Into the Gut Microverse

Nissy Milcia William, Shourya
School of Biological Sciences
National Institute of Science Education and Research Bhubaneswar
Jatani, Khordha, Odisha, PO: 752050
nissymilcia.w@niser.ac.in, shourya.2021@niser.ac.in

Abstract

The Microbiome is a complex assembly of microbes, such as bacteria, fungi, viruses, and their genes, that naturally live inside and outside our bodies. Microbiomes thriving in the gut play a huge role in regulating health, behaviour, and physiology. We aimed to create a machine learning model that could predict human gastrointestinal diseases based on gut microbiome composition. Our dataset was obtained from a github repository. This data was then pre processed, followed by feature selection using filter and wrapper methods, assessed by dimensionality reduction models. Multiple models were trained using our data and the optimal one, Random Forest, was chosen. Models trained on individual datasets and those trained on a combined dataset were compared and contrasted. We found that models trained on individual datasets tended to have more accurate predictions as opposed to models trained on the combined datsaset, and our feature reduction attempts appear to be successful. Further experiments will use neural network models and integrate an analysis of metabolome profile.

1 Introduction

The Microbiome is a complex assembly of microbes, such as bacteria, fungi, viruses, and their genes, that naturally live inside and outside our bodies. Microbiomes thriving in the gut play a huge role in regulating health, behaviour, and physiology. It is one of the most studied microbiomes, harbouring an interface connecting the external and internal milieu.

Microbiomes readily undergo dysbiosis by abiotic and biotic factors, which leads to several gut-related disorders. Studies showed that diseased human faecal samples have less diversity of microorganisms, primarily caused by dysbiosis.[1]

The Microbiome itself, as well as the interaction of the Microbiome with the host, secrete metabolites, which can act as epigenetic factors that regulate gene-related gut disorders. So, in this machine learning project, we focused on gut-related disorders and tried to establish a relationship between different microbial abundance and metabolites present in the gut, estimated by metagenomic analysis of the faecal samples, taking Microbial abundance and different metabolites as features.

We are expecting our model will learn from various datasets and will be able to differentiate whether the test is healthy or not.

2 Methods

2.1 Data Collection

Data was collected from a GitHub repository (https://github.com/borenstein-lab/microbiome-metabolome-curated-data.git) containing bacterial abundance data and patient metadata. The dataset was extracted, comprising features related to microbial abundance and patient characteristics.

2.2 Data Preprocessing

The collected data underwent several preprocessing steps. Firstly, missing values were addressed through imputation for rare occurrences and feature elimination for commonly missing values. Additionally, certain continuous numerical features were binned into categorical classes to simplify analysis. Non-numerical features were encoded as necessary for compatibility with machine learning algorithms. Summary statistics and exploratory data analysis were conducted to understand the dataset's distribution.In order to ensure the validity of subsequent statistical analyses, the normality of the data distribution was assessed using the Shapiro-Wilk test. This test is a commonly employed method for evaluating the normality assumption of a dataset. Following the Shapiro-Wilk test, the data was found to deviate significantly from a normal distribution, and hence non-parametric tests such as the Mann-Whitney U test and the Kruskal-Wallis method were deemed appropriate for subsequent analysis, as they do not rely on the assumption of normality.

Moreover, the balance of the dataset with respect to the labels was evaluated to ascertain whether there were any significant disparities in the distribution of samples across different groups or categories.

2.3 Data Scaling

To ensure uniformity in feature scaling, the data was normalized using min-max scaling, scaling values between 0 and 1.

2.4 Feature Engineering

Firstly, for feature selection, we employed two distinct methods: the filter method and the wrapper method. The filter method evaluates features based on their intrinsic properties, such as distribution and sample adequacy for each group. To this end, we utilized the Mann Whitney U test and the Kruskal Wallis method. These non-parametric tests are particularly suited for scenarios where the assumptions of normality and homogeneity of variance are violated. The Mann Whitney U test is appropriate for comparing two independent groups, while the Kruskal Wallis method extends this comparison to more than two groups. These methods rank features based on their discriminatory power and facilitate the selection of relevant features for subsequent analysis.

We then employed the wrapper method, specifically Recursive Feature Elimination with Cross-Validation using Random Forest (RFECV-RF). Unlike the filter method, which relies solely on intrinsic feature properties, the wrapper method evaluates features based on their contribution to the predictive performance of a specific model criterion. RFECV-RF iteratively selects subsets of features and assesses their performance through cross-validation, thereby identifying the optimal subset that maximizes model performance.

Additionally, Principal Component Analysis (PCA) was employed to assess whether further dimensionality reduction could be achieved beyond the selected features. PCA identifies patterns in data and reduces its dimensionality while retaining most of its variance, thereby aiding in visualizing high-dimensional data and potentially simplifying subsequent analyses.

2.5 Model Selection

The K-nearest neighbors (kNN) algorithm was initially considered. However, its application encountered limitations, primarily due to low separation between data points within the feature space. This resulted in lower accuracy, prompting a reassessment of alternative models.

Subsequently, both the XGBoost and Random Forest algorithms were evaluated for their suitability in handling the dataset. Both models demonstrated higher accuracy in predictive performance; nevertheless, Random Forest emerged as the preferred choice for several reasons.

Random Forest was selected due to its capacity to effectively handle large datasets, a notable advantage in the context of our study's data volume. Additionally, its non-parametric nature circumvents assumptions regarding data distribution, thereby reducing bias and enhancing the model's robustness across diverse datasets. Furthermore, Random Forest exhibits faster computational speed compared to XGBoost, an attribute crucial for expedited analyses and efficient resource utilization.

3 Results

3.1 The number of features underwent a substantial decrease after processing and feature selection

Following the processing and feature selection utilizing the Mann-Whitney U test and Kruskal-Wallis test, a substantial reduction in the number of features was achieved across all studies. This reduction is depicted in Table 1, detailing the number of features before and after processing for each study.

Table 1: Number of Features Before and After Processing

Study	# Features Before Processing	# Features After Processing
Erawinjintari	10529	621
Franzosa	11722	45
iHMP	9696	1176
Kim	501	23
Mars	2897	146
Sinha	88	88
Yachida	11944	734

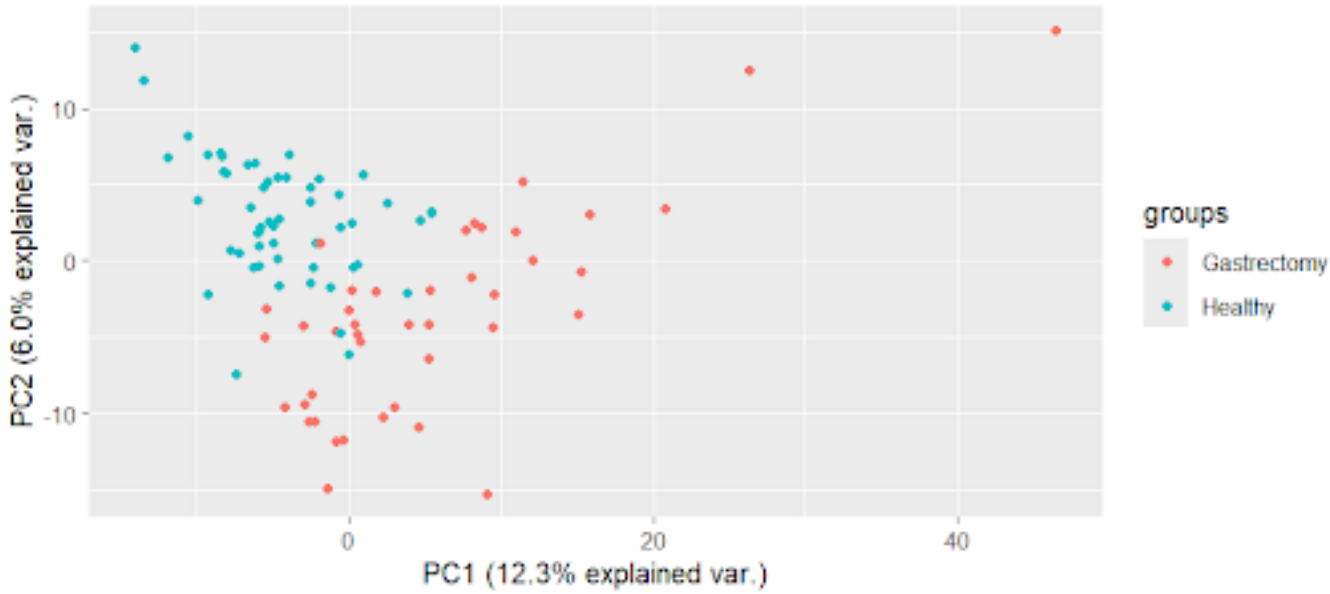
The Mann-Whitney U test, a non-parametric statistical method, compares the variance of a feature across different labels or groups within a dataset. Here we used it for studying microbial community differences between health states, such as comparing the abundance of bacterial genera between "Healthy" and "Disease".

The null hypothesis assumed that there is no difference in the distribution of bacterial abundance between the Healthy and Disease groups, i.e, the probability of a randomly selected value from one group (e.g., Healthy) being greater than a randomly selected value from the other group (e.g., Disease) is equal to the probability of a randomly selected value from the Disease group being greater than that from the Healthy group.

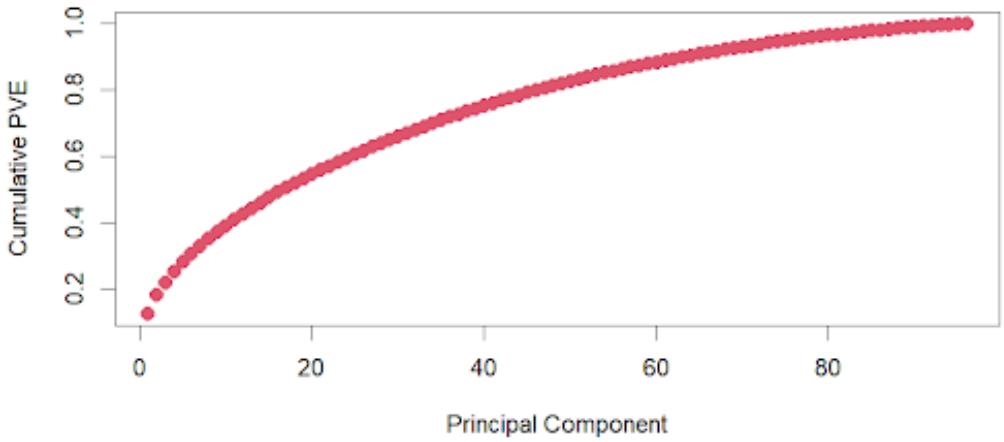
If the null hypothesis was true, it suggested that the bacterial genus did not vary in abundance between the health states and thus was not relevant for distinguishing between these conditions. In contrast, the alternative hypothesis was that there was a difference between the groups. This could manifest as one group consistently exhibiting higher or lower abundance of the bacterial genus compared to the other. If the alternative hypothesis was supported by the data, it implied that the abundance of this bacterial genus was statistically significant and hence was used to predict the disease or health state. The Kruskal-Wallis method similarly evaluated whether there are significant differences in the distributions of a feature across multiple independent groups.

Upon application of these feature selection methods, a notable reduction in the number of features was observed across all studies. For instance, in the Erawinjintari study, the number of features decreased from 10,529 to 621 after processing, showcasing a substantial reduction in dimensionality. Similar reductions were observed in the other studies, with the number of features decreasing significantly after processing.

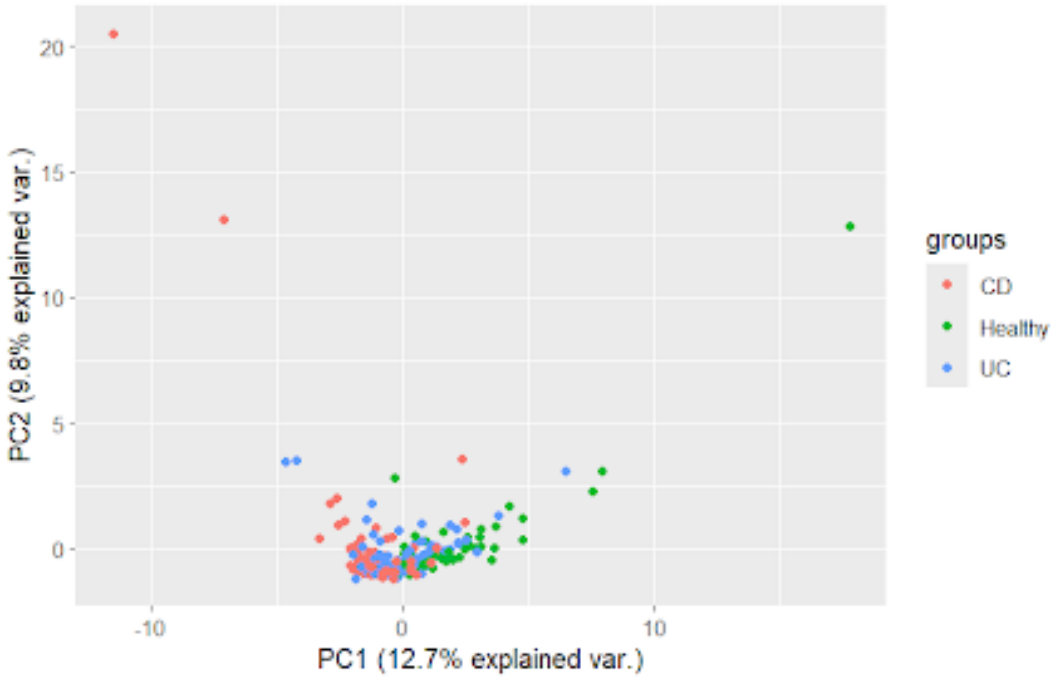
3.2 Principal Component Analysis (PCA) was utilized to evaluate whether additional reduction in the number of features was feasible



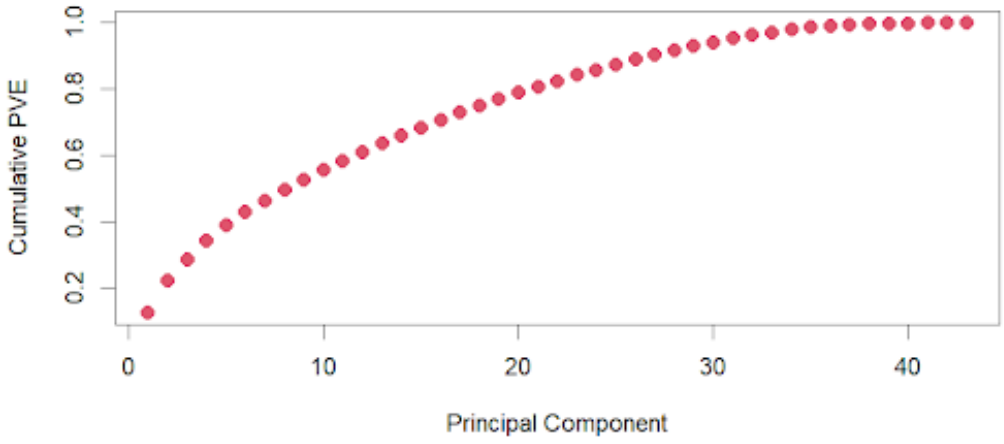
(a) Erawinjintari PCA



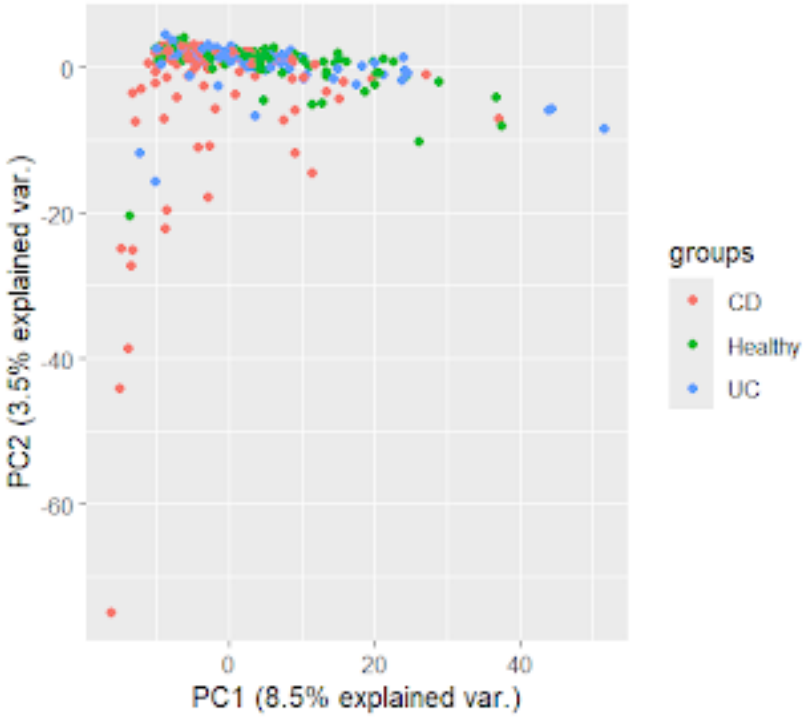
(b) Erawinjintari PCA Cumulative sum plot



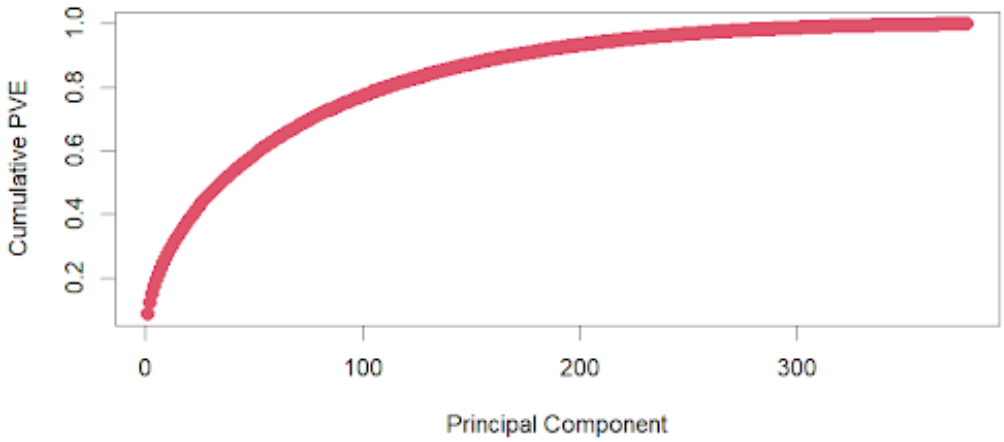
(a) Franzosa PCA



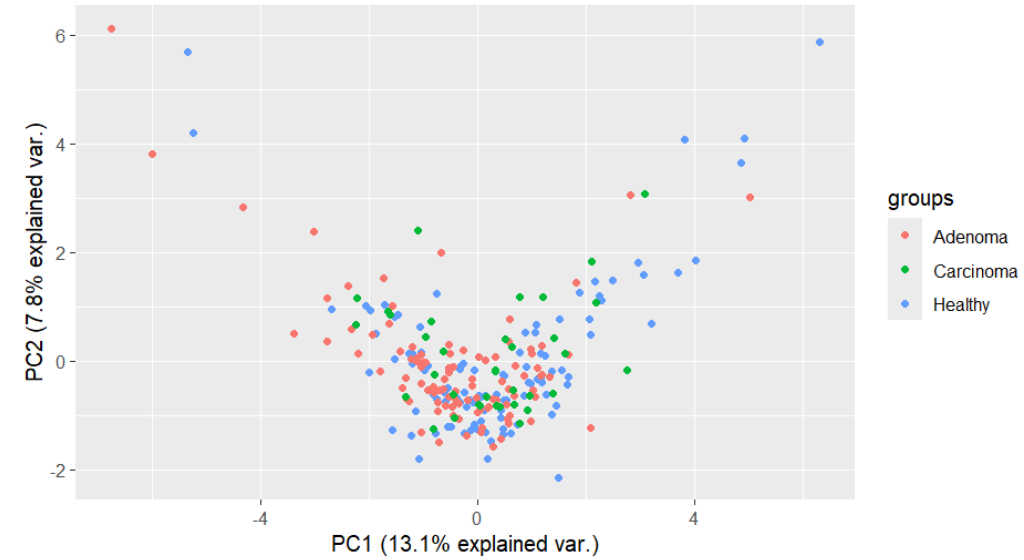
(b) Franzosa PCA Cumulative sum plot



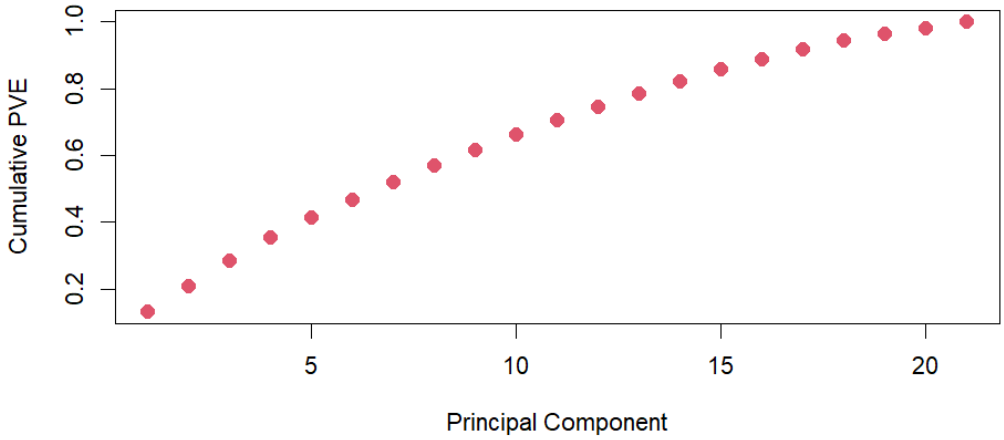
(a) iHMP PCA



(b) iHMP PCA Cumulative sum plot

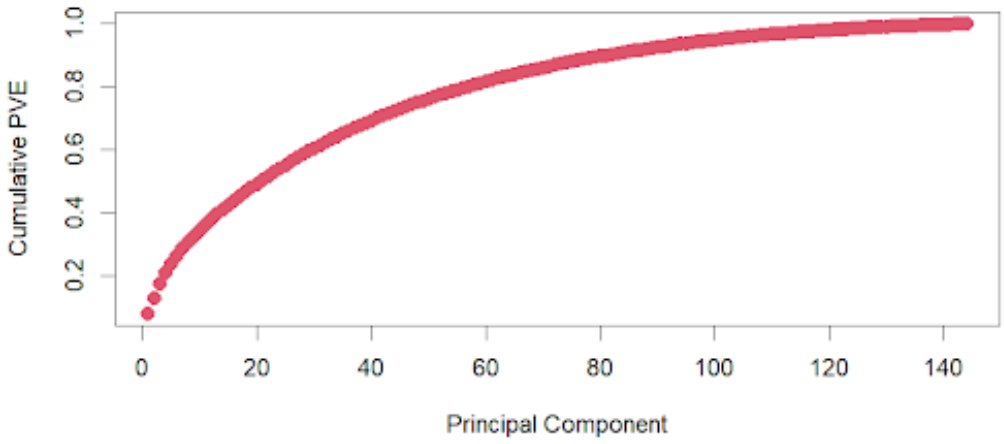


(a) Kim PCA

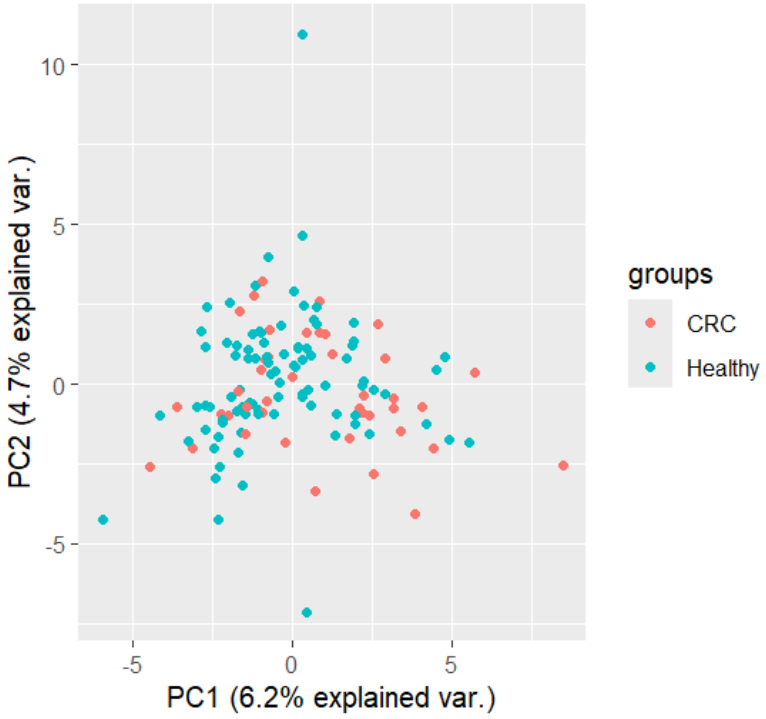


(b) Kim PCA Cumulative sum plot

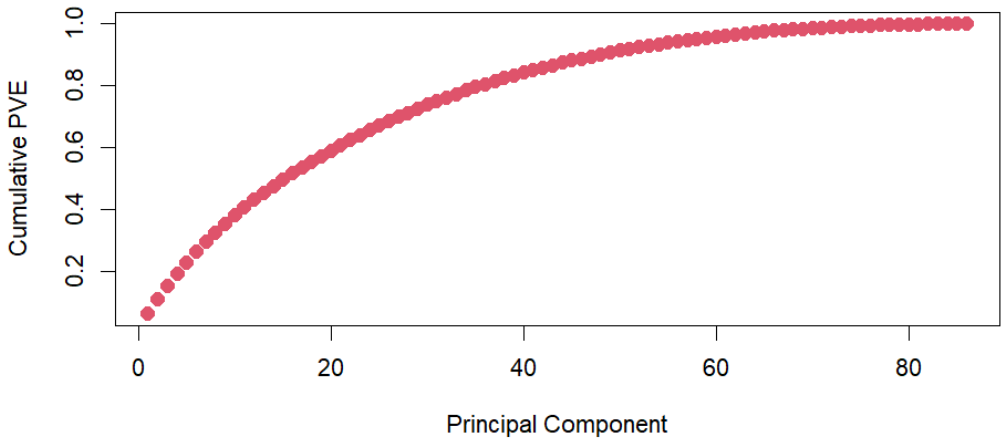
(a) Mars PCA



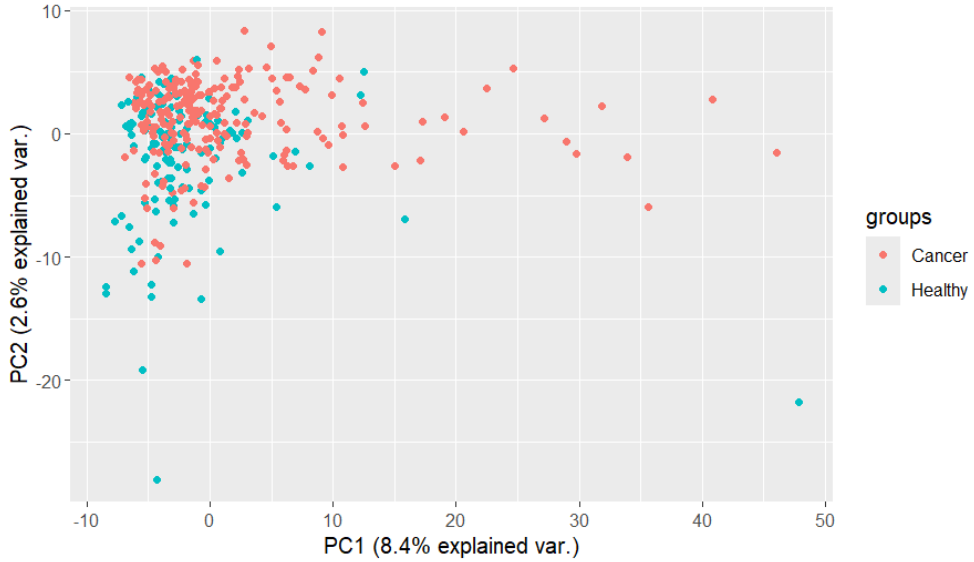
(b) Mars PCA Cumulative sum plot



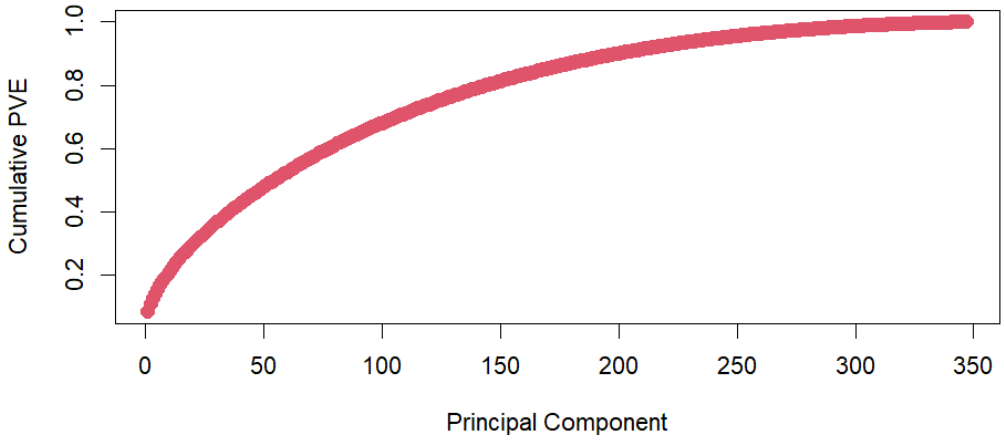
(a) Sinha PCA



(b) Sinha PCA Cumulative sum plot



(a) Yachida PCA



(b) Yachida PCA Cumulative sum plot

Before and after processing, we applied PCA to the dataset. Directly applying PCA to high-dimensional datasets can be computationally expensive, especially when dealing with large datasets with many features. Hence we selected features using the methods described above and followed them with PCA to check if the number of features could be further reduced after feature selection.

3.3 Comparison of Processed and Combined Datasets versus Unprocessed and Uncombined Datasets

F1 score is a metric used to evaluate the performance of a classification model, particularly when dealing with imbalanced classes. It is the harmonic mean of precision and recall, providing a single numerical value that takes into account both false positives and false negatives.

$$Precision = \frac{TruePositives}{FalsePositives + TruePositives}$$

$$Recall = \frac{TruePositives}{FalseNegatives + TruePositives}$$

$$F1Score = \frac{2Precision + Recall}{Precision \times Recall}$$

Table 2: F1 Scores

	Erawinjintari	**Franzosa**	**iHMP**	**Kim**	**Mars**	**Sinha**	**Yachida**
Unprocessed, combined	0.332	0.511	0.565	0.34	0.762	0.581	0.601
Unprocessed, uncombined	0.85	0.607	0.523	0.403	0.757	0.697	0.452
Processed, combined	0.231	0.314	0.397	0.204	0.322	0.227	0.305
Processed, uncombined	0.796	0.747	0.631	0.473	0.903	0.635	0.578

The comparison of F1 scores between processed and unprocessed datasets, both in combined and uncombined forms, reveals notable variations in performance across different datasets. In the unprocessed datasets combined based on common columns, the F1 scores varied from 0.332 for Erawinjintari to 0.601 for Yachida, with an average F1 score of 0.527. Conversely, in the unprocessed datasets that were not combined but retained only common columns, the F1 scores ranged from 0.85 for Kim to 0.613 on average. Moving to processed datasets, where all columns were combined, the F1 scores notably decreased, ranging from 0.204 for Kim to 0.397 for iHMP, with an average F1 score of 0.286. However, in the processed datasets that were not combined, the F1 scores exhibited a substantial improvement, ranging from 0.473 for Kim to 0.903 for Mars, with an average F1 score of 0.680. Notably, the maximum F1 score was achieved in the unprocessed datasets not combined but retaining only common columns, and in the processed datasets that were not combined.

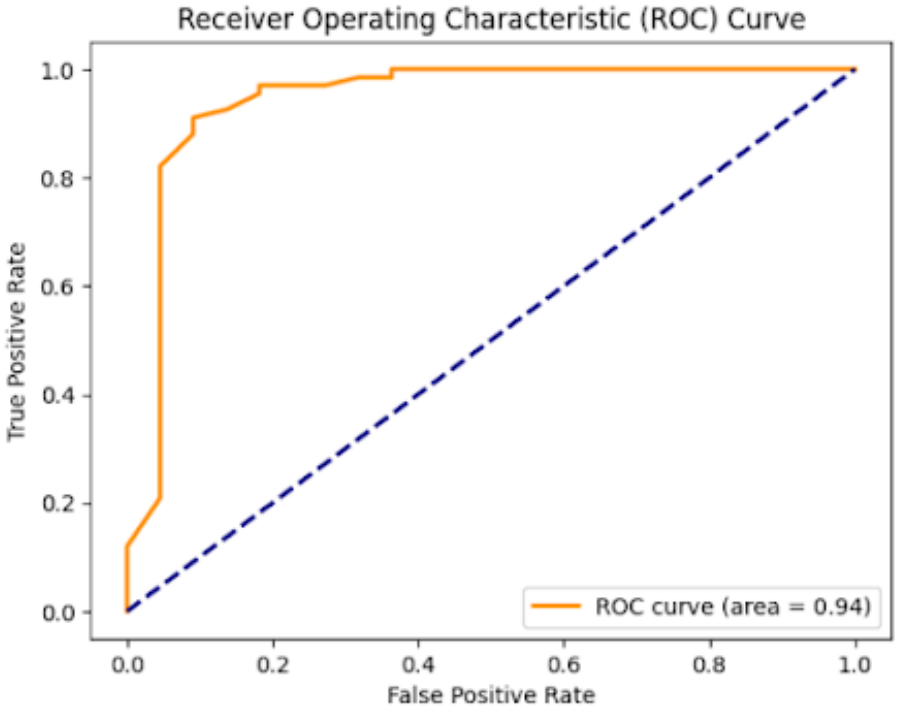


Figure 8: ROC curve for Random Forest classifier of Mars Dataset

An AUC of 0.94 suggests our model is good at correctly classifying positive and negative cases. It can accurately identify a high proportion of true positives while keeping the number of false positives low.

4 Discussion

Our machine learning project investigated the relationship between gut microbiota and gut-related disorders. The analysis employed various statistical and machine learning techniques to extract meaningful insights from the data.

A key aspect of the project involved feature selection, which plays a crucial role in improving model performance and interpretability. This work highlights the effectiveness of non-parametric statistical tests like the Mann-Whitney U test and Kruskal-Wallis test in selecting informative features.

While the Mann-Whitney U test is a valuable initial step for feature selection in microbiome studies, it assumes independence between features, potentially overlooking crucial interactions between bacterial genera and metabolites. These interactions are essential for understanding gut health, as certain microbes can create favorable conditions for others, or their metabolites can influence each other's growth and function. To address this limitation to capture the interactions, we planned to use correlation matrices as they can offer clues about relationships, but they have limitations. Pearson method has inductive bias of linear relationship and spearman assumes the relation is monotone, In gut microbiome studies, relationships might be more complex and nonlinear. The correlation matrix might miss these important relationships and it also doesn't account for cause-and-effect. Additionally, interpreting a large number of correlations in these matrices can be challenging.

Hence, we explored alternative methods. Wrapper methods like Recursive Feature Elimination with Cross-Validation (RFECV) using a Random Forest can inherently consider feature interactions during model building, and analyzing feature importance can reveal which features influence each other more strongly. Since working with raw data containing thousands of features can be computationally

expensive, we employed the Mann-Whitney U test as a filter method to select a more manageable subset of features before applying RFECV.

We conducted hyperparameter tuning for the Random Forest model, configuring it with 100 trees in the forest and setting the maximum number of features to consider for the best split as the square root of the total number of features. This process aimed to optimize the model's performance by systematically exploring various combinations of hyperparameters.

F1 scores provide a more balanced evaluation by considering both precision and recall. The F1 score is the harmonic mean of precision and recall. It balances these two metrics, making it useful when there's a trade-off between them. A high F1 score indicates good performance in both precision and recall, whereas a high accuracy can sometimes mask poor performance in one of these metrics. Hence we opted to utilize the F1 scores as our primary metrics as opposed to accuracy for comparing the performance of different models.

Our strategy of employing filter methods followed by wrapper methods for feature selection led to a substantial decrease in computation time. This enabled us to perform the analysis using readily available CPU resources.

Further attempts at dimension reduction as well as introducing a threshold for the number of bacterial genera considered in the studies did not appear to increase the accuracy of predictions.

5 Future prospects

In future endeavors, we intend to explore the versatility of neural network architectures, notably multilayered perceptrons (MLPs). Neural networks offer inherent advantages in capturing complex patterns and relationships within high-dimensional datasets, such as metabolomic profiles. By harnessing the capabilities of MLPs, we anticipate enhanced model performance and robustness, thereby facilitating more accurate disease classification and prognosis.

Secondly, we plan to conduct comprehensive analyses and model training procedures utilizing metabolome data. Metabolomic profiles provide a holistic representation of an individual's biochemical status, offering unique insights into underlying physiological processes and disease mechanisms. By integrating machine learning techniques with metabolomic data we aim to refine predictive models.

6 References

1. Lloyd-Price, J., Arze, C., Ananthakrishnan, A.N. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569, 655–662 (2019). https://doi.org/10.1038/s41586-019-1237-9
2. Franzosa, E.A., Sirota-Madi, A., Avila-Pacheco, J. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. Nat Microbiol 4, 293–305 (2019). https://doi.org/10.1038/s41564-018-0306-4
3. Yachida, S., Mizutani, S., Shiroma, H. et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. Nat Med 25, 968–976 (2019). https://doi.org/10.1038/s41591-019-0458-7
4. "Identification of Biomarkers to Diagnose Diseases and Find Adverse Drug Reactions by Metabolomics". Drug Metabolism and Pharmacokinetics, 1 Jan. 2020, https://www.sciencedirect.com/science/article/pii/S1347436720304341.
5. Fecal Microbiota, Fecal Metabolome, and Colorectal Cancer Interrelations. (2016, March 31). PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152126
6. Faecal microbiome-based machine learning for multi-class disease diagnosis. (2024, April 20). Faecal microbiome-based machine learning for multi-class disease diagnosis. Nature Communications. https://www.nature.com/articles/s41467-022-34405-3
7. Learning representations of microbe–metabolite interactions. (n.d.). Learning representations of microbe–metabolite interactions. Nature Methods. https://www.nature.com/articles/s41592-019-0616-3

8. https://github.com/borenstein-lab/microbiome-metabolome-curated-data
9. https://www.ibm.com/topics/exploratory-data-analysis
10. ttps://www.stat.berkeley.edu/ breiman/randomforest2001.pdf
11. https://scikit-learn.org/stable/