**Machine Learning Prediction of Obesity from Taxa and Functional Profiles in Human Gut Microbiota**

**Danesh Moradigaravand**

**Under the supervision of Prof Babak Khalaj**

**Sharif University of science and technology**

**Summary:**

The human gut harbors a diverse community of microorganisms collectively known as the gut microbiota, which plays a pivotal role in various aspects of human health, encompassing metabolism and immune function. Recent research has begun to unravel the potential links between alterations in the gut microbiota and the development of metabolic disorders, including obesity. This article delves into the intricate relationship between the gut microbiota and obesity, shedding light on the existing evidence and potential mechanisms involved. Emerging evidence highlights the association between changes in gut microbiota composition and function and obesity, thereby suggesting that the gut microbiota's composition could hold significant predictive information for the early diagnosis and prognosis of obesity. In this project, we harnessed the power of machine learning algorithms to predict obesity status using taxa and functional information derived from the gut microbiota of 247 human samples. Our findings demonstrated high predictive accuracy (AUC-ROC = 0.95) and allowed for the identification of key taxa and pathways implicated in obesity. To make our predictive model accessible to the research community, we developed a user-friendly data science solution featuring both a command line interface and a graphical interface.

**Introduction:**

**>>> Microbiota and its Role in Metabolic Disease**

The gut microbiota consists of trillions of microorganisms, including bacteria, viruses, fungi, and other microbes. These microorganisms interact with each other and with the host's body, influencing various physiological processes. One of the key roles of the gut microbiota is its involvement in metabolism. The gut microbiota contributes to the breakdown and absorption of dietary nutrients, produces metabolites, and influences energy balance. Dysregulation of the gut microbiota can disrupt these functions, leading to metabolic imbalances and contributing to the development of metabolic diseases.

**>>> Definition of Obesity**

Obesity is a complex and multifactorial condition characterized by excess body fat accumulation. It is commonly defined based on body mass index (BMI), which is calculated as weight in kilograms divided by the square of height in meters. However, obesity is not solely determined by an individual's body weight. It involves an intricate interplay of genetic, environmental, and lifestyle factors. Obesity is associated with a range of health complications, including type 2 diabetes, cardiovascular disease, and certain types of cancer. The Middle East, including Iran, is facing a significant and growing challenge in the form of obesity. In recent years, there has been a notable rise in obesity rates across the region, with Iran being particularly affected. The combination of rapid urbanization, sedentary lifestyles, unhealthy dietary patterns, and cultural factors has contributed to the alarming increase in obesity prevalence. Obesity in Iran poses a multifaceted problem with far-reaching consequences for public health. The burden of obesity-related diseases, such as diabetes, cardiovascular disorders, and certain cancers, has increased substantially in recent years. Addressing the issue requires a comprehensive approach that encompasses awareness campaigns, policy interventions, and the promotion of healthy lifestyles to combat the escalating rates of obesity and its associated health risks in Iran.

**>>> Links and Evidence between Obesity and Microbiota**

Mounting evidence suggests that alterations in the gut microbiota composition and function are associated with obesity. Studies have shown that obese individuals often have an altered gut microbiota compared to lean individuals, characterized by reduced microbial diversity and an overrepresentation of certain bacterial species. Furthermore, animal studies have demonstrated that transplantation of gut microbiota from obese individuals into germ-free mice can induce weight gain and metabolic abnormalities, highlighting the potential role of the gut microbiota in obesity development.

**>>> Sequencing technologies and machine learning**

The composition of the gut microbiota can be analyzed using advanced sequencing technologies, which offer valuable insights into the structure and diversity of the microbial community. By combining this information with genomic data, such as genetic variants and gene expression profiles, researchers can uncover potential interactions between the host and microbiota that influence complex traits. Furthermore, pathway analysis enables the examination of biological processes and molecular pathways through which the gut microbiota may impact human physiology and disease. In this study, we created a curated database containing information on taxa and pathway abundance for both obese and non-obese groups of healthy individuals. Using this labeled dataset, we trained machine learning models to predict the labels of obese versus non-obese individuals. Subsequently, we utilized interpretability features to identify the most influential features for obesity prediction.

**Methods**

**Input data**

In this project, we used the a standardized database of HumanMetagenomeDB. The main goal of the HumanMetagenomeDB is to simplify the identification and use of public human metagenomes of interest. HumanMetagenomeDB version 1.0 contains metadata of 69 822 metagenomes. The database standardized 203 attributes, based on standardized ontologies, describing host characteristics (e.g. sex, age and body mass index), diagnosis information (e.g. cancer, Crohn's disease and Parkinson), location (e.g. country, longitude and latitude), sampling site (e.g. gut, lung and skin) and sequencing attributes (e.g. sequencing platform, average length and sequence quality). Further, HumanMetagenomeDB version 1.0 metagenomes encompass 58 countries, 9 main sample sites (i.e. body parts), 58 diagnoses and multiple ages, ranging from just born to 91 years old. We fetched the short read sequencing for 247 individuals, consisting of 130 obese and 147 non-obese individuals. The non-obese group included normal and underweight groups. The data was fetched with sratoolkit using the accession number. An average each sample had 457Mb size. The samples were from individuals from various countries and age groups to remove any confounding effect. We also took samples sequenced on Illumina platforms. Accession numbers and specification of the samples are provided in the GitHub directory of the project.

**Metagenomic pipeline**

We then fetched the short-read sequencing and subject them to pipelines for taxa and pathway abundance identification. The bioBakery pipeline, along with the MetaPhlAn and HUMAnN tools, represents a powerful suite of computational tools designed for analyzing metagenomic data and unraveling the functional potential of microbial communities. The pipeline begins with MetaPhlAn, which provides taxonomic profiling by accurately quantifying the abundance of microorganisms at the species level based on marker genes. MetaPhlAn utilizes a unique set of clade-specific marker genes, enabling high-resolution profiling even for low-abundance species.

Following taxonomic profiling, HUMAnN (HMP Unified Metabolic Analysis Network) is employed to predict functional profiles of the microbial community. HUMAnN integrates MetaPhlAn's taxonomic abundance estimates with a comprehensive reference database of microbial gene families. By aligning metagenomic reads to this database, HUMAnN can quantify the abundance of functional gene families and pathway modules. This approach allows researchers to gain insights into the potential metabolic capabilities and functional diversity of the microbial community. We curated the output of the two tools as tabular data for the machine learning pipeline. We considered the phylum and classes from taxa for prediction. The output files were deposited to the GitHub directory of the project.

**Machine learning pipeline**

In our study, we employed a standard approach to split the data for machine learning into 80% training and validation sets, and 20% test set. This division allowed us to assess the performance and generalization ability of our models effectively. Additionally, we utilized 5-fold cross-validation during the training phase to further evaluate and validate the models' performance, reducing the risk of overfitting. This approach helped us optimize our models and select the best performing one for further analysis.

To prepare the input data for machine learning, we vectorized the features to transform them into a numerical representation that could be processed by the models. This process involved encoding categorical variables, scaling numerical variables, and handling missing values appropriately. By vectorizing the input data, we ensured that the models could effectively learn from the available features and make accurate predictions.

As baseline models, we utilized logistic regression classifier, a widely used and interpretable method for binary classification tasks. To compare its performance, we also employed a gradient boosted decision tree model using the scikit-learn library. We performed a hyperparameter grid search to tune the tree depth and the number of iterations in the gradient boosted model, and the regularization penalty term in logistic regression. This search allowed us to find the optimal combination of hyperparameters that maximized the area under the receiver operating characteristic curve (AUC-ROC), a common evaluation metric for binary classification tasks.

After selecting the best model based on AUC-ROC, we conducted SHAP (SHapley Additive exPlanations) feature importance analysis. SHAP is a model-agnostic technique that quantifies the contribution of each feature to the model's predictions. By analyzing the SHAP values, we gained insights into the importance and impact of each feature, i.e. taxa and pathways, on the model's decision-making process. This analysis helped us understand the underlying mechanisms and drivers of the predictive performance, providing valuable insights for feature selection, interpretation, and further investigations.

**Deployment of machine learning**

To enhance the usability and accessibility of our machine learning model, we transformed it into both a command-line application and a Streamlit-based web application that can predict the output for any arbitrary input. By creating a command-line application, users can interact with the model directly from the terminal by providing input data and receiving the predicted output. This command-line interface enables automation and integration into existing workflows.

Furthermore, we developed a Streamlit application, a user-friendly web framework, to provide a visually appealing and intuitive interface for users to input their data and obtain predictions instantly. The Streamlit application allows users to upload a file or manually enter data through a user interface, and then displays the predicted output along with any additional visualizations or insights generated by the model.

By offering both a command-line application and a Streamlit-based web application, we cater to different user preferences and usage scenarios. The command-line application suits users who prefer a command-driven workflow, while the Streamlit application appeals to those who prefer a graphical interface. These applications allow users to make predictions using our machine learning model with ease and convenience, regardless of their technical expertise or familiarity with coding.

**Code and data availability**

All the codes and row and cleansed datasets are deposited to the GitHub directory of this project:

**Results and Discussion:**

**>>> Abundance of classes in the input predictors**

We predicted obese and non-obese labels from the microbiota data for 247 individuals. The input dataset consisted of 259 features, including 23 features related to taxa abundance and the remaining features representing pathways (Figure 1). Figure 1A displays the most frequently occurring taxa in the dataset. At the phylum level, the majority of strains belonged to Firmicutes, Bacteroidetes, and Actinobacteria. At the next taxonomic level, class, Clostridia and Bacteroidia were dominant, as expected for the human microbiota. This confirms the reliability of the taxa and pathway abundance pipelines. Regarding pathway abundance, we identified multiple pathways with varying frequencies. The different frequencies of these pathways provided valuable prediction information that could be utilized for machine learning-based predictions.

**>>> Performance of the machine learning model**

We trained the machine learning model on the taxa abundance and presence/absence of pathways in the samples. The dataset was split into training and test datasets, and two models were trained. The performance of the baseline logistic regression and gradient boosted decision trees indicates that both models can accurately predict the obese label (Figure 2). The gradient boosted decision trees outperformed logistic regression, with an AUC-ROC value of 0.90 compared to 0.83 (Figure 2A). However, for other prediction metrics, both models seemed to exhibit overfitting, as demonstrated by higher metrics (F1, precision, and recall) for the training and test datasets. On average, the difference was 0.15 (95% CI 0.18-0.2). The logistic regression model showed a higher recall value, while the gradient boosted decision tree yielded better precision (Figure 2B). To further explore these metrics, we extracted them from confusion matrices for gradient boosted and logistic regression models, revealing comparable accuracy. However, the precision and recall for the obese label were higher for the gradient boosted model (Figure 3). Conversely, the recall value for the non-obese label was higher for the logistic regression model. These results overall suggest that the more complex model performs better and justify the choice of a more complex model. Although both models exhibited some level of overfitting, potentially due to the higher number of predictive features relative to the sample size, a larger dataset would be warranted. The high accuracy of predictions on the small dataset indicates strong prediction signals in the microbiota data, which were effectively utilized by the models.

**>>> Feature importance analysis**

We proceeded with the more complex gradient boosted decision model for prediction. However, these models do not provide a mechanistic understanding of the features. To gain insights and quantify the importance of features in the taxa and pathway abundance data, we employed the SHAP (Shapley Additive exPlanations) metric (Figure 4). The SHAP values illustrate the contribution of each feature to the final prediction for each sample, as denoted by the dot points. It assumes the absence of interactions between features (Figure 4A). Furthermore, these values indicate the direction of the effect, whether the predictor predicts the obese label or the non-obese label. Among the top 20 predictor features for the labels, 11 were taxa features and 9 were pathways. Out of the 11 taxa features, 6 were at the class level and 5 were at the phylum level, designated by the prefix "c" for classes and "p" for phylum (Figure 4A).

**>>> Top taxa predictor features**

We next examined the top five taxa features and contextualized them with existing evidence. The most predictive feature was found to be the Bacilli group, whose presence strongly correlates with the obese trait (Figure 4A and 4B). Several studies have observed higher levels of certain Bacillus species in individuals with obesity compared to those with a healthy weight, suggesting a potential positive correlation between Bacilli abundance and obesity. Bacilli encompass various species known for their metabolic capabilities and potential influence on host physiology. Research indicates that the abundance and diversity of Bacilli in the gut microbiota may play a role in obesity development and metabolic dysregulation. Notably, Bacillus subtilis and other Bacilli species have been found to produce short-chain fatty acids (SCFAs) and influence the production of hormones and enzymes involved in fat storage and metabolism. Changes in Bacilli abundance and activity can alter SCFA production and signaling, leading to disruptions in energy balance and fat accumulation.

**>>> Top taxa predictor features**

In contrast to Bacilli, the second predictor taxa, namely Clostridia, exhibited a negative association with obesity (Figure 4A). Clostridia are anaerobic bacteria that encompass several species known for their diverse metabolic capabilities. They play a crucial role in the fermentation of dietary fibers, production of SCFAs, and modulation of host immune responses. Studies consistently show that the abundance of certain Clostridia species is altered in individuals with obesity compared to those with a healthy weight. Specifically, reduced levels of butyrate-producing Clostridia, such as Faecalibacterium prausnitzii and Roseburia spp., have been consistently observed in obese individuals. Our findings align with this negative link between Clostridia and obesity.

**>>> Other predictor taxa**

The next three predictive features for obesity were the class of Negativicutes, the phylum of Actinobacteria, and the class of Deltaproteobacteria. The Negativicutes features were positively associated with obesity, while the link for Deltaproteobacteria and Actinobacteria was either negative or unclear (Figure 4A, 4B). Negativicutes, a class of bacteria, has been found to be increased in individuals with obesity compared to those with a healthy weight. These bacteria are known to produce endotoxins, such as lipopolysaccharides (LPS), which can trigger inflammation and metabolic dysfunction[1]. The higher abundance of Negativicutes in obesity suggests a potential role in promoting low-grade chronic inflammation, a hallmark of obesity-related metabolic disorders.On the other hand, Deltaproteobacteria, another class of bacteria, has been found to be decreased in individuals with obesity compared to lean individuals. These bacteria have diverse metabolic functions, including sulfate reduction, and play a role in maintaining gut homeostasis. The reduced abundance of Deltaproteobacteria in obesity may disrupt sulfate metabolism and gut barrier function, contributing to metabolic dysregulation. For Actinobacteria, a complex link was reported between the presence of only certain strains in obese individuals, making it a less robust biomarker for obesity. Overall, the findings from the most predictive features align with existing reports and suggest that these taxa, particularly Negativicutes, could serve as potential biomarkers for obesity.

**>>> Pathway dominance**

Similar to taxa, some of the pathway abundance features were found to be strong predictors of obesity. However, linking these features to obesity may not be as straightforward as for taxa, as the abundance might only reflect the over-abundance of a particular taxa that harbors the pathways. Nevertheless, the most predictive pathways align with their confirmed metabolic roles. Among these, the most predictive pathway feature for obesity was PWY-7198 (Figure 4A and 4B), which corresponds to Pyrimidine deoxyribonucleotides de novo biosynthesis IV. This specific metabolic pathway, involved in the synthesis of DNA building blocks, has been implicated in obesity and metabolic dysfunction. The pathway plays a crucial role in providing the necessary components for DNA replication and cellular proliferation. Dysregulation of pyrimidine deoxyribonucleotide biosynthesis IV has been observed in individuals with obesity, suggesting a potential involvement in the development of the condition.

In contrast to PWY-7198, the pathway PWY-5189 was found to be a strong predictor of obesity (Figure 4A and 4B). Tetrapyrrole biosynthesis II, specifically from glycine, is a metabolic pathway that plays a vital role in the production of heme, chlorophyll, and other essential compounds involved in various cellular processes. Emerging evidence suggests a potential link between disturbances in tetrapyrrole biosynthesis II and obesity. Dysregulation of this pathway can disrupt the synthesis of heme, which is a critical component of hemoglobin and numerous enzymes involved in energy metabolism. Heme also plays a role in the regulation of adipocyte differentiation and lipid metabolism. Alterations in tetrapyrrole biosynthesis II may lead to imbalances in heme levels, affecting adipose tissue function and lipid handling. Our results suggest that the absence of this pathway is prevalent in obese individuals. However, the exact mechanism by which this effect is exerted requires further analysis at the metabolome level. Similar to the findings related to taxa, the pathway data show strong predictive power for obesity and could provide biomarkers at another cellular level. Further analysis of the metabolomic signatures could reveal the exact mechanisms of action.

**>>> deployment of the model**

We deployed our model as a data science prototype solution for end users. Both the command line and graphical interface accept the outputs of the biobackery pipeline and predict the associated obesity label, distinguishing between obese and non-obese individuals. Figure 5 illustrates the three features within the graphical interface. Once the user uploads the files, they are fed into the best-performing gradient boosted model. In addition to providing predictions, the tool visualizes the importance of features for the prediction of the label using SHAP local importance plots (Figure 5B). We also allow users to obtain tabular data showcasing the values positively and negatively linked with obesity. This model demonstrates the practicality of machine learning models in facilitating the rapid diagnosis of complex human diseases.

**Discussion and conclusion**

In this project, we conducted training and validation of machine learning models to predict obesity based on the genomic profile of the gut microbiome. The link between gut microbiota and obesity is supported by several proposed mechanisms. These include increased energy extraction from the diet, enhanced fat storage, altered gut barrier function, chronic low-grade inflammation, and modulation of appetite-regulating hormones. Although the precise mechanisms are still under investigation, it is evident that the gut microbiota plays a significant role in energy metabolism and may contribute to the development and progression of obesity. Our study identified and analyzed well-characterized taxa and pathways as predictors in the model.

There are two areas in which future improvements can be made to enhance this study. Firstly, incorporating more genomic data into the predictors would be beneficial. Currently, standardized datasets in the field are limited due to the heterogeneity of data generated under different circumstances. The need for such data is widely recognized, and there is a growing demand for it. With larger datasets, the accuracy of the models can potentially reach a level that would allow for the adoption of these tools as diagnostic aids in clinical settings. Secondly, integrating other omics data as predictors in the model is crucial. This would enable a comprehensive understanding of the complex environment of the gut microbial ecosystem. However, addressing the challenges associated with large-scale and diverse datasets, as well as the integration of multi-omics data, remains a priority for advancing the field.

The study of the gut microbiota and its impact on human health has gained significant attention in recent years. Researchers are increasingly exploring the potential of using gut microbiota composition, genomic information, and pathway analysis to predict complex human traits. These traits, including obesity, diabetes, inflammatory bowel disease, and mental health disorders, result from a combination of genetic, environmental, and microbial factors. By integrating data from gut microbiota composition, genomics, and pathway analysis, we gain insights into the intricate relationships between these factors and develop predictive models for various complex traits. These diseases pose acute challenges in Iran, and the diagnosis and development of new therapies can greatly contribute to improving public health in the country.

**Discussion**

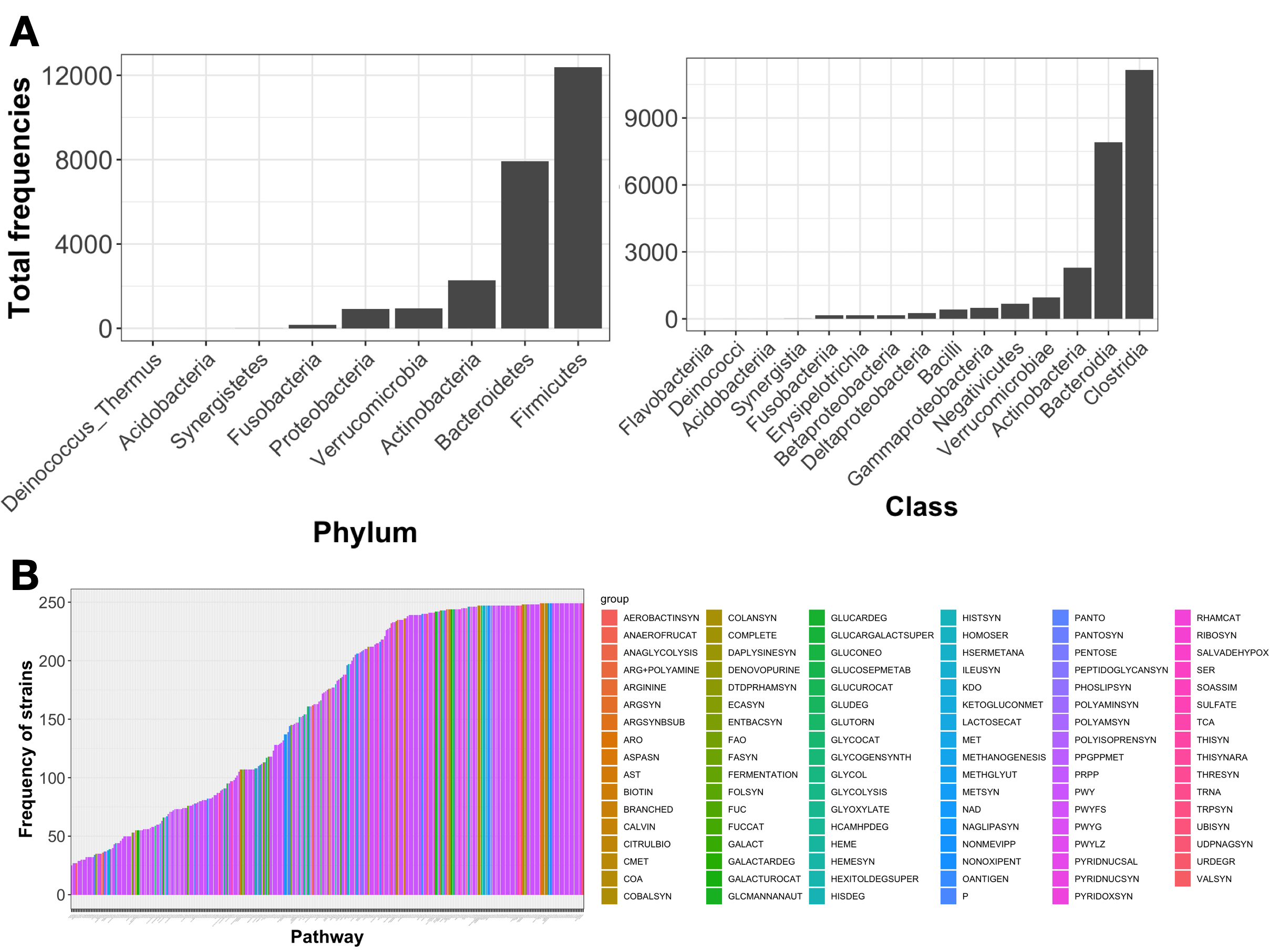


Figure 1 depicts the composition of the input dataset, which includes pathways and taxa. In panel A, the distribution of taxa is presented at both the phylum and class levels. The y-axis represents the summation of the relative frequencies observed across all samples. Panel B illustrates the distribution of pathways, with colors indicating their respective pathway classes. The frequency axis indicates the count of strains that contain each pathway.

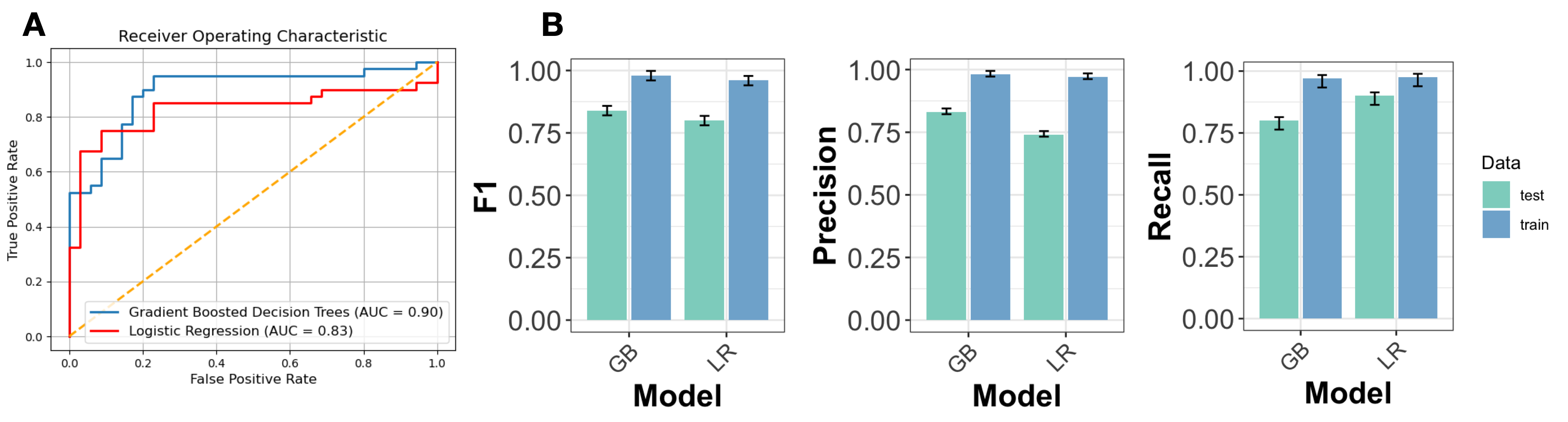


Figure 2 displays the performance of logistic regression and gradient boosted decision trees. Panel A showcases the ROC curves, while panel B presents the classification performance of the two models on both the test and training datasets. The error bars in panel B represent the 95% confidence interval derived from ten random implementations of the models on ten randomly split training and test datasets.

A white background with black text

Description automatically generated with low confidence

Figure 3 presents the confusion matrix and prediction performance for a single split of the training and test dataset, comparing the gradient boosted decision trees and logistic regression classifiers.

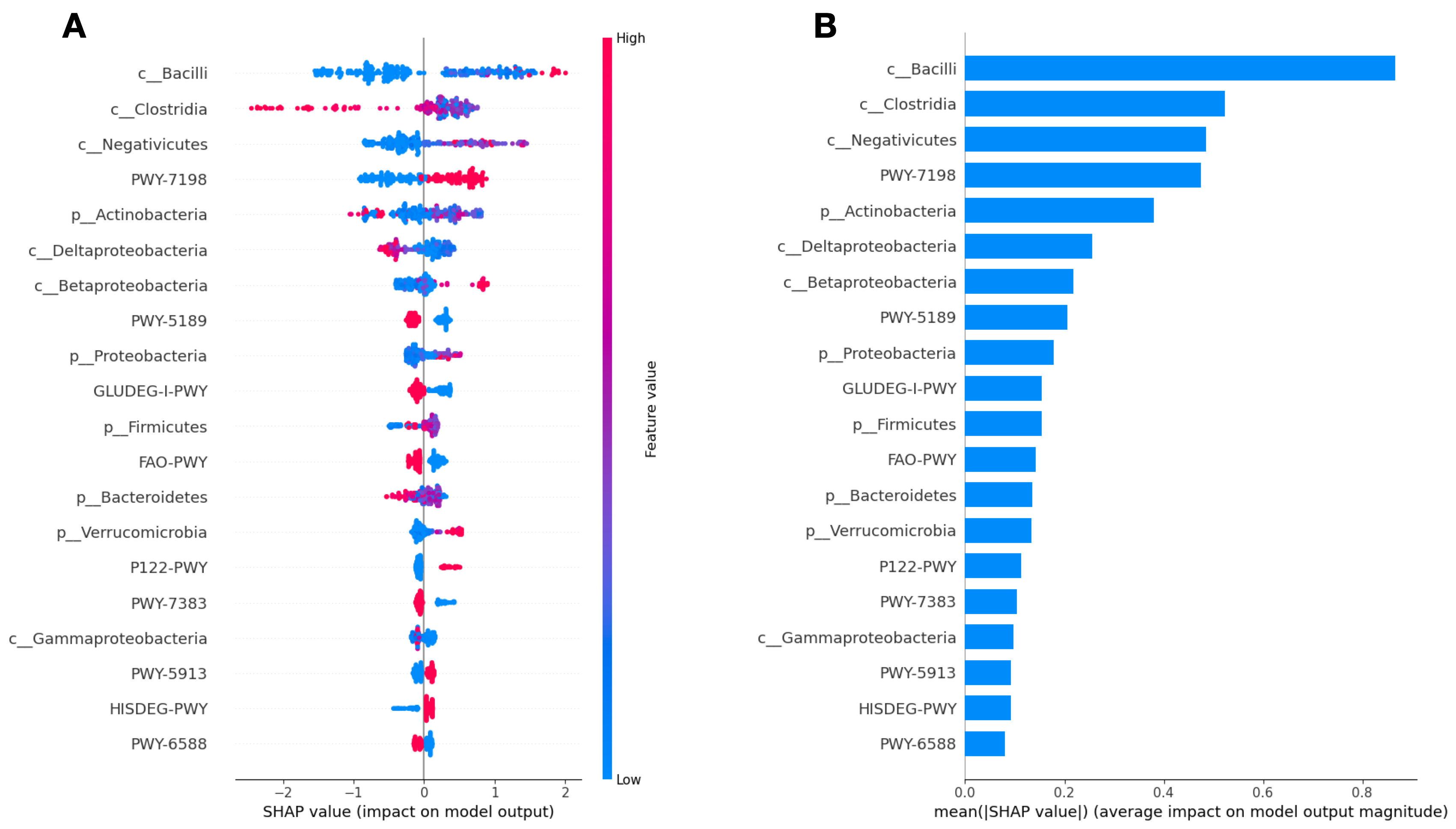


Figure 4 illustrates the results from the feature importance analysis. A) It shows the distribution of SHAP values, with each dot representing one sample. The colors red and blue indicate the inclination towards obese and non-obese statuses, respectively. B) The distribution displays the averaged SHAP values across all samples.

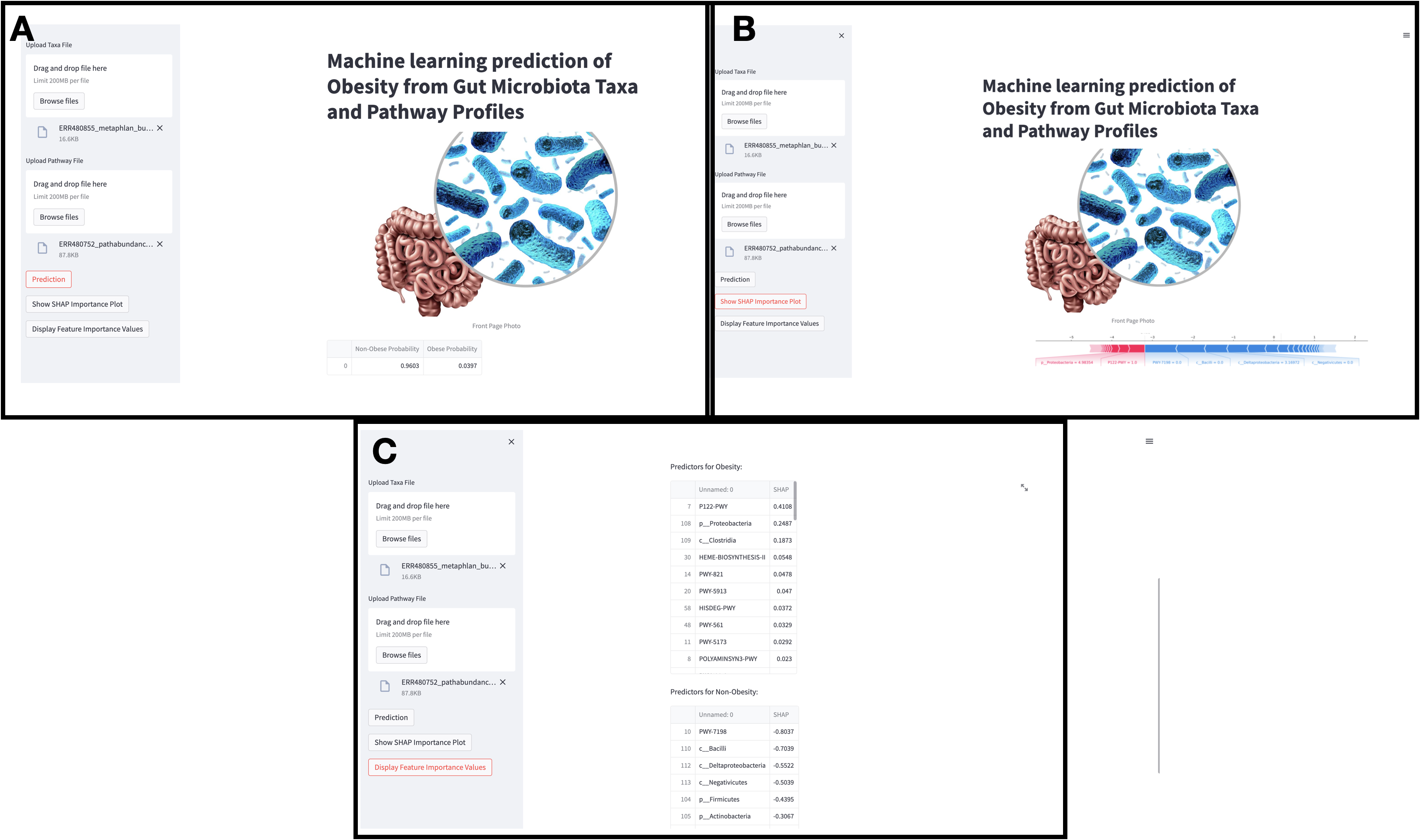


Figure 5 Showcases of the interactive features implemented for the user through the Streamlit application. A) The application displays the probabilistic outcome of the prediction based on the input files, indicating the probabilities for obese and non-obese status. B) The feature importance analysis for the input data is presented, with the red and blue colors representing the inclination towards obese and non-obese status, respectively. C) The table presents the most important features for the obese and non-obese predictions, inferred from the SHAP values.

**References**

1. Hu, J., et al., *Gut Microbiota Signature of Obese Adults Across Different Classifications.* Diabetes Metab Syndr Obes, 2022. **15**: p. 3933-3947.