonomic output and classifier accuracy.

Within the queried metadata, there are some attributes that are directly influential on others. The full set of attributes used include race, irritable bowel syndrome, autism spectrum disorder, migraine, autoimmune, thyroid, sibo, appendix removed, chickenpox, csection, acid reflux, diabetes, liver disease, lung disease, kidney disease, birth year, cardiovascular disease, sex, seasonal allergies, lactose, cancer, epilepsy or seizure disorder, host age, pregnant, contraceptive, host height, BMI cat, and host weight. Some of these attributes are encoded with Boolean values, others are categorical indicating positive status, negative status, and uncollected data. Birth year and age are incorporated into the metadata, but they are different forms of the same information and are thus dropped. The other categories may not have obvious influence on each other, so we examine the categories' correlations with each other. Cramer’s V is a measure of association between two categorical variables, so we calculate the Cramer’s V value for each pair of metadata attributes. The categorical attributes are converted to numerical values for our investigation of the metadata.

$$V = \sqrt{\frac{\chi^2}{n \cdot min(k-1, r-1)}}$$

Equation 1: Cramer’s V Equation. Where χ^2 is the chi-square statistic obtained from the contingency table, n is the total number of observations in the table, k is the number of columns in the table, and r is the number of rows in the table.

Cramer’s V is an extension of Pearson's chi-squared test and is used to quantify the strength of association between two categorical variables in a contingency table. Values range from 0 to 1, where 0 indicates no association and 1 represents a perfect association. Using these values, we generate a heatmap to quickly view strong associations.

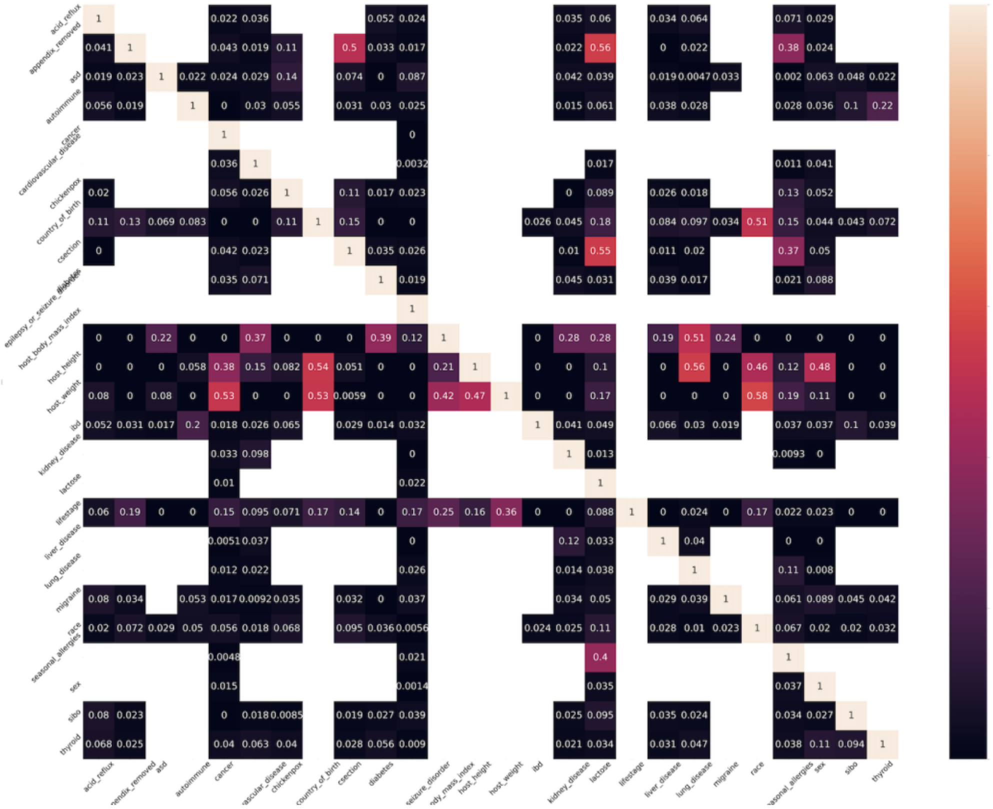


Figure 3: Cramer's V Attribute Heatmap. Our exploratory data analysis for thyroid disease using Cramer's V is plotted in a heatmap, allowing for intuitive understanding of feature correlation. Notably, host weight and cancer are correlated with a Cramer's V of 0.53. Another visible correlation that is intuitively expected is that between race and both weight and height. Blank spots represent null values.

The absence, in such a heatmap, of Cramer's V values near 1 indicates that there are no metadata attributes in this analysis with perfect correlation. The ranges observed in the heatmap are from 0 to 0.56, This means we can safely train random forest classifiers for each of the metadata attributes using all other metadata attributes.

Disease Status Prediction

For reasons noted previously, we sample the data using under sampling of the majority class and over sampling of the minority class to gather equivalent counts of diagnosed and negatively diagnosed classes. Our input data contains categorical variables which cannot be used as input for random forest classifiers, so the categorical variables are one-hot encoded. One-hot encoding takes each label; "Diagnosed by a medical professional (doctor, physician assistant)", "I do not have this condition", and "not provided" our data, and generates a binary column where a value of 1 indicates the patient had this label, and 0 indicates the patient did not. The original categorical column is removed, and the metadata table contains Boolean columns for the three labels to use as random forest input.

With our input data prepared, the random forests are trained using Randomized Search Cross Validation (RSCV) for broad parameter tuning. RSCV is the random sampling of combinations of hyperparameter values from predefined ranges to evaluate their performance using cross-validation to enable a more efficient exploration of the hyperparameter search space.

To examine the microbiome composition and its influence on disease we utilize frequency tables which tables that contain the frequencies of given features, in our case unique reads, as seen in Table 2.

Unique read	10317.000147908.135979	10317.000058724.131314	10317.000101010.128285	10317.000029168.128304
TACGGAAGGTCCGGGCGTTATCCGGATTTATTGGGTTTAAAGGGGGCGCAGGCGGACTCTTAAGTCAGTTGTGAAATACGGCGGCTCAAC	0	0	0	0
TACGTAGGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGGGTGCGTAGGCGGGTTATCAAGTCTTTGGTTAAAATACGGTGCTCAAC	0	0	0	0
TACAGAGGATGCAAGCGTTATCCGGAATGATTGGGCGTAAAGCGTCTGTAGGTGGCTTTTCAAGTCCGCCGTCAAATCCCAGGGCTCAAC	0	0	0	0
TACGTAGGGAGCGAGCGTTATCCGGATTTACTGGGTGTAAAGGGCGTGTAGGCGGGGAGGCAAGTCAGGCGTGAAAACTCAGAGCTCAAC	0	0	0	0

Table 1: Example Feature Table File. This table shows an example of a qiime2 feature table file. The column names are sample IDs while the row labels are the features, which are, in our case, 90 bp long sequences. We have gathered this data by querying redbiom.

The feature table is manually converted to representative sequence files. Representative sequence files are stored into text files with each sequence on a new line delineated by the character “>”, also known as FASTA files, containing all features (reads) from the feature table. With read data prepared in representative sequence files, the reads are aligned to a reference database. The reference database used is Greengenes. To ensure reads align only to the 16S region of the genome, the 16S forward and reverse primers are used to extract the 16S V4 region (shown in the schematic in Figure 4) of the genome.

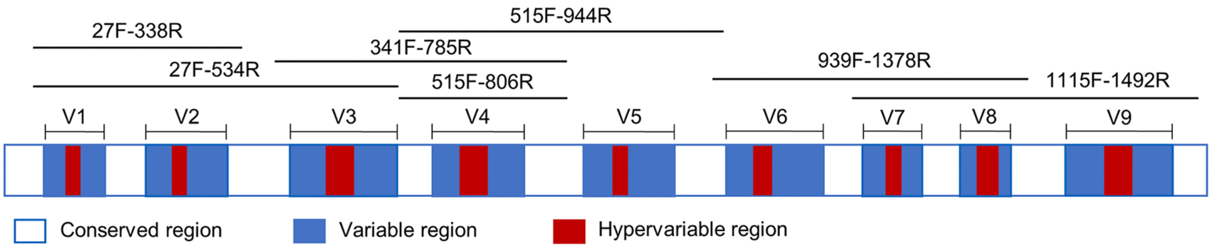
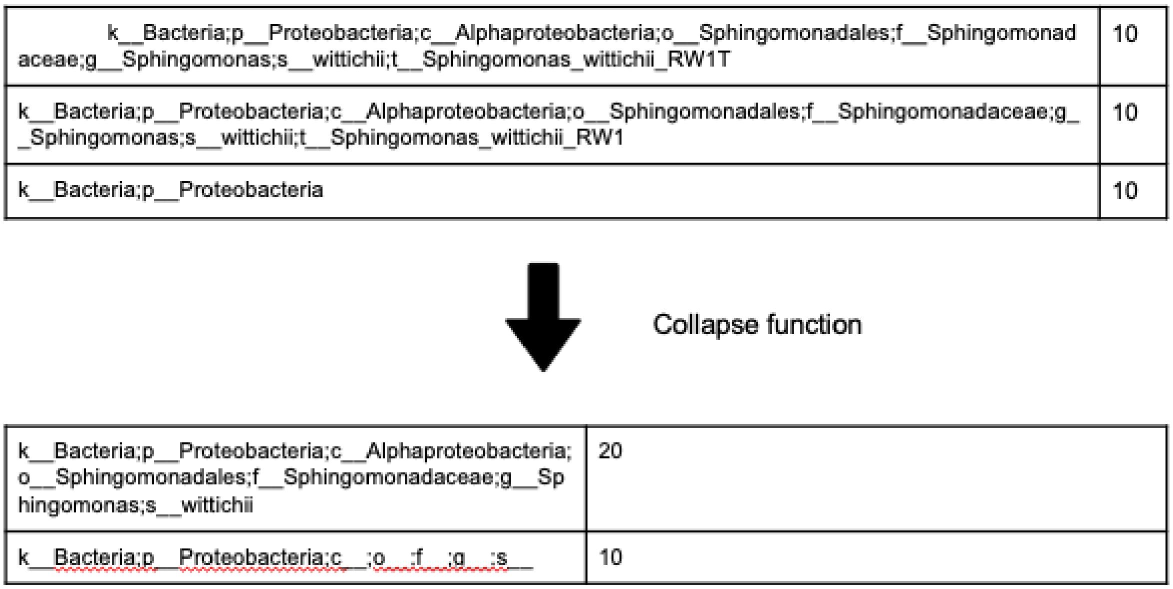


Figure 4: Regions of the 16S genome that are identified as hypervariable. 515F and 806R are the primers used for sequencing the V4 region of the 16S gene. [12]

We will use each of these extracted reads and generate a prediction as to what taxonomic classification the read is most similar to through a $k - mer$ based taxonomic classification method. Taxonomic classification uses hierarchical categories based on shared evolutionary characteristics, generated through sequence alignment. The reads previously extracted from the Greengenes database with forward and reverse primers are used to train a multinomial Naive Bayes machine learning classifier by relating the extracted Greengenes sequence data to the taxonomic labels. Specifically, the extracted Greengenes sequence data are broken into $k - mer$, where a $k - mer$ is a k-length portion of the sequence. Each read sequence is broken down into all possible $k - mer$. The frequencies of each $k - mer$ are recorded for each read sequence. Then, a Naive Bayes classifier is trained using the frequencies of the k-mers and the associated taxonomic labels probabilities to generate a probabilistic model for disease status.

To classify our dataset, the AGP queried reads are classified with the trained multinomial Naive Bayes classifier. This, again, extracts k-mers from the sequence and assigns a taxonomic classification based on the k-mer counts with the pretrained probabilistic naive bayes model. Once a sequence is classified to a specific taxonomic label (species or strain), the information is propagated up the taxonomic hierarchy to higher taxonomic levels. The taxonomic hierarchy from bottom to top in our case contains strain, species, genus, family, order, class, phylum, and kingdom. If a sample contains read sequences that align to multiple species that are in the same order but do not share a family or genus, the species or strain level k-mer classifications will be propagated up to the order level, resulting in a large quantity of k-mers assigned to that level of classification. That will increase the chances that a give sequence will be classified at the order level, which is what we want to capture as there is conflicting information at more specific levels of taxonomic classification.

This method of classification can result in granular results with classifications down to strain level that are highly specific. To mitigate the complexity of the random forests we will train later, the taxonomic classifications are collapsed down to species level, merging strain level classifications to the species level such that strain counts are added to the species counts. Reads that are classified up the tree are assigned blank values as place holders for levels beyond the granularity of the specific assignment.



k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__wittichii;t__Sphingomonas_wittichii_RW1T	10
k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__wittichii;t__Sphingomonas_wittichii_RW1	10
k__Bacteria;p__Proteobacteria	10

Collapse function

k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas;s__wittichii	20
k__Bacteria;p__Proteobacteria;c__;o__;f__;g__;s__	10

Figure 5: Transformation of Taxonomy Feature Table to Collapse to Species Level. The arrow denotes the collapse function. Before the collapse, there are two strains of *Phingomonas wittichii*, and after the collapse, the two strain counts are added together and the classification is the species level classification of *Phingomonas wittichii*. Some features are identified as Proteobacteria with no further granularity in classification level. Those features are collapsed down to species level, but the collapse function does not add in new information or change the counts when collapsing increases granularity.

Sampling

With our taxonomic information gathered, we move on to training the microbially informed random forests. Similar to our metadata informed random forest training, our data contains highly imbalanced diagnosis statuses. To deal with our imbalanced data, we shall train forests for two different sampling methods and one without. The first random forest is trained with the full unsampled data. In the case of the next forest, we under sample the majority class followed by over sampling the minority class. Lastly, in the third forest, we under sample the majority class and then over sample the minority class using SMOTE, an algorithm which generates new synthetic samples for the training dataset by selecting random samples and its k nearest neighbors, and averaging the values for the features of the cumulative samples.

There are nine diseases examined, Irritable Bowel Disorder, Thyroid disease, Liver Disease, Cardiovascular Disease, Kidney Disease, Cancer, Lung Disease, Autism Spectrum Disorder, and Migraine. The microbiome varies greatly by gender, so random forest classifiers were trained twice per disease, one classifier per gender. This results in 18 disease-gender combinations. Each of these combinations is examined with several random forests using different sampling methods and different random forest parameters.

Outlier Detection

Now that the random forests are trained, some more interpretable metrics are examined to compare the results with. To identify abnormality in diseased microbiomes, beta diversity matrixes are examined with CLOUD [12], an outlier detection tool. Microbiome beta diversity is a concept used in microbiome research to describe the variability or dissimilarity in microbial community composition between different samples or environments. High beta diversity indicates substantial dissimilarity between microbial communities, suggesting that different factors influence the composition of the communities in the compared samples. This is an interpretable metric used to view the differences between microbiome compositions.

There are multiple beta diversity metrics; UniFrac distance is a widely used phylogenetic metric, while Bray-Curtis dissimilarity and Jaccard distance are examples of compositional metrics. Weighted UniFrac is used in this analysis.

$$WU = \frac{\sum_{I=1}^{S} \quad w_i d_i}{\sum_{i=1}^{S} \quad w_i}$$

Equation 2: Weighted UniFrac Formula.

To analyze beta diversity, a distance matrix is generated which will be used as input for CLOUD. Using the Weighted UniFrac formula, pairwise comparisons are run to fill a distance matrix. The Weighted UniFrac distance is computed by summing the product of the abundance difference (w_i) and the phylogenetic distance (d_i) for each shared taxonomic classification in the sample microbial communities. This calculation emphasizes both phylogenetic relationships and abundance differences. This sum is then divided by the total abundance, effectively normalizing the result.

The ecological distances are input into Cloud-based LOcally linear Unbiased Dysbiosis (CLOUD), an outlier detection test that functions using a clustering algorithm. The goal of implementing CLOUD outlier detection test is to identify if diseased samples are outliers with respect to beta diversity, which is a more interpretable metric than our classifiers.

CLOUD uses locally linear embedded ecological distances, and functions by creating a reference "cloud" (hence the name) of microbiome samples with a diagnosis status: in our case, "negative diagnosis" with regard to IBD. On a high level, CLOUD is taking target samples and identifying if the cloud of the target sample is highly different from the reference data clouds. If it is significantly different, this indicates a difference in beta diversity compared to the healthy references.

CLOUD takes the target samples and calculates whether or not the target sample's cloud has a higher diameter compared to the average reference cloud. If the target sample's cloud is larger by a set threshold, we identify the sample as a CLOUD outlier.

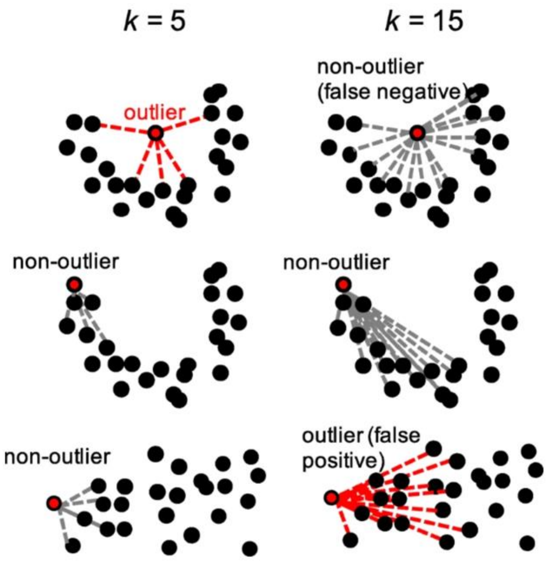


Figure 6: Cloud High Dimensional Visual Interpretation. In this plot, the outlier detection is based on distance to a samples k nearest neighbors. False positives can be generated when the quantity of neighbors used is too high, if a nonoutlier sample lays on the edge of high dimensional space compared to the other samples. This makes parameter tuning an interesting opportunity for future work.

Results

Random Forests Trained with Metadata

The metadata informed random forest model scores for the nine target disease accuracies and Receiver Operator Characteristic Area Under the Curve values (ROC AUC values) are gathered and compared in Table 2 below. Accuracy is defined as correct predictions over total predictions and the ROC AUC is a metric for overall performance of the data. ROC AUCs perform better than the accuracy metric for imbalanced data. However, because the data used is balanced using SMOTE or over sampling and under sampling, the simple and intuitive accuracy metric is used for identification of the best model. The best classifier by accuracy score was the model trained to identify IBD, and its ROC Curve is shown in Figure 7.

Target Feature	Accuracy	Train Roc Auc	Test Roc Auc
Irritable Bowel Disorder	0.85	0.920	0.841
Thyroid disease	0.78	0.867	0.825
Liver Disease	0.754	0.943	0.725
Cardiovascular Disease	0.734	0.938	0.852
Kidney Disease	0.692	0.923	0.756
Cancer	0.671	0.889	0.817
Lung Disease	0.663	0.828	0.746
Autism Spectrum Disorder	0.648	0.974	0.702
Migraine	0.593	0.785	0.736

Table 2: Metadata Informed Random Forest Accuracy Scores.

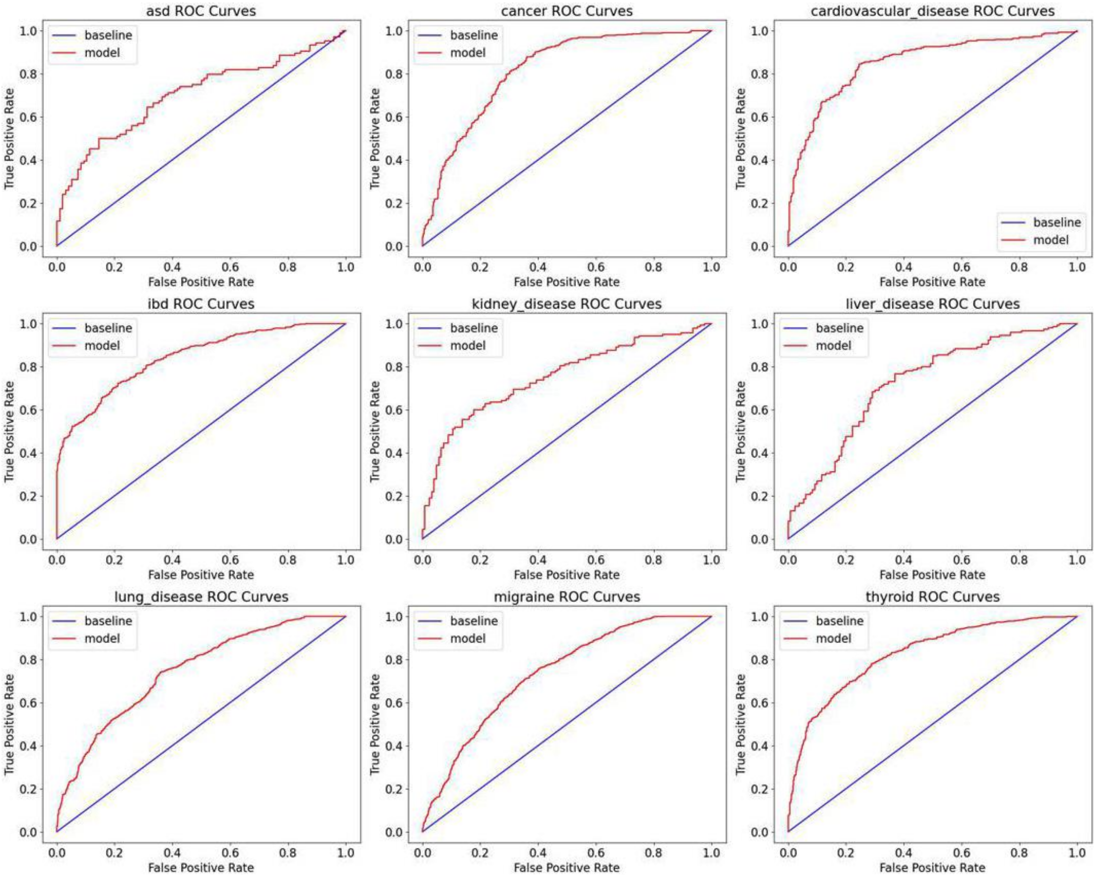
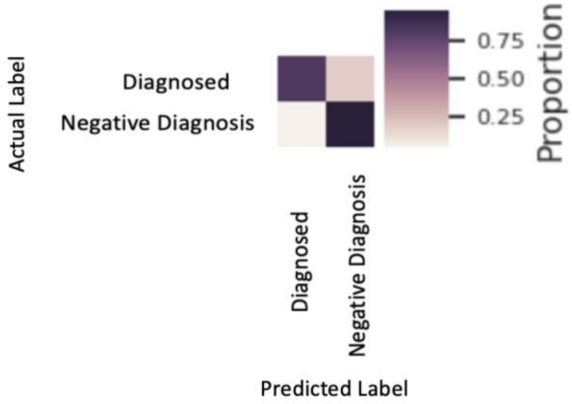


Figure 7: Metadata Informed ROC Plots. The IBD classifier trained with metadata had the highest accuracy (0.85) of all the metadata classifiers. The training ROC AUC was 0.92 and the testing ROC AUC was 0.84. A curve aligning with the top left corner of the ROC plot indicates that the classifier is performing well compared to baseline, the diagonal line.

The best model is quite informative when predicting diagnosis with only metadata: the accuracy of the model is 85%. The training ROC AUC Score is 0.92, compared to the testing ROC AUC of 0.84, indicating the possibility of overfitting of training data.

Microbially Informed Random Forests

For all the disease-gender combinations for the microbially informed random forest classifiers, the random forest classifiers using resampling for males with IBD positive and negative statuses performed the best with an accuracy of 91.8%. The forests were not trained specifically to reduce false negatives, an important thing to consider for clinical samples. A false negative result would indicate to a patient they are not sick, and therefore might not seek treatment.



	Diagnosed	Negative Diagnosis
Diagnosed	0.82101	0.17899
Negative Diagnosis	0.0407	0.9593

Figure 8: Confusion Matrix of Microbially Informed Random Forest Predicting IBD for Males Trained with Randomly Sampled Data. The bottom right corner of the confusion matrix is the percent of times that a Negative Diagnosis was correctly predicted is 95% accurate, compared to the diagnosed prediction accuracy of 82%. In future work, weighing these classifiers toward predicting disease status is important. This would result in higher false positive counts but mitigate false negative classifications. The upper left corner indicates true positives and the lower right corner contains true negatives.

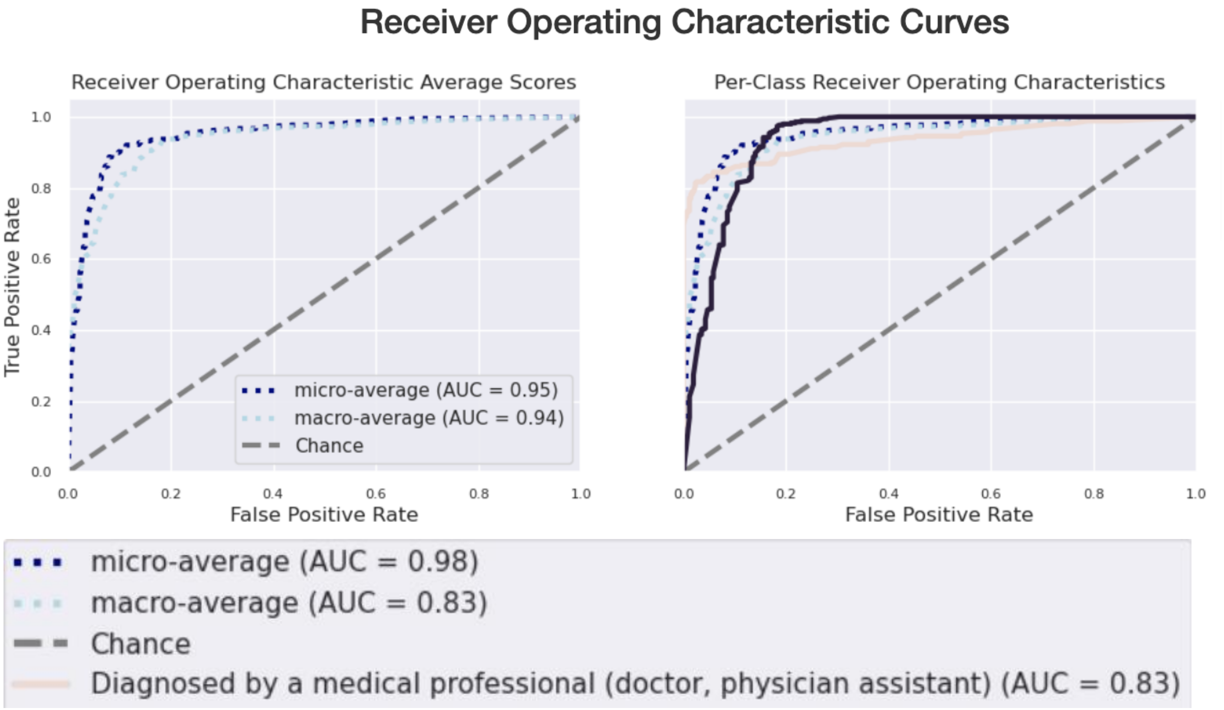


Figure 9: IBD Random Forest ROC Curves. The ROC curve is a graphical representation of the performance of a binary classification model at various discrimination thresholds. It is used visually identify where we would like to set threshold values. Sensitivity reduces the quantity of false negatives which is highly important in the diagnosis of diseases, so using this curve visually to increase sensitivity is a possible future work. The micro-average line aggregates the contributions of all diagnoses by considering the total number of true positives, false positives, and false negatives across all classes. The macro-average calculates the performance metrics separately for each diagnosis and then calculates the average.

The random forest accuracies using two different sampling methods were visualized and compared, seen in Figure 10.

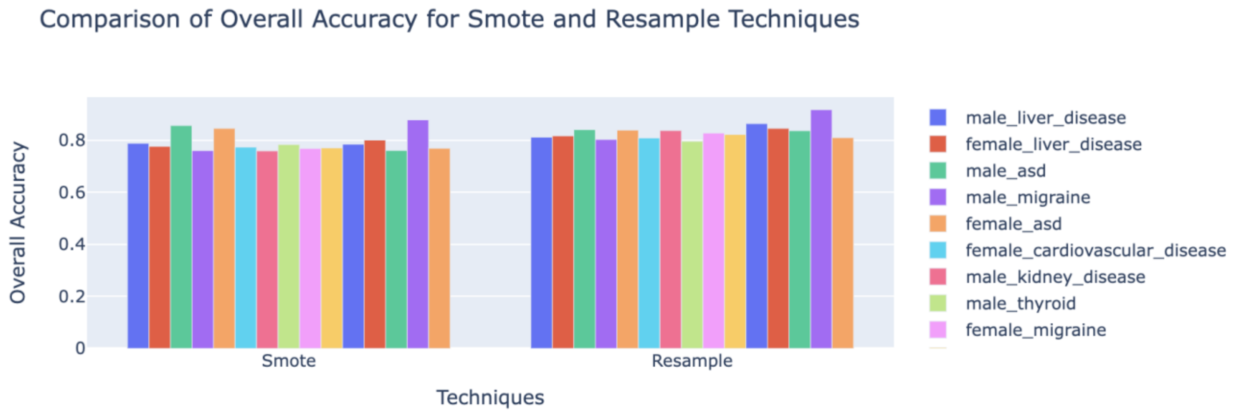


Figure 10: Comparison of Average Sampling Accuracies. The random forest accuracy scores show that simpler resampling (right hand plot) performs better for the classification accuracy of the samples. The use of SMOTE to identify disease in a random forest may not be working as well in this examination. It may through the averaging of feature information be losing information, as examined in Figure 13.

Across the random forest models, we extract the 10 most commonly identified microbes. From this list of the 10 most commonly identified microbes, we examine the microbes with the highest feature importance score. We examined found that *Turicibacter sanguinis* [13] (suspected in having influence in IBD), *Peptococcus niger* (suspected in human infections) [14], *Dialister invisus* (a pathogen), *Odoribacter splanchnicus* (decrease in abundance has been associated with inflammatory bowel disease), and *Fenollaria massiliensis* [15](a common vaginal microbe) were the most important pathogens in the classification of diseases.

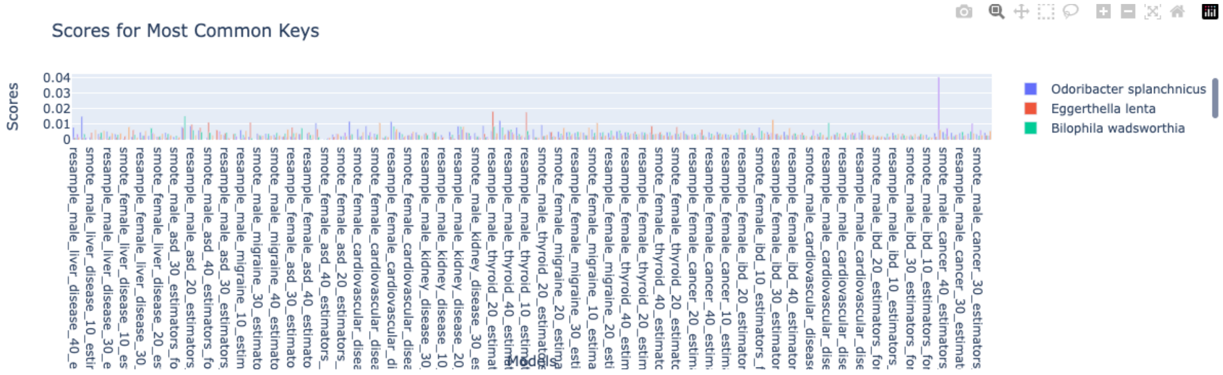
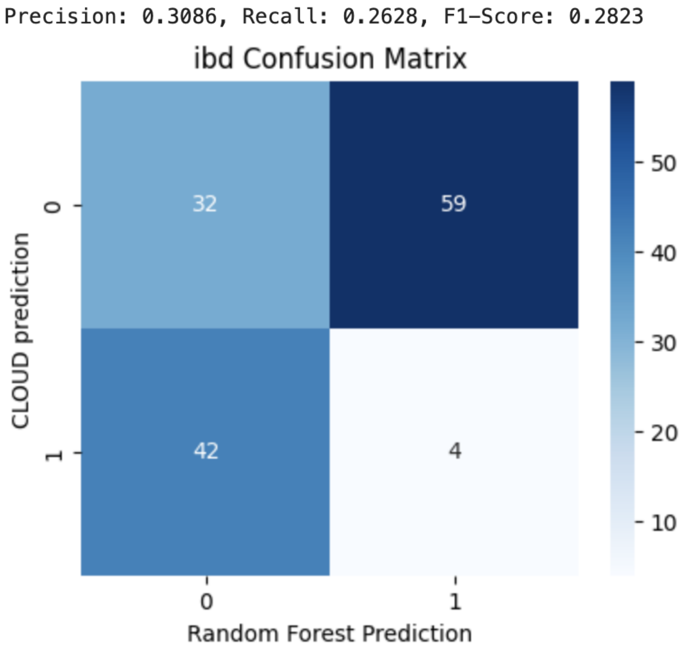


Figure 11: Most Commonly Found Microbes Average Importance Scores. The most important nonpathogenic microbes were *Eggerthella lenta* [16], and *Bilophila wadsworthia* [16]. Both are common and viewed to be related to disease when imbalanced.

CLOUD Outlier Detection Test Comparison

For a given gender:disease dataset, we run CLOUD’s outlier detection test to determine whether CLOUD detects as an outlier what the random forest predicts as “diagnosed” (in other words, diseased). To view this, we generate confusion matrices to compare the random forest classification results and the CLOUD outlier detection test results.

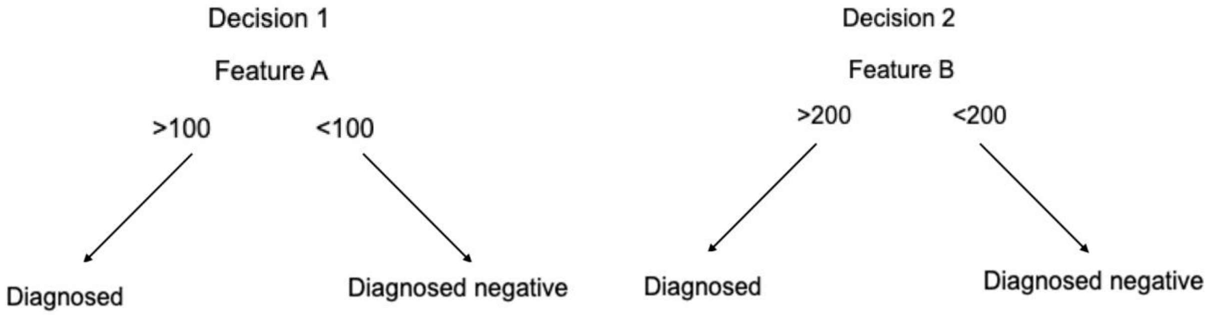


```
Precision: 0.3086, Recall: 0.2628, F1-Score: 0.2823
```

Figure 12: Microbially Informed Irritable Bowel Disorder Confusion Matrix. Each sample in the confusion matrix represents a patient diagnosed with IBD. 0 indicates a prediction of negatively diagnosed, 1 indicates a prediction of diagnosed. It is clear that both tests are weighted toward not diagnosing the disease. However, using the low proportion that give both the correct value for CLOUD and for the random forest, we infer that there may be interference that prevents both from identifying a diseased target. Perhaps due to the trimming of the random forest decision trees to prevent over tuning, these outliers do not have clear decision tree pathways to be classified accurately. This indicates that there is possible value in using CLOUD to identify microbiome outliers. Using outlier tests can indicate to patients that there is irregularity in their gut microbiome and that their random forest classification may be inaccurate. For the diseased patients in the confusion matrix, the random forest was unable to predict disease status for 42 of them. CLOUD identified these samples as outliers, so a negative classification for an outlier can be considered less reliable.

Discussion

The use of SMOTE as a sampling technique did not appear to give interesting results. It would be interesting to see if samples used for training the classifier are misinforming the classifier. It instead may result from averaging of the feature counts in the synthetically created diagnosed samples. Average values are less informative in differentiating, perhaps resulting in the random forests inability to classify as well.



Three Nearest Neighbors				
	Sample 1 (diagnosed)	Sample 2 (diagnosed)	Sample 3 (diagnosed)	Sample 4 (diagnosed generated by SMOTE)
Feature A	159	107	10	92
Feature B	10	20	400	143

Figure 13: Random Forest Decision and Smote Feature Table Example.
The average values generated by SMOTE results in a classification of diagnosed for a sample that should be diagnosed negative. This can be caused by a disease where multiple features and not necessarily the combination of those features are informative of disease. In the example, because SMOTE generated averaged values for the testing data, the classifier loses accuracy.

In Figure 13, Decision 1 is an example of a simple decision tree with only one decision. If Feature A contains read counts less than 100, the sample is diagnosed negative, and the inverse is true for positive. Our example feature table has three samples that are diagnosed positive. The goal of utilizing SMOTE was to improve the decision boundary for highly imbalanced classes. SMOTE is designed to help generate samples for nonlinear and complex data, but if the microbiome is not informative enough, then it is possible using SMOTE might not be advisable.

In this analysis, we compared accuracies of random forest results. An analysis of the accuracy ratios may also be valid, but not necessary in our case. The accuracy ratio of the random forest classifiers is the accuracy divided by the baseline accuracy. The baseline accuracy is the proportion of the majority class, which in the case of this project is roughly equivalent to the minority class, meaning that the accuracy ratio is not as informative a metric for balanced data.

In future work, the likelihood of a given classification's result will be compared across metrics to generate a comprehensive report of diagnosis likelihoods for each of the diseases examined here. In other words, for a given classification, the likelihood of a correct diagnosis will be linked with the result. For example, if a random forest is weighted to diagnose a sample as positive, we want to additionally inform the patient that although our metrics have identified the sample as positive, it over estimates positive diagnoses 10% of the time. If the random forest and cloud outlier detection test agree and diagnose a sample as positive, then we want to identify how much more likely that is as well. The implementation of several metrics in this project—CLOUD, microbially informed forests, and forests implemented with metadata—will each be used in combination to identify the frequency of sample misclassification. The likelihood of missing a diagnosis across these metrics in combination will then eliminate more samples from the pool of false negatives. The use of a patient's microbial data and metadata are then examined by disease to provide an overarching diagnosis.

Future works include an examination into pathogens that are identified where they should otherwise not be found are of interest. For example, examining further whether *Fenollaria massiliensis* is found in the male microbiome is of interest, because it was found to be one of the most important features in random forest training for this analysis. Additional expert analysis for each disease analysis report is necessary for the identification of pathogens of interest for disease. A count of feature identifications in a database for querying would allow users to find the most commonly found pathogens or microbes for a given disease, and allow users to plot the rise of pathogens over time with associated or new diseases.

A further examination into the filtering of samples as described in the Data Acquisition and Preprocessing subsection above is of additional interest for the analysis of disease. A robust measure of what samples are informative to the classifiers trained would be necessary to identify at what thresholds samples can be filtered at. The idea would be to automate this filtering method so that with new datasets and new diseases the informative samples would be updated by disease.

Conclusions

Looking ahead, future work should involve a comprehensive comparison of likelihoods for given classifications across different metrics. The goal aims to enhance our understanding of diagnosis likelihoods and provide more nuanced insights into the performance of classifiers. The proposed inclusion of multiple metrics; CLOUD, microbially informed forests, and forests implemented with metadata offers a multifaceted evaluation of sample misclassifications.

Furthermore, an examination by an expert for the analysis for each disease report, particularly in the context of identifying pathogens is important. The investigation into the presence of *Fenollaria massiliensis* in the male microbiome exemplifies the potential significance of individual pathogens in disease diagnosis.

More future work includes a closer examination into the filtering of samples, with the goal of establishing an automated method for identifying informative samples tailored to each specific disease. This approach ensures the continued effectiveness of the classification model with new datasets and emerging diseases.

This study contributes insights into the complexities of utilizing SMOTE in microbiome-based disease diagnosis. The challenges identified underscore the importance of a nuanced and context-specific approach when integrating synthetic data augmentation techniques into classifier training for clinical applications.

The analysis indicates that metadata-informed disease predictions offer valuable insights, in addition to microbially informed data. Additionally, CLOUD emerges as a promising avenue, providing valuable insights into samples that may be susceptible to misclassification.

References

[1] N. Thevaranjan, A. Puchta, C. Schulz, A. Naidoo, J. Szamosi, C. P. Verschoor, D. Loukov, L. P. Schenck, J. Jury, K. P. Foley, J. D. Schertzer, M. J. Larché, D. J. Davidson and E. F. Verdú, 12 April 2017. [Periodical]. Cell Host & Microbe 21, no. 4: 1931-3128. https://doi.org/10.1016%2Fj.chom.2017.03.002. [Accessed 2023].

[2] R. D. Hills, B. A. Pontefract, H. R. Mishcon, C. A. Black, S. C. Sutton and C. R. Theberge, "Gut Microbiome: Profound Implications for Diet and Disease," 16 July 2019. [Periodical]. Nutrients 11, no. 7: 1613. https://doi.org/10.3390/nu11071613 https://doi.org/10.3390%2Fnu11071613.

[3] D. McDonald, Y. Jiang, M. Jiang and et al., "Greengenes2 unifies microbial data in a single reference tree," 27 July 2023. [Periodical]. Nat Biotechnol (2023). https://doi.org/10.1038/s41587-023-01845-1.

[4] S. Chakravorty, D. Helb and M. Burday, "A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.," 22 February 2007. [Periodical]. J Microbiol Methods. 2007 May;69(2): 330-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2562909/.

[5] A. Amir, D. McDonald, J. A. Navas-Molina, E. Kopylova, J. T. Morton, Z. Xu, E. P. Kightley, L. R. Thompson, E. R. Hyde, A. Gonzalez and R. Knight, "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns," 7 March 2017. [Periodical]. mSystems 2, no. 2: 10.1128/msystems.00191-16. https://doi.org/10.1128/msystems.00191-16.

[6] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson and S. P. Holmes, "DADA2: High-resolution sample inference from Illumina amplicon data," 16 May 2016. [Periodical]. Nat Methods 13, 581–583 (2016). https://doi.org/10.1038/nmeth.3869.

[7] D. McDonald, B. Kaehler, A. Gonzalez, J. DeReus, G. Ackermann, C. Marotz, G. Huttley and R. Knight, "redbiom: a Rapid Sample Discovery and Feature Characterization System," 5 June 2019. [Periodical]. mSystems 4, no. 4: 10.1128/msystems.00215-19. https://doi.org/10.1128/msystems.00215-19.

[8] Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kos, "Qiita: rapid, web-enabled microbiome meta-analysis," 1 October 2018. [Periodical]. Nat Methods 15, 796–798 (2018). https://doi.org/10.1038/s41592-018-0141-9.

[9] D. McDonald, E. Hyde, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov and R. Knight, "American Gut: an Open Platform for Citizen Science Microbiome Research," 15 May 2018. [Periodical]. mSystems 3, no. 3: 10.1128/msystems.00031-18. https://doi.org/10.1128/msystems.00031-18.

[10] Mario, "Annotate the quartiles with Matplotlib in a normal distribution plot," 16 October 2021. [Online]. Available: https://stackoverflow.com/questions/43360414/annotate-the-quartiles-with-matplotlib-in-a-normal-distribution-plot/69595392#69595392.

[11] J. M. Russell, "Significant Statistics: An Introduction to Statistics," 2020. [Book]. Available: https://pressbooks.lib.vt.edu/introstatistics/chapter/measures-of-the-spread-of-the-data/.

[12] I. Abellan-Schneyder, M. S. Matchado, S. Reitmeier, A. Sommer, Z. Sewald, J. Baumbach, M. List and K. Neuhaus, "Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing," [Periodical]. mSphere 6, no. 1: 10.1128/msphere.01202-20. http://dx.doi.org/10.1128/mSphere.01202-20.

[13] E. A.-G. G. H. B. e. a. Montassier, "CLOUD: a non-parametric detection test for microbiome outliers," 6 August 2018. [Periodical]. Microbiome 6, 137 (2018). https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0514-4.

[14] C. N. Bernstein and J. D. Forbes. [Periodical]. Inflamm Intest Dis 8 November 2017; 2 (2): 116–123. https://doi.org/10.1159%2F000481401.

[15] "UK Standards for Microbiology Investigations," 2 April 14. [Periodical]. UK Standards for Microbiology Investigations. ID 14 Issue 3. https://assets.publishing.service.gov.uk/media/5a801172ed915d74e622c47e/ID_14i3.pdf.

[16] M. T. France, J. Clifford, S. Narina, L. Rutt and J. Ravel, "Complete Genome Sequences of Ezakiella coagulans C0061C1 and Fenollaria massiliensis C0061C2," 5 July 2022. [Periodical]. Microbiology Resource Announcements 11, no. 7: e00444-22. https://journals.asm.org/doi/10.1128/mra.00444-22.

[17] B. J. Gardiner, A. Y. Tai, D. Kotsanas, M. J. Francis, S. A. Roberts, S. A. Ballard, R. K. Junckerstorff and T. M. Korman, "Clinical and Microbiological Characteristics of Eggerthella lenta Bacteremia," J Clin Microbiol. 2015 Feb;53(2):626-35. [Periodical]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4298500/.

[18] S. Finegold, P. Summanen, S. H. Gerardo and E. Baron, "Clinical importance of Bilophila wadsworthia," November 1992. [Periodical]. J Clin Microbiol. 2015 Feb;53(2):626-35. https://pubmed.ncbi.nlm.nih.gov/1295759/.