Finding Biomarkers for Schizophrenia

Can Machine Learning algorithms identify schizophrenia-related biomarkers within metagenomic data derived from the human gut microbiome?

Timothy Bastow

Supervisors: Thomas Abeel, Eric van der Toorn, David Calderón Franco

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Timothy Bastow
Final project course: CSE3000 Research Project
Thesis committee: Thomas Abeel, Eric van der Toorn, David Calderon Franco

Abstract

There is mounting evidence indicating a relationship between the gut microbiome composition and the development of mental diseases but the mechanisms remain unclear. Shotgun sequenced data from 90 schizophrenic patients and 81 sex, age, weight, and location matched controls was used for three machine learning models: Logistic Regression, Random Forests, and XGBoost. The 20 most relevant species in the decision making of each classifier was retained and the overlap between models recorded. There is a total 19 overlapping species between the models' top 20 most relevant species, with 10 species overlapping on all three models. *Bifidobacterium bifidum*, *Akkermansia muciniphila*, *Eubacterium siraeum*, *Alistipes finegoldii*, *Intestinibacter bartlettii*, *Bifidobacterium pseudocatenulatum*, and *Streptococcus thermophilus* are of particular interest as they are reported as enriched in schizophrenia samples in existing literatures. *Phoceicola vulgatus* has been found to play a significant role in the classifiers decisions and is enriched in healthy samples in the literature. One species, *Ruthenibacterium lactatiformans*, and one co-abundant gene group, *Eubacterium sp. CAG:180*, consistently ranked as the most important features across all three classifiers, despite the absence of reporting in existing literature. This study could be expanded by using genus-level data. Further research should be done to validate the species mentioned above as potential biomarkers for schizophrenia.

1 Introduction

In recent years, increased research efforts have provided mounting evidence that mental diseases crucially affect the gut-brain axis (Appleton, 2018). A microbiome refers to a collection of microorganisms that reside in a given environment, in the case of this particular study the human gut. The gut microbiome plays a crucial role in human health and is influenced by several factors such as diet, lifestyle, and environmental exposures (Ahn and Hayes, 2021). The composition and function of the gut microbiome has gained significant scientific interest due to its potential implications for diagnosing, preventing, and treating various diseases (Shreiner et al., 2015). There are several methods to obtain microbiome information, also referred to as metagenomic information, from the human gut. One of the most common methods is to extract DNA information from stool samples and sequence it in order to determine the genetic information. While several sequencing methodologies are available, the current research focuses on samples that have undergone shotgun metagenomic sequencing, given its capacity for delivering accurate taxonomic data (Ranjan et al., 2016). Shotgun sequencing has several advantages over other methods such as 16S rRNA gene amplicon sequencing, namely enhanced detection of bacterial species, increased detection of diversity and increased prediction of genes. The longer read lengths in shotgun also improves the accuracy of species detection (Ranjan et al., 2016). These advantages are particularly important when leveraging taxonomic abundance data in machine learning models.

Unfortunately, shotgun metagenomic data from treatment-naive patients (patients without a history of prior treatment) remain scarce thereby limiting our understanding of the complex interactions between the gut microbiome and the brain when it comes to mental diseases (Zhu et al., 2020). One mental disorder that has received relatively limited attention in this field is Schizophrenia (Szeligowski et al., 2020). Affecting approximately 21 million individuals worldwide, schizophrenia remains a significant challenge in terms of early diagnosis and intervention (Charlson et al., 2018). Schizophrenia is a debilitating psychiatric condition, usually progressing gradually through various stages, characterized by hallucinations, delusions, and thought disorder and thereby distorting perception and hampering social interaction (Andreasen and Flaum, 1991). Diagnosis of schizophrenia predominantly relies on psychological assessments of patients, often leading to identification of the disorder only after it has reached an advanced stage (Lee et al., 2021). This highlights the need for novel diagnostic approaches to facilitate earlier detection and intervention. The extent to which the gut microbiome contributes to schizophrenia remains unclear and only a handful of studies have attempted to utilize machine learning to investigate schizophrenia-related biomarkers derived from shotgun sequenced data from the human gut microbiome (Wang et al., 2023).

In light of the potential role of the gut microbiome in mental disorders and the current limitations in early diagnosis of schizophrenia, there is a pressing need for novel methods to identify and validate biomarkers. The present research aims to combine the precision of shotgun sequencing with machine learning in order to identify schizophrenia-related biomarkers within metagenomic data derived from the human gut microbiome. To achieve this, three machine learning models will attempt to identify biomarkers with the same data. The structure of the paper is as follows: Section 2 details how the experiments were set up and the implementation process. Section 3 presents the results and then analyses and compares them to existing literature. Finally, Section 4 summarises the findings and provides suggestions on what future research could be done to find and verify biomarkers for schizophrenia.

2 Methodology

2.1 Language and Frameworks

The entirety of the data was procured through the CuratedMetagenomicData R package distributed through the Bioconductor ExperimentHub platform SOURCE, made available by Pasolli et al., (2017). The R programming framework 4.3.0 (R Core Team, 2021) was used to extract the relative taxonomic abundance of species data from the study performed by Zhu et al., (2020). The data was represented as a TreeSummarizedExperiment and further manipulations using R commands were done in order to extract the relative taxonomic abundance and convert into table format. The metadata was also extracted in a similar fashion.

The implementation of the machine learning process was done in Python 3.10 (Van Rossum and Drake, 2009), supplemented by a suite of additional libraries. These included the sci-kit learn 1.2.2 package (Pedregosa et al., 2011) as the main machine learning library, pandas 2.0.2 for manipulating the data tables (team, 2023), seaborn 0.12.2 for visualising data (Waskom, 2021), matplotlib 3.7.1 for generating graphs (Hunter, 2007), and numpy 1.24.3 for arithmetic operations (Harris et al., 2020).

2.2 Data Processing and Feature Extraction

The dataset created by Zhu et al., (2020) was used and made available through the CuratedMetagenomicData R package. This dataset encompassed gene families, marker abundance, marker presence, pathway abundance, pathway coverage, and relative abundance for each sample in the collection. Taxonomic abundance of bacteria, fungi, and archea in each sample were determined using MetaPhlAn3, while metabolic functional potential was calculated with HUMAnN3. All collected samples were sourced from various cities within the Shaaxi province, China, amassing a total of 171 samples, including 81 control samples. Each sample was accompanied by associated metadata features including gender, age, and Body Mass Index (BMI). Exploratory data anlaysis of the metadata revealed a balanced distribution across both the control and schizophrenia cohorts. In terms of gender, both cohorts exhibited an equal distribution between male and female participants. The age of the majority of the patients was bracketed within the 20-40 year range, and most recorded BMI values fell within the 18 to 24 range, considered as the healthy bracket. This data was used exclusively to verify whether the data was well balanced and did not have any inconsistencies and was not used for model training. This study focused exclusively on extracting relative species abundance information.

Regarding relative abundance data, each sample was represented by 481 features, each expressing the relative abundance of a particular species as a percentage value between 0 and 100. Features where 90% or more of the samples had a value of 0, and the remaining samples had a value of 0.01 or less, were discarded due to their negligible predictive utility, reducing the number of features from 476 to 179. The data was scaled using the Z-score normalization scaling method, implemented through Sci-kit's standardScaler function. The target labels, initially classified as ”control”, ”first-episode schizophrenia”, and ”repeated-episode schizophrenia”, were first simplified to either ”control” or ”schizophrenia” (merging first-episode and repeated-episode patients into one category) and then encoded into binary categories (0 for ”control” and 1 for ”schizophrenia”).

2.3 Model Training

This sub-section describes the methodology employed to collect results using the preprocessed data with the feature extracted, as detailed in the preceding segment. The specific parameters and random states implemented throughout the process are documented in Appendix A. The relative abundance of species table post-feature extraction was divided into a training and validation set following a 80/20 split with stratification. This ensured uniform representation of each class in both sets. Three pipelines were initialized with sklearn's pipeline package to implement three classification algorithms: Logistic Regression (LR), Random Forests (RF), and XGBoost (XGB). The sklearn's $StandardScaler$ was utilized in every pipeline for Z-score normalization scaling. A set of hyperparameters for each model (outlined in Appendix A) was then created for subsequent hyperparameters tuning processes.

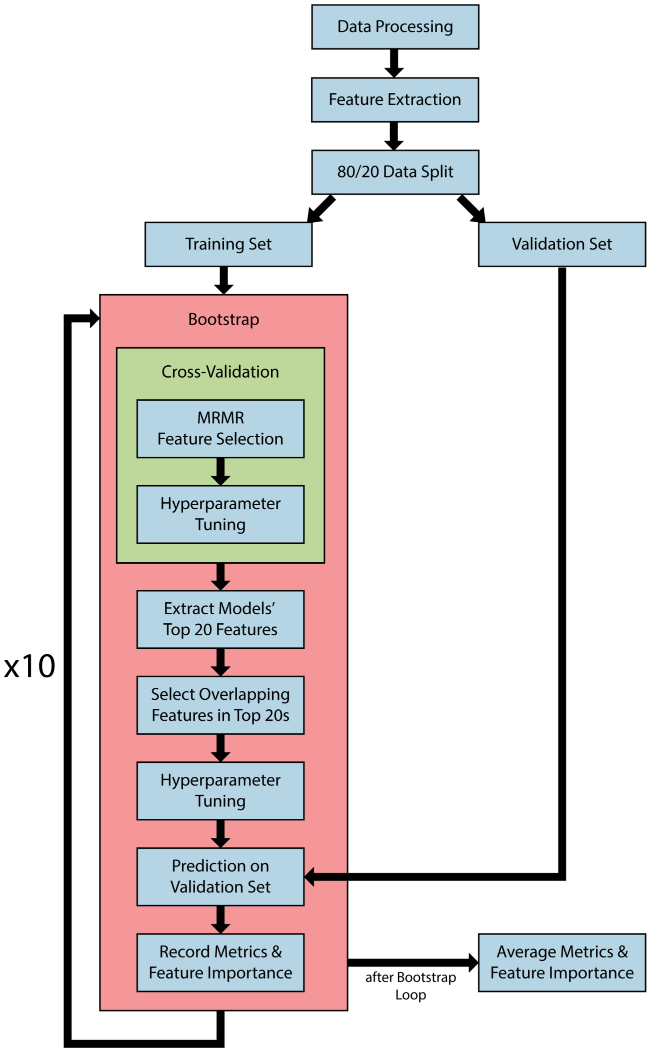

Figure 1: Diagram describing the overall process used in this study. The goal is to find the most important features across three classifiers. In order to avoid overfitting and to obtain accurate values for feature importance, the main process is bootstrapped (repeated 10 times with a different random state for each iteration). The feature importances are then averaged across all iterations.

The subsequent two paragraphs describe the bootstrapping technique used in the implementation. This technique was performed 10 times, each iteration using a unique random state for the cross-validation process to minimize overfitting and achieve a more precise representation of each model's feature importance. The classifier's random state remained constant throughout to ensure consistency across iterations.

The first step of the bootstrapping process consisted of performing stratified 5-fold cross-validation with MRMR feature selection and hyparameter-tuning. Stratification ensures preservation of the sample proportion for each class within the folds. For each fold's training segment, MRMR feature selection was independently applied to identify 50 relevant features. A random search was then conducted to fit 5 folds across 50 candidates, totalling 250 fits per model. The best parameters returned by the random search were applied to the model and the MRMR-selected 50 features were used for predictions on the validation fold. The accuracy, precision, and ROC-AUC scores as well as the confusion matrix of each fold were recorded. For features selected via MRMR, their importance was directly extracted for the RF and XGB models. In the case of LR, the absolute value of each feature's coefficient was recorded. Features not selected during the MRMR process were assigned an importance of zero for that specific fold. Upon completion of all folds, the metric scores and the feature importances were averaged, and the confusion matrices were cumulatively summed.

Following the completion of the cross-validation process, and the recording of the metric scores in global dictionaries, rankings for each classifier's most important features were compiled using the averaged feature importances calculated during the cross-validation process. Each ranking contained the 20 most important features for the given classifier. Each ranking was recorded to a global dictionary as to record the rankings of every bootstrap iteration. The classifiers were trained again on the updated training set containing only the overlapping features between the classifier's top 20 rankings — features appearing in at least two classifier rankings. Hyperparameter-tuning was performed once again for each classifier using the updated training set. Predictions were then made on the validation set using parameters returned by the random search for each model. The results of the predictions were recorded, marking the end of the bootstrap iteration. This entire sequence was repeated 10 times, each iteration using a different random state for the cross-validation process.

Upon completion of the bootstrapping loop, two separate rankings were generated for each classifier: one detailing the top 20 features across all cross-validation runs and another highlighting the 10 most important features across all validation set predictions. Similarly, two metric sets were developed for each model: one encompassing the average metrics across all cross-validation runs and another encompassing the average metrics from all validation set predictions. An Analysis of Variance (ANOVA) and a Tukey's Test were performed with the three classifiers to identify performance differences among the models. The ROC curve was plotted for each classifier using the predictions made across all bootstrap iterations in order to compare the models' ability to distinguish between classes.

2.4 Model Validation and Evaluation

Model validation and evaluation were done using an iterative cross-validation strategy, ensuring robust evaluation of predictive performance. A bootstrap process was implemented with different random states for each iteration, creating a degree of randomness that still permitted reproducibility. In each iteration, the three models - Logistic Regression, Random Forests, and XGBoost - underwent a stratified 5-fold cross-validation with feature selection on the training data. This stratified cross-validation strategy was used to optimise model hyperparameters and gather preliminary performance metrics. Additionally, confusion matrices were generated for each model. Post cross-validation, feature importance was computed for each model. The top 20 features were identified and selected for each model. Models were then retrained using these selected features, offering an opportunity for each model to learn from the most influential features. This feature selection process was consistently applied across all cross-validation iterations. The models were used to predict the outcomes on an validation set. These predictions permitted the evaluation of the models' ability to generalise to unseen data. ROC curves were plotted for each model using these predictions.

Finally, a statistical comparison between the models was conducted using an Analysis of Variance (ANOVA) and Tukey's post-hoc test. These tests offered a statistical perspective on the differences in the performance of the models. ROC curve plots were also generated for each model in order to further observe the difference in performance. The feature importance of each model were visualised using Venn diagrams, providing an overview of the comparative performance and feature utilisation of each model.

2.5 Responsible Research

Data Authenticity

The set used in this paper was initially created by Zhu et al. (2020) for their study titled *"Metagenome-wide association of gut microbiome features for schizophrenia"*. The study was supported by the Clinical Research Award of the First Affiliated Hospital of Xi'an Jiaotong University, Shenzhen Municipal Government of China, Innovation Team Project of Natural Science Fund of Shanxi Province, and Key Program of Natural Science Fund of Shanxi Province (Zhu et al., 2020). The raw data is publicly accessible and has been diposited in the China National Gene Bank (CNGB) and the European Nucleotide Archive (ENA). Furthermore, the processed data used in their study can be directly accessed from the published paper. The authors also provide an inventory of the software and tools employed throughout their research. Moreover, their publication underwent peer-review prior to being published. It is reasonable to assume that the legitimacy of the data due to the author's transparency regarding data availability, supporting institutions, and the peer-review process. However, the original publication offers limited insight into the specific procedures of data collection. The only detail provided is that shotgun sequencing was performed on fecal samples. Further data about the samples, such as demographic and clinical characteristics, can be found in the accompanying metadata.Additionally, the data was procured via the CuratedMetagenomicData package (Pasolli et al., 2017), designed to offer uniformly processed human microbiome data to users possessing minimal bioinformatic knowledge, implying that were was an external validation of the data. Even though it remains impossible to ascertain the integrity of the sample collection process, the aforementioned reasons permit a reasonable level of confidence in the data's authenticity.

Reproducibility of Methods

The present paper aims for transparency by using publicly accessible data and providing comprehensive documentation of the conducted experiments. Nevertheless, certain experiments that did not yield any contribution to the results featured in this paper have been excluded. This primarily includes discarded experiments that neither improved the models' performances nor yielded conclusive results. Ideally, these experiments would be documented as well, but constraints related to the paper's length made this impossible. Regardless, all experiments that did contribute in any way to the presented results are documented in this paper.

3 Results & Discussion

3.1 Feature Importance

Two rankings were created for each classifier, one listing the 20 most important features for each model across all cross-validation runs and another across predictions made on the validation set. The classifiers do not use the same set of features for training during the cross-validation runs and the predictions on the validation set and therefore the rankings are not directly comparable. During the cross-validation process, the set of features used for training is obtained by applying MRMR feature selection individually for each fold and feature importance is extracted from the selected features. For the predictions on the validation set, the feature set is obtained by merging the sets of overlapping features between the top 20 most important features of each classifier found during the cross-validation process. More details about the calculation of feature importance can be found in section 2.3.

Table 1 presents the combined rankings of each classifier across all cross-validation runs, indicating the most important features in the overall decision making process of each model. he Venn diagram in *Figure 2* illustrates the overlap between the top 20 overall rankings of each model, revealing the most relevant features for all three classifiers. During the bootstrapping process, the cross-validation ranking is used to create a new set of training features for the models. Consequently, the features listed in *Figure 2* are most relevant for the subsequent prediction on the validation set. There is a total of 19 overlapping features out of 31 distinct features present across all rankings. . A substantial overlap can be seen between the Random Forests and XGBoost rankings as they share 16 features out of their respective top 20. The Logistic Regression ranking diverges the most with at least 8 unique features not present in any other ranking. It also has the least individual overlap with other rankings.

For the Logistic Regression ranking across cross-validation runs, *Phocaeicola vulgatus* is the feature with the highest

Table 1: Table listing the most important species for the Logistic Regression (LR), Random Forests (RF), and XGBoost (XGB) classifiers across all cross-validation runs. For each model, the 20 most important species were retained (based on the methodology described in section 2.3). Overlapping species between rankings are highlighted in bold. The table is organised by overlapping group and alphabetical ordering of the species. LR uses coefficients to calculate a feature's importance whereas RF and XGB use feature importance. For each model, the scores of the 3 most important species are highlighted in bold.

Species	LR	RF	XGB
Alistipes finegoldii	0.044	0.032	0.017
Anaerostipes hadrus	0.049	0.033	0.02
Bifidobacterium bifidum	0.047	0.032	0.023
Eggerthella lenta	0.025	0.02	0.018
Eubacterium sp. CAG:180	**0.072**	**0.079**	**0.056**
Flavonifractor plautii	**0.06**	0.032	0.023
Intestinibacter bartlettii	0.051	0.021	0.019
Phocaeicola vulgatus	**0.094**	**0.036**	0.02
Ruthenibacterium lactatiformans	0.034	**0.052**	**0.04**
Streptococcus thermophilus	0.033	0.019	**0.031**
Akkermansia muciniphila	-	0.03	0.031
Clostridium innocuum	-	0.012	0.024
Eubacterium rectale	-	0.028	0.019
Eubacterium siraeum	-	0.019	0.021
Fusicatenibacter saccharivorans	-	0.025	0.014
Ruminococcus lactaris	-	0.015	0.024
Hungatella hathewayi	0.027	-	0.016
Roseburia sp. CAG:471	0.032	-	0.013
Bifidobacterium pseudocatenulatum	0.038	0.013	-
Actinomyces sp. ICM47	0.028	-	-
Anaerotruncus colihominis	0.026	-	-
Bacteroides salyersiae	0.021	-	-
Citrobacter portucalensis	0.026	-	-
Coprobacillus cateniformis	0.053	-	-
Firmicutes bacterium CAG: 110	0.054	-	-
Enterocloster asparagiformis	0.031	-	-
Blautia wexlerae	-	0.017	-
Ruminococcus bicirculans	-	0.009	-
Roseburia intestinalis	-	0.007	-
Citrobacter youngae	-	-	0.01
Parabacteroides goldsteinii	-	-	0.011
Mean Score	0.042	0.027	0.026

absolute coefficient by a clear margin. The co-abundant gene group *Eubacterium sp. CAG:180* and the species *Flavonifractor plautii* both have significantly higher coefficient values compared to the rest of the ranking. In the Random Forests ranking, *Eubacterium sp. CAG:180* is the feature with the highest importance. *Ruthenibacterium lactatiformans* also displays a considerably higher importance value relative to the rest of the ranking. Finally, in the XGBoost ranking, *Eubacterium sp. CAG:180* again leads as the feature with the highest importance score, although not as dominant as in the Random Forests ranking. *Ruthenibacterium lactatiformans* is also a high-ranking feature. Thus, *Eubacterium sp. CAG:180* and *Ruthenibacterium lactatiformans* appear to be crucial features during the cross-validation process.

Table 2 displays the combined rankings of each classifier across all predictions made on the validation set, revealing the most important features for each model after combining each model's most important features found during the cross-validation process. Since the set of training features consists of the overlapping features in the model's top 20 feature importance rankings, the pool of available features for prediction is significantly reduced. Therefore, only the top 10 most important features are retained for each model. *Figure 3* is a Venn diagram representing the overlap between each model's overall top 10 ranking, showing a total of 10 overlapping features out of 15 unique features present across all rankings.

Comparing *Table 1* with *Table 2*, we can observe that each classifier ranks feature importance for predictions on the validation set similarly to how they rank feature importance during the cross-validation process. The top three most important features for each classifier remain the same across the two tables. For the Random Forests and XGBoost classifiers, all the features listed in *Table 2* also appear in the top 10 from *Table 1*. For the Logistic Regression classifier, only seven features listed in *Table 2* are in the top 10 from *Table 1*. This inconsistency in the Logistic Regression's rankings could further suggest that using only features present in every classifier's top 20 most important feature rankings increases the performance of the Logistic Regression classifier.

Table 2: Table listing the most important species for the Logistic Regression (LR), Random Forests (RF), and XGBoost (XGB) classifiers across all predictions on the validation set. For each model, the 10 most important species were retained (based on the methodology described in section 2.3). Overlapping species between rankings are highlighted in bold. The table is organised by overlapping group and alphabetical ordering of the species. LR uses coefficients to calculate a feature's importance whereas RF and XGB use feature importance. For each model, the scores of the 3 most important species are highlighted in bold.

Species	LR	RF	XGB
Alistipes finegoldii	0.061	**0.066**	0.046
Eubacterium sp. CAG:180	**0.069**	**0.122**	**0.083**
Flavonifractor plautii	**0.07**	0.064	0.047
Phocaeicola vulgatus	**0.084**	0.056	0.042
Ruthenibacterium lactatiformans	0.051	**0.097**	**0.073**
Akkermansia muciniphila	-	0.057	0.061
Streptococcus thermophilus	0.045	-	**0.081**
Anaerostipes hadrus	0.057	0.057	-
Bifidobacterium bifidum	0.054	0.054	-
Eggerthella lenta	0.046	-	-
Intestinibacter bartlettii	0.053	-	-
Eubacterium rectale	-	0.046	-
Fusicatenibacter saccharivorans	-	0.052	-
Clostridium innocuum	-	-	0.044
Eubacterium siraeum	-	-	0.046
Ruminococcus lactaris	-	-	0.048

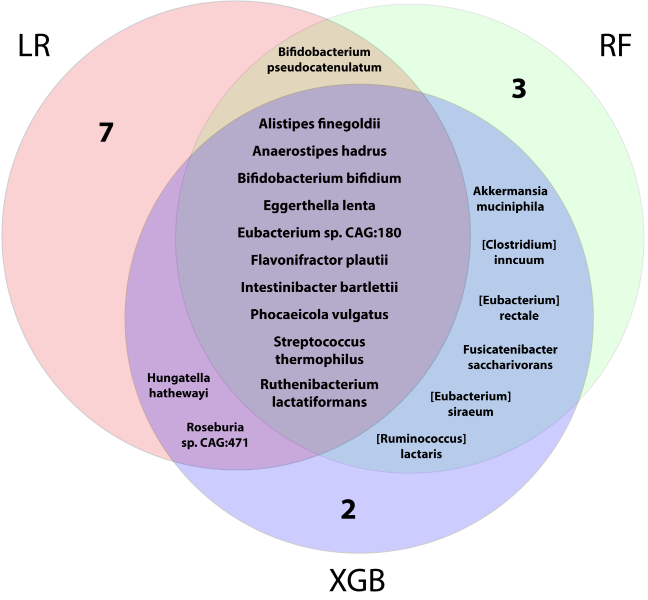

Figure 2: Venn Diagram showing the sets of overlapping features between the Logistic Regression (LR), Random Forests (RF), and XGBoost (XGB) classifiers' top 20 features across all cross-validation runs. Only the names of overlapping features are displayed, features that do not overlap between models are represented by their count.

3.2 Model Comparison

Statistical Tests

An Analysis of Variance (ANOVA) and a Tukey's Test were performed for the accuracy, precision, and ROC-AUC metrics with the three classifiers to identify performance differences among the models (ANOVA). For each of the three metrics, the ANOVA test has a p-value of $1.7e^{-3}$ or less, therefore the null hypothesis that all classifiers perform equally is rejected. A set of Tukey's tests was then used to enable pairwise comparisons between models. The statistical tests suggest that the Logistic Regression classifier outperforms both the Random Forests and XGBoost classifiers in terms of accuracy, precision, and ROC-AUC. However, there is no significant difference in the performance of the Random Forests and XGBoost classifiers according to these metrics. This could mean that either Random Forests or XGBoost is redundant, however, the objective is identify relevant features in the data. Thus, even if models perform similarly, it is beneficial to explore their decision making process. The ANOVA scores and the Tukey tests tables can be found in appendix B.

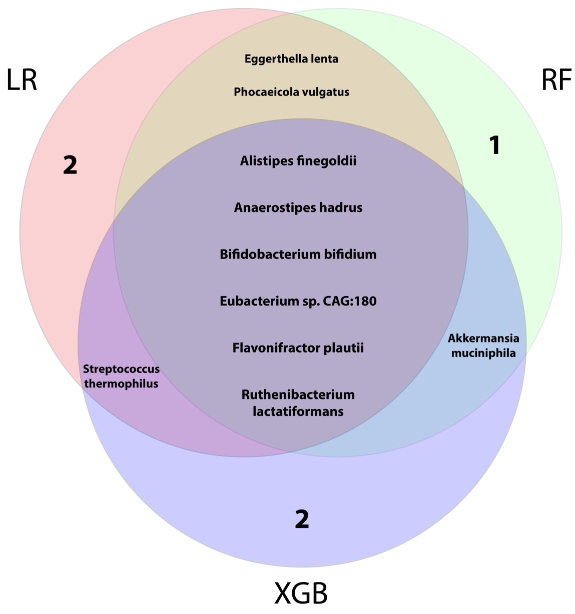

Figure 3: Venn Diagram showing the sets of overlapping features between the Logistic Regression (LR), Random Forests (RF), and XGBoost (XGB) classifiers' top 10 features across all predictions on the validation set. Only the names of overlapping features are displayed, features that do not overlap between models are represented by their count.

ROC Curves

The ROC curves for each of the classifiers - Logistic Regression, Random Forest, and XGBoost - are displayed in *Figure 4*. These curves serve as a graphical representation of sensitivity (true positive rate) and specificity (1 - false positive rate) at varying threshold levels. The curves in the Logistic Regression plot, though more varied in shape, generally approach the top left of the graph. This reflects a superior trade-off between sensitivity and specificity compared to the other classifiers. Despite the observable disparity among the Logistic Regression curves, suggesting a higher level of instability, this classifier's potential for high true positive rates at low false positive hints at a good overall performance. The overall performance of the models is discussed in the next sub-section.

On the other hand, the Random Forest and XGBoost classifiers demonstrate consistency across bootstrap iterations, as showed by their similar ROC curves. However, these curves approach the diagonal more compared to Logistic Regression, which suggests a less favorable trade-off between sensitivity and specificity. This observation implies that for these classifiers, an increase in sensitivity might coincide with a less acceptable increase in false positives, potentially making them less optimal for predicting schizophrenia where minimizing false positives is critical.

Metrics Discussion

This sub-section provides a comparison of the performance metrics of cross-validation runs and predictions made on the validation set for each model. *Table 3* presents the average accuracy, precision, and ROC-AUC scores with their re-

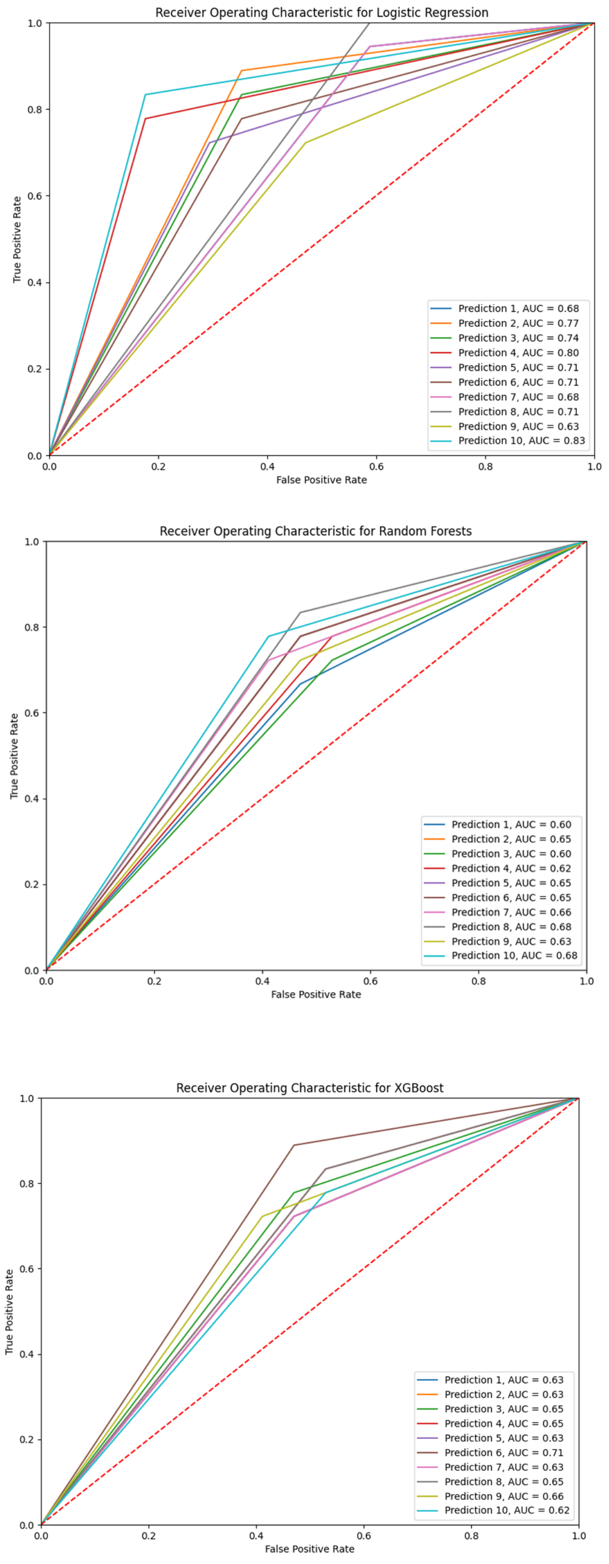

Figure 4: Receiver Operating Characteristic curve plots for every classifier across the ten predictions on the validation data set. The ten predictions are the result of the bootstrapping process, where the selected features for training the models vary from one iteration to the next.

Table 3: Averaged accuracy, precision, and ROC-AUC scores with their respective standard deviation values across ten 5-fold stratified cross-validation runs for each model, Logistic Regression (LR), Random Forests (RF), XGBoost (XGB).

	LR	RF	XGB
Accuracy	**0.54** ± 0.03	**0.67** ± 0.05	**0.67** ± 0.02
Precision	**0.58** ± 0.03	**0.69** ± 0.04	**0.71** ± 0.03
AUC-ROC	**0.57** ± 0.02	**0.74** ± 0.03	**0.74** ± 0.02

Table 4: Averaged accuracy, precision, and ROC-AUC scores with their respective standard deviation values obtained from ten predictions on the validation set for each model, Logistic Regression (LR), Random Forests (RF), XGBoost (XGB).

	LR	RF	XGB
Accuracy	**0.73** ± 0.06	**0.63** ± 0.04	**0.64** ± 0.03
Precision	**0.7** ± 0.07	**0.62** ± 0.03	**0.62** ± 0.03
AUC-ROC	**0.73** ± 0.06	**0.62** ± 0.04	**0.64** ± 0.03

spective standard deviations across ten 5-fold stratified cross-validation runs with feature selection and hyperparameter tuning. *Table 4* displays the same metric scores averaged from ten predictions made on the validation set. These predictions, for each model, are made utilizing the 20 most relevant features identified during the cross-validation process.

Table 4 reveals a significant increase in the Logistic Regression (LR) classifier's performance on the validation set compared to its cross-validation performance. Although the standard deviation of the mean metrics is higher in comparison to the standard deviation of other models or of the cross-validation process, the performance boost remains significant. On the other hand, both Random Forests (RF) and XGBoost (XGB) models show a slight performance dip on the validation set. Given that the training feature set is updated after the cross-validation process, this could imply that the LR classifier benefits from using features deemed important by the RF and XGB classifiers, while RF and XGB are negatively impacted by the omission of features not listed in the LR's top 20 features. The fact that the LR classifier's ranking differs on at least 8 features with the other classifiers' rankings, while RF and XGB's rankings only disagree on 4 features, could be an indication that the LR classifier benefits from utilizing features considered important by the other two models.

It is important to note that the LR classifier is a linear classifier whereas the Random Forests and XGBoost models are decision-tree based classifiers, which could explain why these two models share similar rankings. LR assumes linear relationship between the input variables and the log odds of the output variable. If the relationship is non-linear in the data, it could explain why Logistic Regression performs worse. However, it does outperform the other two classifiers when trained on overlapping features, suggesting that the linearity of the relationship between input and output variables may not be the predominant factor influencing the classifier's performance.

3.3 Comparison with Existing Literature

From the data presented in *Table 1* and *Table 2*, *Eubacterium sp. CAG:180* plays the most significant role. It consistently ranks first for both the Random Forests and XGBoost classifiers, and consistently within the top three features for the Logistic Regression classifier. Other notably significant species include *Phocaicola vulgatus* and *Ruthenibacterium lactatiformans*, both of which demonstrate exceptionally high importance or coefficient values for at least two classifiers. Other important species include species that are deemed important by all three classifiers across cross-validation runs and on predictions on the validation set. These include *Alistipes finegoldii*, *Akkermansia muciniphila*, *Streptococcus thermophilus*, *Anaerostipes hadrus*, and *Bifidobacterium bifidum*.

The study by Zhu et al. (2020), from which the data was obtained and which also implements machine learning methodologies, did not report *Ruthenibacterium lactatiformans* nor *Eubacterium sp. CAG:180* as significantly enriched in schizophrenia. However, the bacterial presence of *Phoceicola vulgatus* was significantly higher in healthy samples. This could indicate that the lack of *Phoceicola vulgatus* could be a potential indicator for schizophrenia (Zhu et al., 2020). *Bifidobacterium bifidum*, *Akkermansia muciniphila*, and *Eubacterium siraeum* were also reported as being enriched in schizophrenia. *Alistipes finegoldii* and *Intestinibacter bartlettii* were noted as being slightly enriched in healthy guts. All these species are recurrent across the Random Forests and XGBoost rankings. The Logistic Regression ranking does not include *Akkermansia muciniphila* and *Eubacterium siraeum*. Overall, the Random Forests and XGBoost rankings show greater correspondence with the findings presented by Zhu et al. (2020) compared to the Logistic Regression ranking. This correspondence is to be expected given that the study also used Random Forests as its main classifier.

The study by Castro-Nallar et al., (2015), which conducted a statistical shotgun metagenomic analysis on 16 individuals with schizophrenia and 16 controls, presented different findings. They noted a remarkably high abundance of *Streptococcus thermophilus* and *Bifidobacterium pseudocatenulatum* in schizophrenia samples, with *Streptococcus thermophilus* especially enriched. However, there is minimal overlap between this study's findings and those from Castro-Nallar et al., (2015) as the other eight species deemed relevant in their study are absent from any of the classifiers' rankings in the present study. This divergence could stem from numerous factors. Firstly, the studies originate from different locations, the data used in the present study originates from China whilst Castro-Nallar et al. (2015) conducted their study in the United States. The disparities in lifestyle, diet, or genetics could influence the composition of the gut microbiome. Secondly, the majority of the schizophrenia patients in the Castro-Nallar et al. (2015) study were smokers, which may potentially influence gut microbiome composition (Castro-Nallar et al., 2015). Finally, fundamentally different analyses methods were used between the two studies, one uses machine learning and the other statistical tools.

4 Conclusion and Future Work

The objective of this study was to validate potential biomarkers for schizophrenia, using data obtained from the human gut microbiome and processed via shotgun sequencing. Three machine learning classifiers, namely Logistic Regression, Random Forests, and XGBoost, were employed to analyze the relative species abundance using data sourced from Zhu et al. (2020). The importance of features was extracted from each classifier and subsequently combined with each other and compared to existing literature. In total, eight species that had high importance value across according to all three classifiers corresponded with findings documented in published literature. These species are: *Phoceicola vulgatus*, *Bifidobacterium bifidum*, *Akkermansia muciniphila*, *Eubacterium siraeum*, *Alistipes finegoldii*, *Intestinibacter bartlettii*, *Bifidobacterium pseudocatenulatum*, and *Streptococcus thermophilus*. One species, *Ruthenibacterium lactatiformans*, and one co-abundant gene group, *Eubacterium sp. CAG:180*, consistently ranked as the most important features across all three classifiers, despite the absence of reporting in existing literature.

During the cross-validation process, Logistic Regression notably underperformed relative to the other classifiers. Nevertheless, it excelled in its performance on the validation set in contrast to the other classifiers. Given that the feature set employed for training the classifiers is refined post-cross-validation to include only features considered significant by every classifier, it could be suggested that Logistic Regression benefits from the features selected based on their importance by the Random Forests and XGBoost classifiers. Despite this, the performance of Random Forests and XGBoost did not improve as a consequence of the updated feature set. Analyses via ANOVA and Tukey tests indicate a considerable similarity in the performance of both classifiers for given metrics.

Relatively few studies use the relative species abundance in the gut microbiome derived from shotgun metagenomic data, to analyze schizophrenia samples. To verify the relevance of the species identified in this study and those highlighted in Zhu et al. (2020) and Castor-Nallar et al. (2015), further research using the relative species abundance should be conducted. Moreover, further work should be undertaken to examine the importance of *Ruthenibacterium lactatiformans* and *Eubacterium sp. CAG:180*, given their absence in current literature. This study should be expanded to include more comprehensive statistical testing and analysis of results, in order to understand the classifiers' performance and to explain the correlation between species presence and schizophrenia. Furthermore, obtaining results using the genus level would offer a more robust basis for comparison with other studies, as the majority focus on the genus level in the gut microbiome.

References

Ahn, J., & Hayes, R. B. (2021). Environmental Influences on the Human Microbiome and Implications for Noncommunicable Disease. *Annual review of public health*, *42*, 277–292. https://doi.org/10.1146/annurev-publhealth-012420-105020

Andreasen, N. C., & Flaum, M. (1991). Schizophrenia: The Characteristic Symptoms. *Schizophrenia Bulletin*, *17*(1), 27–49. https://doi.org/10.1093/schbul/17.1.27

Appleton, J. (2018). The Gut-Brain Axis: Influence of Microbiota on Mood and Mental Health. *Integrative Medicine: A Clinician's Journal*, *17*(4), 28–32. Retrieved April 30, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6469458/

Castro-Nallar, E., Bendall, M. L., Pérez-Losada, M., Sabuncyan, S., Severance, E. G., Dickerson, F. B., Schroeder, J. R., Yolken, R. H., & Crandall, K. A. (2015). Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls [Publisher: PeerJ Inc.]. *PeerJ*, *3*, e1140. https://doi.org/10.7717/peerj.1140

Charlson, F. J., Ferrari, A. J., Santomauro, D. F., Diminic, S., Stockings, E., Scott, J. G., McGrath, J. J., & Whiteford, H. A. (2018). Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophrenia Bulletin*, *44*(6), 1195–1203. https://doi.org/10.1093/schbul/sby058

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy [Number: 7825 Publisher: Nature Publishing Group]. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment [Publisher: IEEE COMPUTER SOC]. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Lee, D., Seo, J., Jeong, H. c., Lee, H., & Lee, S. B. (2021). The Perspectives of Early Diagnosis of Schizophrenia Through the Detection of Epigenomics-Based Biomarkers in iPSC-Derived Neurons. *Frontiers in Molecular Neuroscience*, *14*. Retrieved April 29, 2023, from https://www.frontiersin.org/articles/10.3389/fnmol.2021.756613

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., & Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub [Number: 11 Publisher: Nature Publishing Group]. *Nature Methods*, *14*(11), 1023–1024. https://doi.org/10.1038/nmeth.4468

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, *469*(4), 967–977. https://doi.org/10.1016/j.bbrc.2015.12.083

Shreiner, A. B., Kao, J. Y., & Young, V. B. (2015). The gut microbiome in health and in disease. *Current opinion in gastroenterology*, *31*(1), 69–75. https://doi.org/10.1097/MOG.0000000000000139

Szeligowski, T., Yun, A. L., Lennox, B. R., & Burnet, P. W. J. (2020). The Gut Microbiome and Schizophrenia: The Current State of the Field and Clinical Applications. *Frontiers in Psychiatry*, *11*. Retrieved May 2, 2023, from https://www.frontiersin.org/articles/10.3389/fpsyt.2020.00156

team, T. p. d. (2023). Pandas-dev/pandas: Pandas. https://doi.org/10.5281/zenodo.7979740

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Wang, D., Russel, W. A., Sun, Y., Belanger, K. D., & Ay, A. (2023). Machine learning and network analysis of the gut microbiome from patients with schizophrenia and non-psychiatric subject controls reveal behavioral risk factors and bacterial interactions. *Schizophrenia Research*, *251*, 49–58. https://doi.org/10.1016/j.schres.2022.12.015

Waskom, M. L. (2021). Seaborn: Statistical data visualization [Publisher: The Open Journal]. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Zhu, F., Ju, Y., Wang, W., Wang, Q., Guo, R., Ma, Q., Sun, Q., Fan, Y., Xie, Y., Yang, Z., Jie, Z., Zhao, B., Xiao, L., Yang, L., Zhang, T., Feng, J., Guo, L., He, X., Chen, Y., ... Ma, X. (2020). Metagenome-wide association of gut microbiome features for schizophrenia. *Nature Communications*, *11*, 1612. https://doi.org/10.1038/s41467-020-15457-9

A Parameter Settings

A.1 Random States Set

$[42, 1, 1006, 1998, 106, 111117, 1902, 2903, 209, 360]$

A.2 Logistic Regression Hyperparameters

- **C**: $np.logspace(-3, 3.7)$
- **penaly**: $["l1", "l2"]$
- **solver**: $["newton-cg", "lbfgs", "liblinear", "sag"]$
- **l1_ratio**: $np.linspace(0, 1, 5)$

A.3 Random Forests Hyperparameters

- **n_estimators**: $[50, 100, 200, 500, 1000]$
- **max_depth**: $[None, 10, 20, 30, 50]$
- **min_samples_split**: $[2, 5, 10, 12]$
- **min_samples_leaf**: $[1, 2, 4, 10]$
- **max_features**: $["auto", "sqrt", "log2"]$

A.4 XGBoost Hyperparameters

- **learning_rate**': $[0.01, 0.1, 0.2, 0.3]$
- **n_estimators**': $[50, 100, 200, 500]$
- **max_depth**': $[None, 3, 10]$
- **min_child_weight**': $[0.5, 0.7, 1.0]$
- **gamma**': $[0, 0.1, 0.2]$
- **subsample**': $[0.5, 0.7, 1.0]$
- **colsample_bytree**': $[0.5, 0.7, 1.0]$

B ANOVA and Tukey's Test

B.1 ANOVA and Tukey's HSD Test for accuracy

ANOVA test for accuracy: $F = 12.45$, $p = 0.00015$

Tukey's HSD test for accuracy

G1	G2	MeanDiff	p-adj	lower	upper
LR	RF	-0.0886	0.0004	-0.1378	-0.0393
LR	XGB	-0.0829	0.0008	-0.1321	-0.0336
RF	XGB	0.0057	0.9555	-0.0436	0.055

B.2 ANOVA and Tukey's HSD Test for precision

ANOVA test for precision: $F = 8.2$, $p = 0.0016$

Tukey's HSD test for precision

G1	G2	MeanDiff	p-adj	lower	upper
LR	RF	-0.0759	0.0052	-0.1306	-0.0213
LR	XGB	-0.0785	0.0038	-0.1331	-0.0239
RF	XGB	0.0026	0.9925	-0.0572	0.052

B.3 ANOVA and Tukey's HSD Test for ROC-AUC

ANOVA test for ROC-AUC: $F = 12.1$, $p = 0.00018$

Tukey's HSD test for ROC-AUC

G1	G2	MeanDiff	p-adj	lower	upper
LR	RF	-0.0879	0.0005	-0.1377	-0.0381
LR	XGB	-0.0832	0.0009	-0.133	-0.0333
RF	XGB	0.0047	0.9699	-0.0451	0.0546