

FINDING NOVEL MARKERS FOR NEURODEVELOPMENTAL DISEASES BY
MACHINE LEARNING-BASED CLASSIFICATION

by

Atacan Deniz Öncü

B.S., Computer Engineering, Middle East Technical University Northern Cyprus
Campus, 2021

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computational Science and Engineering
Boğaziçi University
2025

ACKNOWLEDGEMENTS

There are so many people to thank for. First of all, I would like to thank my advisors Kutlu Ülgen and Arzucan Özgür for supporting me throughout my education. Without their help, this thesis would be incomplete.

Second gratitude goes to my family, for their never ending support despite our ups and downs.

Third gratitude goes to the friends I have made along the way whom I consider my chosen family.

Last one goes to myself for not giving up despite all the difficulties. I am completely a different person from when I have started this journey and honestly I love this new me.

ABSTRACT

FINDING NOVEL MARKERS FOR NEURODEVELOPMENTAL DISEASES BY MACHINE LEARNING-BASED CLASSIFICATION

Major Depressive Disorder (MDD) is one of the leading neurodevelopmental disorders worldwide and its heterogeneous and complex nature remains a significant challenge for scientists to fully understand and unravel. In recent years, the importance of gut microbiome through the human gut-brain axis has gathered the attention of scientists to analyze and model the bacterial components of the human gut and its effects on the human brain, especially in the case of MDD. In this study, we analyzed the American Gut Project (AGP) dataset using fecal samples of 361 controls and 23 MDD patients. After retrieving the Qiita bioinformatics analysis, the cohort was analyzed for its characteristics and for alpha and beta diversity indexes which did not reveal any statistical significance except for the age of patients. Various differential abundance analysis (DAA) methods were conducted to find potential biomarkers and these results were fed into our machine learning models as an alternative DAA-filtered dataset compared to the raw dataset to find important features. Our best two models, which were Random Forest and XGBoost, at their intersection, have found that despite some inconsistencies in the literature, species *Bifidobacterium adolescentis* and genera *Odoribacter*, *Ruminococcus*, and *Adlercreutzia* could be potential biomarkers for MDD. Our models found that species *B. adolescentis* decreased in MDD patients whereas the rest increased in MDD patients. Despite supporting evidence for these potential biomarkers, factors such as modeling and filtering choices, as well as external influences like sex, stress and diet should also be taken into account when analyzing gut microbiome datasets.

ÖZET

MAKİNE ÖĞRENMESİ BAZLI SINIFLANDIRMA KULLANARAK NÖROGELİŞİMSEL HASTALIKLAR İÇİN AYIRT EDİCİ ÖZELLİKLERİN BULUNMASI

Majör Depresyon Bozukluğu (MDB) dünyada öne gelen nörogelişimsel bozukluklardan birisidir ve kompleks doğası onu günümüz bilim insanları tarafından hala daha tam anlamıyla çözülemeyen bir probleme dönüştürmüştür. Geçtiğimiz yıllarda MDB gibi bozuklukların beyin-bağırsak ilişkisi üzerinden bağırsak mikrobiyomunun önemi analiz ve modelleme yöntemleriyle incelenerek bilim insanlarının dikkatini çekmeye başlamıştır. Bu çalışmada Amerikan Bağırsak Projesi'ne ait 361 Kontrol ve 23 MDB hastanın dışkı örnekleri kullanılmıştır. Qiita üzerinden elde edilen biyoinformatik analizinden sonra veri setinin özellikleri incelenmiştir ve alfa ve beta çeşitlilik indekslerine bakılarak hastaların yaş bilgisi haricinde herhangi bir istatistiksel anlamlılık bulunamamıştır. Çeşitli Diferansiyel Bolluk Analizi (DBA) yöntemleri kullanılarak potansiyel biyolojik belirteçler bulunmaya çalışılmıştır ve buradan elde edilen sonuçlar makine öğrenme modellerine yüklenerek alternatif bir DBA filteresi uygulanmış veri seti kullanılarak orjinal veri setiyle kıyaslanıp önemli belirteçler bulunmuştur. Makine modellerinin arasında en iyi modeller Random Forest ve XGBoost iken, kesişimlerine bakıldığında literatürdeki bazı çelişkilere rağmen bakterilerden *Bifidobacterium adolescentis* türü ve *Odoribacter*, *Ruminococcus*, ve *Adlercreutzia* cinslerinin MDB için potansiyel biyolojik belirteçler olabileceği bulunmuştur. Modellerin sonucuyla *B. adolescentis* türünün MDB hastalarında azaldığını, geriye kalanların ise arttığı bulunmuştur. Bu biyolojik belirteçler için destekleyici bulgular olmasına rağmen, bağırsak mikrobiyom verileri incelenirken alternatif modelleme ve filtreleme seçenekleri ve dış etkenlerden cinsiyet, stres ve beslenme tipi gibi faktörleri de ele alınmalıdır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
ÖZET	v
LIST OF FIGURES	viii
LIST OF TABLES	x
LIST OF ACRONYMS/ABBREVIATIONS	xii
1. INTRODUCTION	1
2. LITERATURE REVIEW	4
2.1. Major Depressive Disorder	4
2.2. Biomarkers for Major Depressive Disorder	5
2.3. Influence of Diet	7
2.4. Understanding Microbiome Data and Machine Learning Models	9
2.5. Machine Learning Performance Evaluation Metrics	15
3. MATERIAL & METHODS	17
3.1. Data	17
3.2. Diversity Analysis	18
3.3. Differential Abundance Analysis	20
3.4. Machine Learning Modeling	21
4. RESULTS	24
4.1. Diversity Analysis on Gut Microbial of MDD Individuals	24
4.1.1. Alpha and Beta Diversity	24
4.1.2. Mean Relative Abundance	25
4.2. Statistical Analysis on MDD Data	25
4.2.1. DESeq2	26
4.2.2. LEfSe	27
4.2.3. Wilcoxon (rarefied)	28
4.2.4. Welch's t.test (rarefied)	29
4.2.5. ALDEx2	29

4.2.6.	MaAsLin2	30
4.2.7.	Summary of DAA Results	30
4.3.	Machine Learning Models on MDD Data	33
4.3.1.	Pilot Experiments	33
4.3.2.	Comparison of All Models	37
4.3.3.	Comparison of Best Two Models	39
4.3.4.	Average of Most Important Features of Random Forest	40
4.3.5.	Average of Most Important Features of XGBoost	40
4.3.6.	SHAP Values of Raw Random Forest Model	41
4.3.7.	SHAP Values of DAA Filtered Random Forest Model	42
4.3.8.	Potential Biomarkers Retrieved from Random Forest Models	43
4.3.9.	SHAP Values of Raw XGBoost Model	45
4.3.10.	SHAP Values of DAA Filtered XGBoost Model	46
4.3.11.	Potential Biomarkers Retrieved from XGBoost Models	47
4.3.12.	Overall Potential Biomarkers	47
5.	DISCUSSION	48
6.	CONCLUSION	56
7.	RECOMMENDATIONS	58
	REFERENCES	59

LIST OF FIGURES

Figure 4.1.	Alpha diversity representation of the groups using (a) Shannon, (b) Simpson, and (c) Observed Species indexes.	24
Figure 4.2.	Beta diversity representation of the groups using (a) NMDS and (b) PCoA ordinations.	25
Figure 4.3.	Mean relative abundances of the cohort on Genus level.	26
Figure 4.4.	DESeq2 results.	27
Figure 4.5.	(a) MA and (b) Volcano plots derived from DESeq2 results.	27
Figure 4.6.	Heatmap from DESeq2 results.	28
Figure 4.7.	LEfSe results.	28
Figure 4.8.	Wilcoxon (rarefied) results.	29
Figure 4.9.	Welch's t.test (rarefied) results.	30
Figure 4.10.	Areas under (a) Receiver operating characteristic (ROC-AUC) and (b) Precision-Recall (PR-AUC) curves for models using raw datasets.	38
Figure 4.11.	Areas under (a) Receiver operating characteristic (ROC-AUC) and (b) Precision-Recall (PR-AUC) curves for models using DAA filtered datasets.	39

Figure 4.12. Most important features of Random Forest model ranked by their average gini importance values.	41
Figure 4.13. Most important features of XGBoost model ranked by their average F-score values.	42
Figure 4.14. SHAP values of raw Random Forest model.	43
Figure 4.15. SHAP values of DAA filtered Random Forest model.	44
Figure 4.16. SHAP values of raw XGBoost model.	45
Figure 4.17. SHAP values of DAA filtered XGBoost model.	46

LIST OF TABLES

Table 2.1.	Confusion Matrix.	15
Table 3.1.	General sample characteristics of the cohort.	18
Table 3.2.	Differential Abundance Analysis methods employed in this thesis. .	20
Table 3.3.	Hyperparameters that were tuned for each model.	22
Table 4.1.	The potential biomarkers identified by applying DAA methods. . .	31
Table 4.2.	Biomarkers identified by all types of differential abundance analysis methods, the arrows represent the decrease/increase of the abun- dance of biomarker in MDD patients.	32
Table 4.3.	Results of all raw ML models during pilot experiments.	33
Table 4.4.	Results of all undersampled ML models during pilot experiments. .	34
Table 4.5.	Results of all oversampled ML models during pilot experiments. . .	35
Table 4.6.	Results of all ratioed ML models during pilot experiments.	35
Table 4.7.	Comparison of all models using undersampling methods NearMiss, RandomUnderSampler (random_state="default"), and RandomUnder- Sampler (random_state=100) respectively.	36
Table 4.8.	Results of ML models with all features.	37

Table 4.9. Results of ML models with DAA filtered features. 38

Table 4.10. Tuned hyperparameter values retrieved from both raw and DAA
filtered models. 40



LIST OF ACRONYMS/ABBREVIATIONS

ADHD	Attention Deficit Hyperactivity Disorder
AGP	American Gut Project
ASD	Autism Spectrum Disorder
BH	Benjamini Hochberg
CNS	Central Nervous System
DAA	Differential Abundance Analysis
DT	Decision Tree
EMP	Earth Microbiome Project
FMT	Fecal Microbiota Transplant
HFP	Human Food Project
HPA	Hypothalamus Pituitary Adrenal
IBD	Inflammatory Bowel Disease
IBS	Irritable Bowel Syndrome
LOOCV	Leave One Out Cross Validation
LR	Logistic Regression
MDD	Major Depressive Disorder
ML	Machine Learning
NMDS	Non-metric Multi Dimensional Scaling
PCoA	Principal Coordinates Analysis
PTSD	Post Traumatic Stress Disorder
RF	Random Forest
XGB	XGBoost

1. INTRODUCTION

Major Depressive Disorder (MDD) is one of the most common neurodevelopmental disorders, seen among 11% of Americans [1] and nearly more than 350 million people globally [2, 3]. MDD's symptoms are characterized by continuous depressed mood, fatigue, changes in sleep patterns and appetite, suicidal ideation, anhedonia and loss of energy [4–8]. The pathophysiology of this disorder is still unclear due to its heterogeneous nature and research in the past used to focus on aspects such as neurotransmitters, stress hormones, neurotrophic factors, and circadian rhythms to characterize MDD [2], [5, 6]. Recent scientific developments highlight the importance of the relationship between the brain and the gut, and how it can show effects upon compositional dysregulation, since putting the emphasis solemnly on genetic heterogeneity, biochemical or neurophysiological changes, or psychosocial variables has not been enough to unravel the pathogenesis of MDD [8, 9].

Microbiota is the collection of microorganisms such as bacteria, viruses, fungi, protozoa, and archaea that reside in the human gut and the genetic materials of such microorganisms make up the microbial composition known as the microbiome [10]. They can influence digestion, vitamin production, permeability of the intestinal barrier and immunity. Many factors could alter the composition of the gut microbiota such as genes, age, sex, diet, life experiences, the environment and stress [4]. These microorganisms make up a complex and dynamic ecosystem functioning as a bidirectional signaling network that interacts with the brain known as the “gut-brain axis” [11]. This bidirectional system has the role of communication between the gut microbiome and the central nervous system (CNS), and microorganisms in the gut can influence the CNS by the microbial byproducts that enter the bloodstream [12]. The gut-brain axis is known to be involved in hormonal, immunological and neural homeostasis [2], [4]. There could be several reasons for compositional dysregulations in the microbiota such as the pile-up of neurotransmitters, alterations in enzyme activity, compromise of the immune system, increase in metabolites of tryptophan, stimulation of the vagal nerve,

alterations in the hypothalamus-pituitary-adrenal (HPA) axis, and increase in the mucosal layer which could result in “leaky gut”, thus causing gut related harmful issues and rendering the blood-brain barrier permeable, potentially influencing the development and functionality of the brain [13,14].

The possible effects of pathogenic and beneficial bacterial species on host pathophysiology have been the interest of many animal studies. In one study, germ-free mice were observed after being colonized with *Escherichia coli* and this resulted in altered behaviors upon exposure to stress compared to pathogen-free mice. This behavior was profiled as a proinflammatory response, and a vaccine that included a strain of *Bifidobacterium* was used to alleviate the behavioral and immunological changes [15]. Such results using single bacterial strains changed the direction of experiments towards using gut microbial contents. Over the years, several research have been conducted to find correlations between the gut microbiota and MDD through the method of fecal microbiota transplantation (FMT) on animals. Numerous results have shown that transplanting germ-free mice with the fecal matter of a depressed patient altered their behaviors and induced symptoms of MDD, indicating a link between the gut microbiome and MDD [2], [6], [16–18]. Despite the promising results of FMT on animals, it is important to remember that human and mice microbiota compositions are different from each other, and external factors such as diet, genetic factors, drugs etc should also be taken into account when examining human microbiomes [16].

Even though studies show conflicting results due to the variability of the methods, human studies demonstrated potential biomarkers that could differentiate a healthy microbial sample from a depressed one. Most common results show that people with MDD contained decreased levels of genera *Bifidobacterium*, *Faecalibacterium*, *Coprococcus*, *Dialister*, *Lactobacillus*, and *Bacterioidetes* and increased levels of *Blautia*, *Enterobacteriaceae*, *Actinobacteria*, *Desulfovibrio*, and *Alistipes* in their gut microbiota [1,2], [6–9], [17], [19].

It is important to mention that some studies show inconsistencies depending on various host factors. For instance, a study focusing on the effects of sex resulted in increased levels of phylum *Actinobacteria* in female patients and phylum *Bacteroidia* in male depressed patients. There was also a disparity in the relative abundance of *Firmicutes* between sexes, which was thought to be caused by the sex factor [1], [8].

Diet is another external factor that is thought to be the cause of such discrepancies in gut compositions. Dietary differences between regions can cause differences in the composition and diversity of the microbiome. Western diet, for instance, is known to have high amounts of fat and can elevate the abundance of gram-negative bacteria in the microbiota which could be a potential indicator of MDD [2].

The shift from the neurochemical and neurological aspect of depression to the intestinal microbiota seems promising and can be used to develop more effective methods for the diagnosis and treatment of MDD and possibly other neurodevelopmental disorders. Up-to-date research methods provide us with numerous analysis and machine learning methods that can bring scientific research up to a new level, providing powerful algorithms and tools for effective analysis, clarification and prediction of the links between gut microbiota and depression [11].

In this thesis, multiple statistical analyses were conducted on MDD data and then various machine learning models were employed to investigate the link between MDD and inherent differences in the gut microbial composition, compared to that in normal controls and then these ML models were used for classifying the state of patients. This thesis may offer significant insights into the comprehension of neurodevelopmental psychiatric disorders like MDD and eventually may lead to improved treatments because unlike many research in the microbiome field, instead of focusing on a single statistical analysis method or machine learning modeling, multiple approaches were tested and compared with each other to validate the results. This approach provides a more comprehensive understanding of the gut microbiome and different results from different methods might show a new direction for the microbiome research.

2. LITERATURE REVIEW

2.1. Major Depressive Disorder

MDD is a complex neurodevelopmental disorder caused by a combined effect of genetic, environmental, psychological and biological factors. Both aetiological (genetic and environmental) factors and mechanisms at a molecular and cellular level (alterations in brain structure and function, inflammation, the gut–brain axis, and the hypothalamus–pituitary–adrenal (HPA) axis) are involved in the onset and continuity of MDD [20].

In modern psychiatry, the common diagnostic methods for MDD are patient observation and interviews with the patients and their immediate family members regarding their medical history. Pharmacotherapy is the leading therapeutic method for depressive symptoms where patients are treated with antidepressants. This not only becomes a financial burden for the patients due to the cost of healthcare but also patients can develop resistance against the treatment (an estimate of up to 35% of people) and repeated exposure to the same medication can cause reduced efficiency. Some patients do not want to continue using drugs due to their side effects. These issues require scientists to explore novel therapies for a better quality of life for depressed patients [21]. Unfortunately, the pathophysiology of MDD is still unclear and there is no biochemical marker that is being used for diagnosis. Currently, there are four biological systems seen as potential biochemical markers for MDD which are; the dopamine system, stress response by the hypothalamic-pituitary-adrenal (HPA) axis and autonomic nervous system (ANS), chronic inflammatory markers, and dysfunction of brain-derived neurotrophic factor (BDNF) [5].

Growing evidence has suggested that altered gut microbiome may also be involved in the pathophysiology of MDD [5]. The human gut is typically dominated by four bacterial phyla which are *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria*,

and a healthy person is expected to have a standard level of diversity between these phyla. However, in cases of MDD, alterations in certain phyla have been observed, indicating a potential causal relationship between MDD and gut microbiome composition [22]. Apart from MDD, changes in the gut microbiota have also been linked to several diseases such as inflammatory bowel disease, type 2 diabetes, colorectal cancer, and schizophrenia through the gut-brain axis. As a result, microbiome science has started to have important implications for drug development and personalized medicine regarding several health problems [23].

2.2. Biomarkers for Major Depressive Disorder

Biomarkers (“biological markers”) are biological characteristics or indicators that can be objectively measured and evaluated for disease diagnosis and pathogenesis [24]. There are hundreds of studies and reviews of potential biomarkers regarding MDD but they show different results due to differences in the methodologies of the reviewed studies and their cohorts. One review by Knudsen et al. (2011), which included 17 studies, frequently observed that the relative abundances of *Eggerthella*, *Atopobium*, and *Bifidobacterium* increased and the relative abundance of *Faecalibacterium* decreased in MDD samples [15].

Another review by Cheung et al. (2019), which included 6 studies, found 21 genera that distinguished the MDD group from the Control group: abundances of *Anaerostipes*, *Blautia*, *Clostridium*, *Klebsiella*, *Lachnospiraceae incertae sedis*, *Parabacteroides*, *Parasuterella*, *Phascolarctobacterium*, and *Streptococcus* were higher in MDD samples, abundances of *Bifidobacterium*, *Dialister*, *Escherichia/Shigella*, *Faecalibacterium*, and *Ruminococcus* were lower in MDD samples, and there were conflicting results for *Alistipes*, *Bacteroides*, *Megamonas*, *Oscillibacter*, *Prevotella*, and *Roseburia* in both directions [25].

A review of 9 studies by Barandouzi et al. (2020) found a high abundance of genera *Oscillibacter*, *Blautia*, *Holdemania*, *Clostridium XIX*, *Anaerostipes*, *Lachnospiraceae*

incertae sedis, *Anaerofilum*, *Streptococcus*, *Gelria*, *Turicibacter*, *Parabacteroides*, *Eggerthella*, *Klebsiella*, *Streptococcus*, *Paraprevotella*, *Veillonella*, *Clostridium IV*, *Eubacterium*, *Erysipelotrichaceae incertae sedis*, *Parvimonas*, *Desulfovibrio*, *Parasutterella*, *Actinomyces*, *Asaccharobacter*, *Atopobium*, and *Olsenella* and a low abundance of genera *Coprococcus*, *Lactobacillus*, *Escherichia/Shigella*, *Clostridium XI*_{Va}, *Dialister*, *Howardella*, *Pyramidobacter*, and *Sutterella* in MDD patients. For the abundance of *Bifidobacterium*, *Roseburia*, *Lachnospiraceae incertae sedis*, *Megamonas*, *Clostridium XI*, *Bacteroides*, *Prevotella*, *Alistipes*, *Phascolarctobacterium*, *Faecalibacterium*, and *Ruminococcus*, however, there were conflicting results across studies [2].

A more recent review of 44 case-control studies by Gao et al. (2023) found that the most consistent changes in the gut microbiota were an increase of genera *Eggerthella*, *Enterococcus*, *Escherichia*, *Flavonifractor*, *Holdemania*, *Lachnoclostridium*, *Paraprevotella*, *Rothia*, and *Streptococcus* and a decrease of genera *Butyrivibrio*, *Coprococcus*, *Faecalibacterium*, *Eubacterium ventriosum* group, *Fusicatenibacter*, *Romboutsia*, and *Subdoligranulum* in patients with MDD relative to their Control counterparts [14].

Studies differ in results as they vary in cohort and metadata characteristics. A study by Chen et al. (2022) found out that the levels of genus *Bifidobacterium* were decreased in depressed patients across 24 cohorts with a total of 18340 participants, which showed consistent results with their previous studies. They also found that genera *Actinobacteria* and *Ruminococcus* were decreased whereas genus *Streptococcus* increased in MDD patients but the authors suggested further investigation into these findings [8].

Another study by Chung et al. (2019) discovered several targets for MDD, 12 genera where *Adlercreutzia*, *Bifidobacterium*, *Blautia*, *Clostridium XI*, *Eggerthella*, *Holdemania*, *Parabacteroides*, *Ruminococcus*, and *Streptococcus* increased whereas *Prevotella*, *Sutterella*, and *Megamonas* decreased in patients with depression [16].

To highlight the differences in gut microbiome due to sex, a sex-based study by Chen et al. (2018) found out that at the genus level, *Actinomyces*, *Bifidobacterium*, *Asaccharobacter*, *Atopobium*, *Eggerthella*, *Gordonibacter*, *Olsonella*, *Eubacterium*, *Anaerostipes*, *Blautia*, *Roseburia*, *Faecalibacterium*, and *Desulfovibrio* were the most abundant genera in female depressed patients and genera *Howardella*, *Sutterella*, and *Pyramidobacter* were the most abundant in female control patients. As for the male patients, the most abundant taxa in depressed patients were *Bacteroides*, *Erysipelotrichaceae incertae sedis*, *Veillonella*, and *Atopobium*, whereas for the control patients, it was *Anaerovorax*, *Gordonibacter*, and *Pyramidobacter* [26].

Another study by Hu et al. (2023) discovered that males had more differentially enriched bacteria in their microbiome and most of those bacteria belonged to *Bacteroides* but they stated that further research is required to find out whether this is a result of sex difference between patients [3].

In this thesis, however, sex was not taken into account during statistical analyses and machine learning modeling due to the imbalanced nature of the dataset. Apart from the imbalance between the sample size of both groups, there was an imbalance in the distribution of sex among each group as well which may potentially cause a bias during machine learning modeling.

2.3. Influence of Diet

The effects of diet on the gut microbiome can be seen starting from childhood where there are clear differences observed between infants that were breast-fed and formula-fed, with a greater prevalence of *Bifidobacteria* in breast-fed infants.

Unlike infant microbiome, an adult microbiome is primarily made up of *Bacteroidetes* and *Firmicutes* with lower amounts of phyla *Proteobacteria*, *Verrucomicrobia*, *Actinobacteria*, and *Cyanobacteria*, and genus *Fusobacteria*. There are also three enterotypes which are *Bacteroides*, *Prevotella*, and *Ruminococcus* based on the dom-

inant bacterial composition of a person. In adulthood, generally, a higher prevalence of *Prevotella* enterotype is observed in individuals reporting a higher intake of carbohydrates and simple sugars (e.g. Mediterranean diet) , and a higher prevalence of *Bacteroides* enterotype is observed in individuals reporting a higher intake of animal protein and saturated fats (e.g. Western diet). *Ruminococcus*, on the other hand, is the most common enterotype overall and it can be seen in people eating a mixture of proteins and sugars. These enterotypes are primarily determined by diet and genetics and they are independent of other environmental factors [25], [27].

A low quality diet is linked with a negative change in gut microbiota diversity, a reduction of intake in essential nutrients, aggravated gut dysbiosis and a decline in the growth of substrated specific microorganisms. A Western diet, which is high in saturated fats and sugars, is linked with cognitive decline and MDD as it can increase the production of endotoxins from commensal bacteria, and promote neuroinflammation and blood-brain-barrier leakage. Healthy dietary patterns like the Mediterranean diet, on the other hand, have been shown to promote healthier gut diversity, reducing inflammation and fighting against depressive symptoms and cause a higher ratio of *Bacteroides* to *Firmicutes* ratio. In addition, Norwegian and Japanese diets which are rich in omega-3, polyunsaturated fatty acids (PUFAs), whole grains, olive oil, greens, fish, soy and lower-fat dairy are linked with reduced risk of MDD [6], [27].

A diet rich in glucose, fructose and sucrose can increase the abundance of *Bifidobacteria* and decrease *Bacteroides*. An increase in lactose intake can also decrease the abundance of *Clostridia* cluster XIVa. A decrease in fiber consumption can result in decreased abundances of *Roseburia*, *Eubacterium rectale*, and *Faecalibacterium prausnitzii*. An increase in fiber consumption, on the other hand, can increase *Bifidobacteria* and *Lactobacillus* abundances [28].

Martin et al. (2023) also studied the relationship between mood & happiness and diet in their study and observed that high consumption of fat and protein can reduce anxiety and depression whereas carbohydrates had the opposite effect on pa-

tients. They also discovered that the total number of calories and fiber intake showed a negative correlation with the diversity of the patient’s gut microbiome, resulting in a decrease in anxiety and depression as the diversity increases. Apart from MDD, for instance, studies on children with epilepsy found that a Ketogenic diet can change their gut microbial composition and be beneficial for controlling severe cases of epilepsy [27].

In this thesis, however, diet was not taken into account during statistical analyses and machine learning modeling due to the vocabulary of the metadata. Since the metadata file that included information about dietary habits of the patients were not grouped by dietary preferences but rather by the frequency of their consumption (i.e., how many times they consumed the given food in a specific period of time) of a list of limited number of foods. As a result, this did not provide an objective space for categorizing the patients by their dietary preferences, leaving no option but to discard these information.

2.4. Understanding Microbiome Data and Machine Learning Models

Microbiome data are generated through the taxonomic identification of bacteria using high-throughput sequencing technologies by primarily profiling the V3-V4 region of the 16S rRNA marker gene to have a better understanding of the microbial communities they belong to but understanding and interpreting these relationships is a complex procedure. Even though this sequencing procedure is cost-effective, 16S rRNA gene sequencing still raises issues regarding how reliable these classifications can be which are mostly possible down to the genus level due to the limited amount of information on bacterial genes and their functions [29].

After the sequencing process, microbiome data are typically organized into forms of matrices where generally rows correspond to observed counts (abundances) of clustered sequences at a certain sequence similarity threshold known as operational taxonomic units (OTUs) which hold bacteria information and columns correspond to the samples (or vice versa). One drawback of OTUs is that the similarity threshold is

arbitrary and it can miss subtle and real biological sequence variations [30, 31]. Furthermore, when the abundance of a bacteria is small, species of the genus *Lactobacillus* for instance, it may fail in quantifying that bacteria on the OTU table [5]. OTUs are also slowly starting to be replaced by amplicon sequence variants (ASVs) which are the un-clustered error-corrected version of OTUs [29]. Certain properties of OTU tables can also cause further problems during the analyses and machine learning modeling if unaddressed. Some of these issues are:

- (i) Uneven sampling/library/read depth. Each sample can be represented by very different number of sequences, reflecting differential efficiency of the sequencing process rather than true biological variation. This variation can be due to differences in platform or technical difficulties in loading the same molar amounts of sequencing libraries on the instrument, or because of random variation. Uneven read sampling can cause problems especially during distance/dissimilarity calculations for multivariate ordinations derived from these distances [30], [32].
- (ii) Sparsity. Almost 90% of OTU tables are full of zeros [23], [30].
- (iii) Compositionality. Total number of reads obtained for a sample does not reflect the absolute number of microbes present, since the sample is just a fraction of the original environment. Thus, compositional data analysis methods must be preferred. Since OTU counts are not usually normally distributed, nonparametric tests are often used to analyze the data but these tests do not take the compositionality of the data into account [30].
- (iv) Dimensionality. Microbiome data is high-dimensional, there are often more features than samples, which can lead to overfitting and poor generalization of performance especially during machine learning modeling [23].

Apart from abundance matrix, drawing reliable and informative insights from microbiome data using metadata such as patient demographics, diet, and use of antibiotics could aid in the process of disease diagnosis but often necessary information in these metadata files is either missing or insufficient use of vocabulary makes it difficult for using them as machine-readable categorical covariates [23]. This was the case in

this dataset and its metadata, not allowing us to use sex and diet related covariates in statistical analyses and machine learning modeling.

In order to solve the issues related to the OTU matrices, most of the time, the dataset is normalized. Normalization is the process of transforming data to eliminate artificial biases. In microbiome data, these biases might occur because the quality of the data can be influenced by experimental, biological, and environmental factors such as differences in sample collection, the lab technician, library preparation, sequence filtering method, clustering threshold, reference of taxonomic assignment, choice of bioinformatic tools and pipelines, and can manifest as incomplete taxonomic ranking, uneven sampling depth and sparsity [23], [30]. One of the most popular normalization methods is rarifying (subsampling) the count data to a fixed depth. This, unfortunately, results in information loss to a certain extent, causing a loss of statistical power and decreasing the sensitivity of the overall performance [33]. Nonetheless, it is still one of the most common practices in microbiome studies and alternatives to rarefying also recommend discarding samples with low sampling depths when they cluster separately from the rest of the data. One way to observe whether normalization has been successful is to do ordination analyses on normalized data to observe whether the sample groupings are based on biological or artificial features [30]. Unfortunately, this can be quite difficult to observe when there is an imbalance between the groups.

Different ways of preprocessing can affect the performance of statistical analyses and machine learning models, possibly introducing biases. To reduce the noise in the dataset, outlier samples or features must be removed and features in low abundance must be eliminated [23]. Depending on the scope of the project, some scientists prefer to keep all of the features in the dataset whereas some focus on specific taxonomic rank for more specific analyses, also removing features that are below a prevalence threshold.

Even though ML modeling is a promising predictive tool for host dysbiosis prediction studies, there are various challenges. The high-dimensional and heterogeneous

nature of the microbiome data can make the model training process difficult, resulting in overfitting with difficulties in developing a robust and accurate prediction model. To prevent this overfitting problem, apart from feature selection, techniques such as cross-validation, reduction of model complexity, augmenting the training data by sample collection or creating new synthetic data, or undersampling (removing data) in certain cases can be implemented into the pipeline [34]. In this thesis, we have used all of these methods to compare different models and validate our final results when finding potential biomarkers from our models of interest.

In gut microbiome studies, several supervised ML algorithms such as Support Vector Machines (SVM), Linear Regression models, Naïve Bayes (NB), and K-nearest Neighbor (kNN) have been used and compared in the literature but decision trees-based ensemble learning methods in particular, such as Random Forest (RF) and Extreme Gradient Boosting Decision (XGBoost), have been applied the most because they tend to have a better performance, they are easier to use and they are more interpretable compared to their counterparts [23], [34,35]. In the RF algorithm, multiple decision tree models are trained using a random subset of training data in a simultaneous manner and then their predictions are aggregated for a final prediction outcome, improving the performance of a single decision tree and reduce the risk of overfitting. In the XGBoost algorithm, instead of the independent construction of decision trees in RF, each weak learner is sequentially trained and improved based on the previous one by minimizing the loss function until a good training score is reached [34]. One reason for the popularity of the RF algorithm in gut microbiome studies is its ability to take a large number of features into account during training and it is resistant to the presence of noise and emissions in the dataset [11]. Even though RF and XGB were our main interest in this thesis due to their model interpretability, we have also implemented Logistic Regression (LR) and Decision Tree (DT) to compare results and observe how different models behave under different circumstances.

Due to the "black-box" nature of machine learning models, model interpretability has been gaining interest among scientists to understand why a model makes a certain

prediction. However, the complex nature of models has been creating a gap between accuracy and interpretability. This becomes an issue in clinical decision-making because these models make it difficult to understand the relationship between cause and effect [35]. Unlike their counterparts, decision tree-based algorithms are not entirely “black-box” in nature because they can identify and rank important features by their scores, using Gini importance scores for RF and F-scores for XGBoost, making them more suitable for gut microbiome research [34]. This is why we have focused more on RF and XGB instead of LR and DT while looking for potential biomarkers.

Even though we can retrieve feature importances ranked by a certain score depending on the model, these scores do not give information about how these features affect the prediction result. Many methods have been developed to understand the contributions of features to the model results so that we can interpret why such predictions are made. SHAP (SHapley Additive exPlanations) is one of them which aims to explain the prediction of our desired classification result by computing the contribution of each feature to that prediction. In other words, the SHAP explanation method computes Shapley values from coalitional game theory where the feature values of a data instance act as players in a coalition. Shapley values tell us how to fairly distribute the prediction among the features by assigning importance values to each feature [36]. It measures the contribution of each feature to the prediction by considering all possible combinations of features fairly and it can work with any kind of model and can be visualized [23]. As a result, by looking at the SHAP value of a feature, we can observe the direction of its contribution to the prediction. Generally, if the value is positive, the feature works in favor of the positive class (MDD class in our case). Our aim was to find SHAP values of potentially important features for all version of models and then to find the intersections of these features, ending up with our potential biomarkers of interest.

Besides MDD, RF modeling also has been studied for analyzing and understanding the gut microbiome and its relation to diseases such as irritable bowel disorder, cardiovascular diseases, obesity, colorectal cancer and hypertension for phenotypic pre-

diction [11], [34]. There are also use cases of RF modeling for phenotyping sponge bacterial density and crop productivity in some studies [35].

Leveraging the feature importance property of certain models has been a recurring method in the literature, as it was also used in this thesis. A study on Myasthenia gravis (a long-term neuromuscular junction disease) used both Random Forest and XGBoost along with other models with leave-one-out cross-validation (LOOCV) to find the intersection of potential biomarkers from filtered and raw model results by retrieving important features [37]. Another study by Angelova et al. (2023), retrieved significant features using an RF model for MDD from a cohort of 74 samples with ages ranging from 18 to 54 years old from Moscow (38 Control, 36 MDD samples). The selected important features were *Bacteroides caccae*, *Coprococcus comes*, *Eggerthella lenta*, *Escherichia coli*, and the most significant one being *Faecalibacterium prausnitzii*. Their RF model has obtained an accuracy rate of 80% with a false-positive prediction rate of 10% and a false-negative prediction rate of 30%, showing consistency with their previous study in terms of their significant features and results [11].

Another study used RF feature importances to differentiate the potential biomarkers between 155 healthy control and 138 patients with depression on three different scales (mild: 24, moderate: 72 and severe: 42) and the researchers found out that *Eubacterium hallii*, *Blautia obeum*, *Clostridium*, *Dorea longicatena*, *Enterobacter cloacae*, *Eubacterium*, *Klebsiella*, *Ruminococcus gnavus*, and *Shigella sonnei* could be potential biomarkers for healthy controls. For the MDD group; *Bacteroides coprophilus*, *Bifidobacterium longum*, and *Parasutterella excrementihominis* for the mild group, *Bacteroides massiliensis*, *Bacteroides stercoris*, *Bilophila wadsworthia*, and *Eubacterium* for the moderate group, and finally *Bacteroides dorei*, *Bacteroides faecis*, *Bifidobacterium longum*, and *Ruminococcus gnavus* for the severe group were found as potential biomarkers. The diagnostic accuracy of the RF models using these important features gave a high area under the curve of receiver operation characteristics (ROC-AUC) with area values around 99% [3].

2.5. Machine Learning Performance Evaluation Metrics

When you train a machine learning model to the validation or test set, in order to measure the performance of your model depending on the preferred scoring metric, you obtain statistical scores using a confusion matrix (Table 2.1). There are 4 elements of this matrix:

- (i) True Positive (TP): an MDD sample correctly predicted as an MDD sample
- (ii) True Negative (TN): a Control sample correctly predicted as a Control sample
- (iii) False Positive (FP): an MDD sample mistakenly predicted as a Control sample
- (iv) False Negative (FN): a Control sample mistakenly predicted as an MDD sample

Table 2.1. Confusion Matrix.

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Using these elements from the confusion matrix, we can compute several scoring metrics to have a better understanding of the model predictions. These metrics are:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2.2)$$

$$Recall \text{ (True Positive Rate)} = \frac{TP}{TP + FN} \quad (2.3)$$

$$False Positive Rate = \frac{FP}{TN + FP} \quad (2.4)$$

$$F1\text{-score} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2.5)$$

In the case of imbalanced datasets, instead of relying on Accuracy or F1-score, it is better to obtain the area under the receiver operating characteristic (ROC-AUC) and area under the precision-recall curve (PR-AUC) to evaluate model performances.

A ROC curve shows the trade-off between true positive rates (TPR) and false positive rates (FPR) across different decision thresholds and tells about how well a model can discriminate between positive and negative samples. TPR measures the fraction of MDD samples that are correctly predicted whereas FPR measures the fraction of Control samples that were classified as MDD samples. Even though ROC-AUC is a useful metric to evaluate the model performance, it still can give misleading results when there is a severe class imbalance because it fails to capture the effect of a large number of negative (Control) samples on the performance. Hence, PR curves (and PR-AUC graphs) are preferred as a solution for the severe imbalance problem to investigate the performance if there is no data augmentation involved [38].

In this thesis, during the evaluation of the models, we have observed both the Precision-Recall scores to find an optimal tradeoff between the metrics while observing both of the AUC scores (PR-AUC and ROC-AUC).

3. MATERIAL & METHODS

3.1. Data

The data used in this paper is from the American Gut Project (AGP), which is essentially a collaborative work between the Earth Microbiome Project (EMP) and the Human Food Project (HFP) that aims to create a cohort of samples for the discovery of microbes and microbiomes by establishing an open source platform for research and analysis [39]. Collections of raw 16S rRNA data samples with sequences that were trimmed to 125 nt have been filtered and analyzed through Qiita with the focus on the V4 region using a 515F/806R primer pair (515F: GTGCTYAGCMGCCGCGGTAA and 806R: GGACTACNVGGGTWTCTAAT). After the sequences were denoised by using deblur, a threshold of 97% OTU sequence identity using the Greengenes 13.8 data bank as a reference was set for the closed reference OTU picking for taxonomic assignment. This threshold is commonly set in practice because in general, instead of treating it like a tunable hyperparameter, 97% is the standard value since it corresponds approximately to species [40].

The samples went through a preliminary filtering process using the metadata. We only took human fecal samples into account which adhered to the following criteria; have not taken probiotics ever, have not used antibiotics in the last year and does not have the following conditions: cancer, alzheimers, epilepsy or any seizure disorder, any eating disorder such as anorexia or bulimia, post traumatic stress disorder (PTSD), inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), diabetes, substance abuse, bipolar disorder, autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD) and schizophrenia. Moreover, samples with no information regarding their age, weight, body mass index and sex were also discarded. At the end of the metadata filtering process, there were 23 MDD and 361 control samples with a total of 384 samples in our cohort (Table 3.1). 65% of the MDD samples were female whereas for the control samples, the rate was 44%. Overall, the cohort skewed

towards the control group causing an imbalance. Demographic variables between MDD and Control group were compared using Chi-Square test for categorical variables (sex) and Mann-Whitney U test for non-normal continuous variables (age, weight, height, body mass index). The normality of the continuous variables were investigated using Shapiro-Wilk test and histograms, indicating non-normal distribution among variables. Differences were considered statistically significant when the p-value was smaller than 0.05 and results showed that only age showed a significant difference (p-value: 0.02).

Table 3.1. General sample characteristics of the cohort.

	MDD (N=23)	Control (N=361)	X^2/W	p-value
Age (years)	54.7 $\pm$ 12.6	45.1 $\pm$ 18.9	2954.5	0.02044
Sex	15F / 8M	159F / 202M	3.1039	0.0781
Body mass index	24.7 $\pm$ 4.21	24.6 $\pm$ 11.6	3676	0.3574
Weight (kg)	73.6 $\pm$ 14.1	71.2 $\pm$ 25.1	3949.5	0.6962
Height (cm)	173 $\pm$ 10.7	170 $\pm$ 24	4163	0.983

Before starting our exploratory data analysis, a prevalence filtering was done with a threshold of 10% which removed OTUs that were found in less than 10% of the samples to reduce sequencing errors by eliminating singletons and other possible contaminants. The taxonomy table was also aggregated at the species level to prevent copies of the same OTUs/features from appearing again and samples unclassified at the genus level were discarded for statistical power. By the end of this process, we ended up with a total number of 129 OTUs/features.

3.2. Diversity Analysis

To understand the characteristics of the community on inter- and intra-levels, two diversity metrics were used. The first one is alpha diversity (also known as species diversity) which summarizes the distribution of species abundances in a given sample into a singular number that depends on species richness and evenness. Richness refers to

the total number of species in a community and evenness refers to how close the number of individuals of different species is in the community. In general, diversity increases together with increasing richness and evenness. Among the several indices used to determine diversity in the sample, Shannon’s index, Simpson’s index and Observed Species index were employed here using the `vegan` [41] package on R. These diversity indices measure the overall community heterogeneity.

The second one is beta diversity which refers to microbiota differences between samples or groups. The index used in this study was the Bray-Curtis dissimilarity index. Beta diversity indexes are represented with distance objects with a triangular structure, thus followed by the process of ordination which aims to reduce the dimensionality of the data for further evaluation or visualization. Various ordination methods have been employed in this paper, such as Non-metric Multi Dimensional Scaling (NMDS) and Principal Coordinates Analysis (PCoA). There are many dissimilarity indexes used in the literature, Bray-Curtis being the most popular one in the microbiome literature. Before calculating the dissimilarity matrix, as a method of normalization, rarefaction takes place. Due to factors like random variation, choice of platform and technical issues in uploading same molar amount of sequences on the device, it is very difficult to obtain an abundance table with exact total number of sequence reads for each sample (read/library depth). This variance in depth is a major confounder for “dissimilarity calculations for multivariate ordinations derived from these distances” [42]. In such cases, rarefaction has been a common normalization practice despite being heavily criticized as it causes loss of information and precision, possibly resulting in an increased amount of false positive rates [43], can reduce statistical power and does not address all challenges of compositional data [23]. To test the statistical significance of beta diversity, we have used PERMANOVA [44], which is a common distance-based method to test the association of microbial composition with covariates of interest by directly using the distance matrix to partition the diversity among sources of variation.

3.3. Differential Abundance Analysis

Differential abundance analysis (DAA) is one of the central tasks of microbiome data analysis to find potential microbial features, also commonly known as biological markers (biomarkers) which show covariance with a variable of interest using tools that compare the differences in relative abundance across samples of high-throughput RNA sequences [45]. One major problem with these tools is that they do not always take into account the compositional nature of the data and have the tendency to be sensitive to their large variability and the negative correlation bias. Moreover, the amount of zeros and small sample size with the inability to adjust covariates can be a limiting factor for the performance of such tools. It is also important to mention that given the prevalence of zeros, the identification of a differentially abundant taxa does not guarantee it to be biologically meaningful [46].

Table 3.2 provides information about the DAA methods used in this thesis and how the dataset was prepared for these methods. The assumed distribution for each method is also given with their default hypothesis test for checking the statistical significance of the results and whether they use False Discovery Rate (FDA) correction or not. The FDA correction method that was chosen for these DAA methods was the Benjamini-Hochberg (BH) method.

Table 3.2. Differential Abundance Analysis methods employed in this thesis.

Method	Input	Normalization	Transformation	Distribution	Hypothesis test	FDA Correction
DESeq2 [47]	Counts	Modified RLE	None	Negative binomial	Wald (default)	Yes
LEfSe [48, 49]	Counts	None	None	Non-parametric	Kruskal-Wallis	No
Wilcoxon (rare)	Rarefied counts	None	None	Non-parametric	Wilcoxon rank-sum	Yes
t.test (rare)	Rarefied counts	TSS	None	Normal	Welch's t.test	Yes
ALDEx2 [32, 50]	Counts	None	CLR	Dirichlet-multinomial	Wilcoxon rank-sum	Yes
MaAsLin2 [51]	Counts	TSS	AST	Normal	Wald	Yes

3.4. Machine Learning Modeling

Machine Learning is a subfield of artificial intelligence in which datasets are used to train and develop predictive models using computational and statistical methods [52]. As the current reliance on traditional bioinformatics methods in the microbiome field has started to involve machine learning modeling, further improvements in clinical applications have emerged. Disease diagnosis, prognosis and target response to drug treatments are one of the few examples of machine learning practices in this area. Furthermore, as this thesis aims to do so, discovering potential biomarkers in the microbiome to unravel their relationship with diseases is one of the use cases for early detection of diseases [23].

Certain machine learning models are more suitable over others because of the properties of microbiome data which are the sparsity (high amount of zeros), high dimensionality and class imbalance. Since we are aiming to find out whether samples have MDD or not, several binary classifiers were employed with additional methods to unravel possible biomarkers. We have built Logistic Regression (LR), Random Forest (RF), Decision Tree (DT) and XGBoost (XGB) models for comparison. All of the models were implemented using the scikit-learn [53] package.

Before the finalized implementation of the models, we have experimented pilot models with the same dataset to investigate the effects of different data augmentation methods to observe which one would give the best results. Models were observed in 4 different cases: raw models (no data augmentation), undersampled models, oversampled models, and ratioed models (groups were augmented to a 2:1 ratio using both oversampling and undersampling).

First, the dataset was split into a train/test set using 70/30 ratio. Then the training set was split into validation/test set using Leave One Out Cross Validation (LOOCV) method where 1 sample was kept for the test set and the rest of the data for the validation set. All models were kept as count values except LR where the

counts were scaled using MinMaxScaler. This happened after the validation split to prevent any data leakage. During this phase, the hyperparameters were tuned using GridSearchCV by using the f1-score as the score metric. The defined hyperparameter values were given in Table 3.3.

Table 3.3. Hyperparameters that were tuned for each model.

	Hyperparameters
LR	penalty: ['l2'] solver: ['newton-cg', 'saga', 'sag', 'liblinear'] C: [0.001, 0.01, 0.1, 1, 10, 100, 1000]
RF	n_estimators: [25, 50, 100] max_depth: [3, 6, 9] max_leaf_nodes: [3, 6, 9] max_features: ['sqrt', 'log2'] criterion: ['gini', 'entropy']
DT	max_depth: [2, 3, 5, 10, 20] min_samples_leaf: [5, 10, 20, 50, 100] criterion: ['gini', 'entropy']
XGB	max_depth: [6, 9, 12, 15] min_child_weight: [1, 10, 100]

After the hyperparameter tuning, the best model was chosen to fit the actual test set to reveal the results for each model. As mentioned before, raw models were not augmented during cross validation. Undersampled and Oversampled models were augmented using NearMiss [54] and SMOTE [55] methods respectively. Ratioed models were augmented with the combination of both aforementioned methods with a 2:1 ratio where the sample size of the minority group was oversampled to the 10% of the majority group and then the majority group was undersampled to the double of minority group's sample size. A little amount of oversampling was done in the first place to shift the bias from the majority group. After comparing all augmentation

methods, as undersampling gave the most reliable results, we also compared it with an alternative undersampling method called RandomUnderSampler [54] which randomly picks samples from the majority group instead of relying on heuristics. We compared the NearMiss method with 2 different versions of the RandomUnderSampler (2 different seeds) method to validate why we chose to move on with NearMiss method.

After finalizing the pilot experiments, in order to get the best results, we have decided to undersample the entire dataset using NearMiss method due to the severe imbalance problem, reaching 46 samples in total. After the undersampling process, the dataset was split into training and test set using 70/30 ratio. Same model architectures from the pilot experiments were used in this case except data augmentation was removed from the hyperparameter tuning process as the groups were already balanced. Feature importance values were retrieved from models with the best area under curve (ROC-AUC and PR-AUC) scores and Precision & Recall scores which were RF and XGB models. SHAP values of features from these models were also retrieved to analyze their impact on model results. Finally, potential biomarkers were found depending on average feature importance scores and their behaviors according to their SHAP values. Same procedure was done for an alternative version of this new dataset with features retrieved from the DAA results, which we called DAA filtered dataset, and similar comparisons were made among potential biomarkers after retrieving average feature importances and SHAP values for both models. Finally, the intersection of biomarkers retrieved from the raw and DAA filtered datasets were decided to be the most reliable biomarkers to be found from this thesis.

4. RESULTS

4.1. Diversity Analysis on Gut Microbial of MDD Individuals

4.1.1. Alpha and Beta Diversity

After completing the data preparation process, alpha diversity indexes were calculated to investigate the microbial biodiversity using the raw abundance data. Shannon, Simpson, and Observed Species indexes were measured respectively (Figure 4.1).

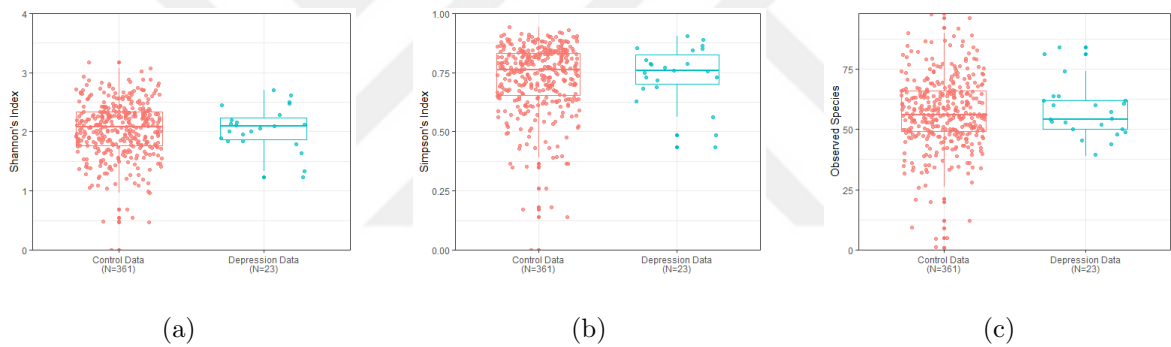


Figure 4.1. Alpha diversity representation of the groups using (a) Shannon, (b) Simpson, and (c) Observed Species indexes.

MDD and Control group do not show a significant difference between their mean values for any of the indexes ($p=0.85$, $p=0.804$, $p=0.963$ respectively).

The Qiita analysis did not provide a phylogenetic tree, therefore, only the Bray-Curtis dissimilarity index was calculated which showed no statistical difference between the groups. Before calculating the dissimilarity matrix, the abundance data was rarefied at a chosen sampling depth ($n=4000$). To visualize the results in a two-dimensional space, NMDS and PCoA were used as shown in Figure 4.2. PERMANOVA showed no statistical significance between the groups ($p=0.901$).

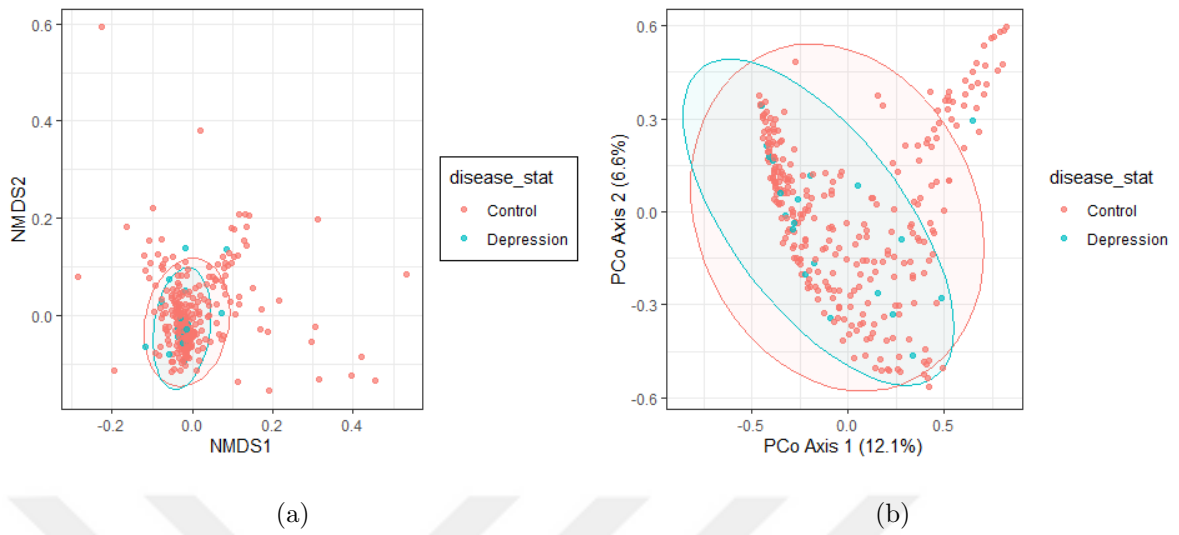


Figure 4.2. Beta diversity representation of the groups using (a) NMDS and (b) PCoA ordinations.

4.1.2. Mean Relative Abundance

A bar chart was applied to visually demonstrate the predominant genera at the genus level using the mean relative abundance values (Figure 4.3). The most abundant genus for both groups was *Bacteroides*. Among the genera with the highest mean relative abundances in the cohort, the abundance of *Prevotella* was higher in the Control group whereas the abundances of *Faecalibacterium*, *Bacteroides*, *Ruminococcus* and *Blautia* were higher in the MDD group

4.2. Statistical Analysis on MDD Data

Several differential abundance analysis methods, as described in Table 2, were used to find potential biomarkers that were statistically significant.

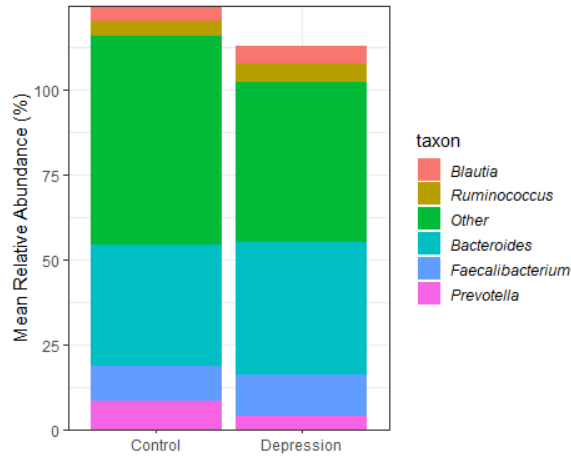


Figure 4.3. Mean relative abundances of the cohort on Genus level.

4.2.1. DESeq2

DESeq2 primarily found statistically significant genera from the phyla *Actinobacteria*, *Bacteroidetes*, *Firmicutes* and *Proteobacteria* for the control group and *Fusobacteria* for the MDD group. MDD samples had an increase in species *Fusobacterium unclassified* and a decrease in unclassified species of genera *Dialister*, *Comamonas*, *Bacillus*, *Bifidobacterium*, *Lactococcus*, *Anaerococcus*, *Acinetobacter*, and species of *Enterobacter radicincitans*, *Prevotella copri*, *Pseudomonas viridiflava*, *Bacteroidetes eggerthii*, *Bifidobacterium adolescentis*, *Bacteroides ovatus*, and *Haemophilus parainfluenzae* respectively (Figure 4.4).

The MA plot (Figure 4.5), the volcano plot (Figure 4.5), and the heatmap (Figure 4.6) for DESeq2 results can be seen respectively.

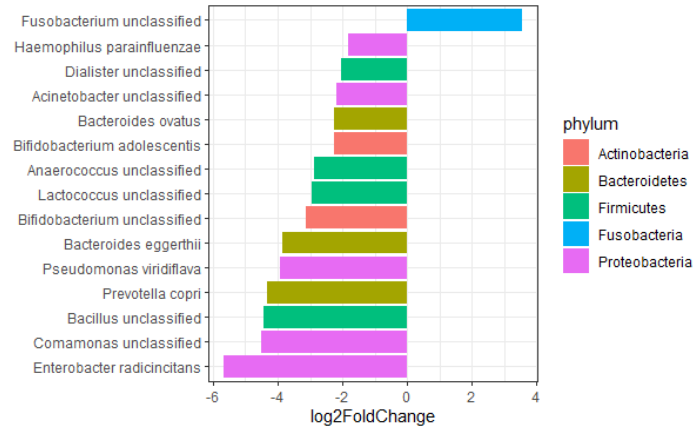


Figure 4.4. DESeq2 results.

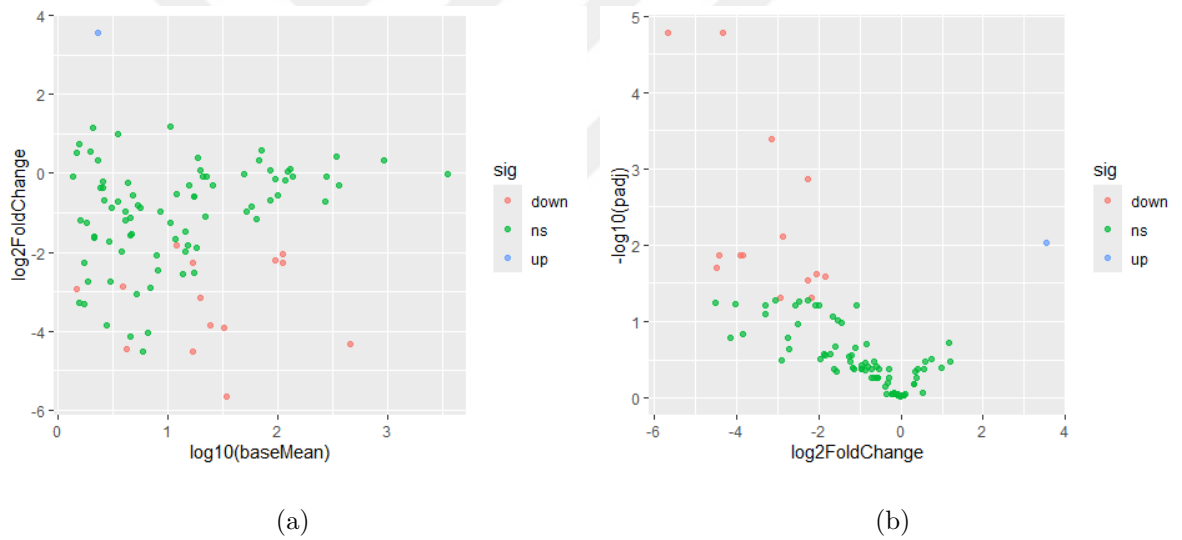


Figure 4.5. (a) MA and (b) Volcano plots derived from DESeq2 results.

4.2.2. LEfSe

After filtering the LDA scores ($LDA > 2$) of LEfSe, MDD samples had an increase in the unclassified species of genera *Ruminococcus*, *Adlercreutzia*, *Anaerotruncus*, and *Blautia producta* and a decrease in species *Coprobacillus unclassified* and *Bifidobacterium adolescentis* (Figure 4.7).

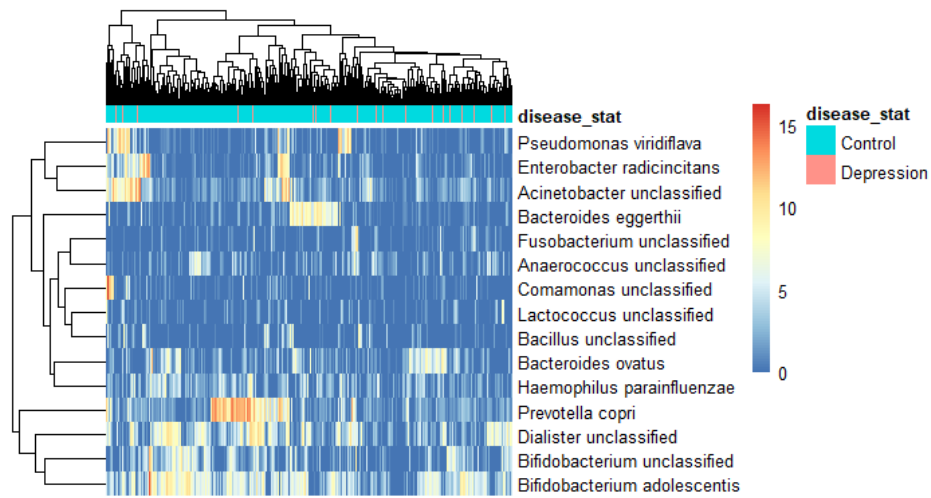


Figure 4.6. Heatmap from DESeq2 results.

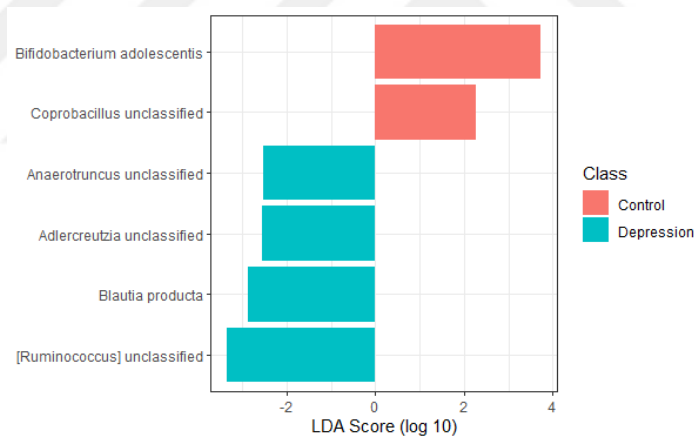


Figure 4.7. LefSe results.

4.2.3. Wilcoxon (rarefied)

Wilcoxon (rarefied) results gave 6 statistically significant taxa which were *Enterobacterium dolichum*, *Adlercreutzia*, *Anaerotruncus*, *Prevotella melaninogenica*, *Coprobacillus*, and *Bifidobacterium adolescentis*. The first two were increased in MDD samples and the rest were decreased in abundance by their mean comparison (Figure 4.8).

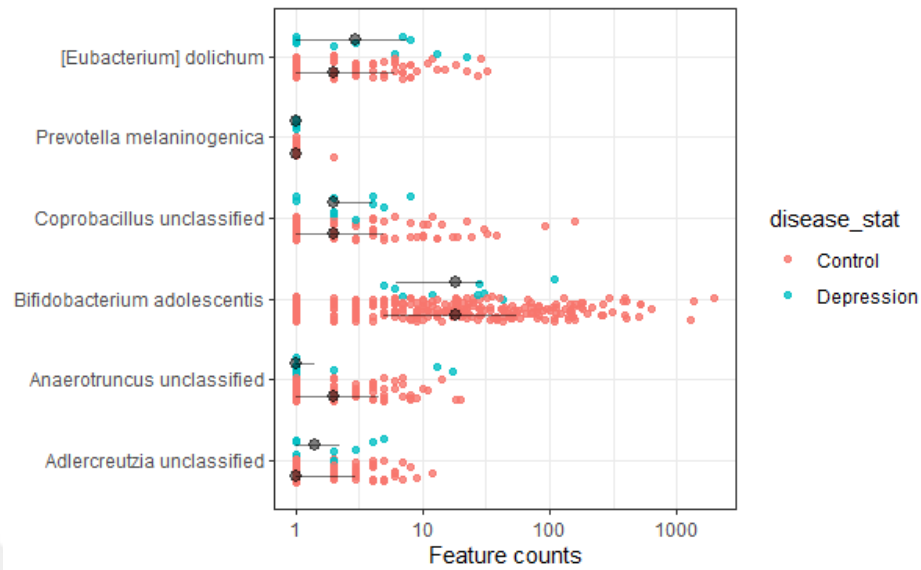


Figure 4.8. Wilcoxon (rarefied) results.

4.2.4. Welch's t.test (rarefied)

Welch's t.test (rarefied) results gave us eleven statistically significant taxa which were *Ruminococcus callidus*, *Turicibacter*, *SMB53*, *Planomicrobium*, *Mogibacterium*, *Haemophilus parainfluenzae*, *Facklamia*, *Enterobacter hormaechei*, *Bifidobacterium*, *Bifidobacterium adolescentis*, and *Aquamonas haywardensis*. Only the first taxon was more abundant in MDD samples, the rest were decreased in terms of their relative abundance (Figure 4.9).

4.2.5. ALDEx2

ALDEx2 gave no statistically significant results after BH correction. This requires further analysis as it could be due to the filtering process and the nature of the dataset.

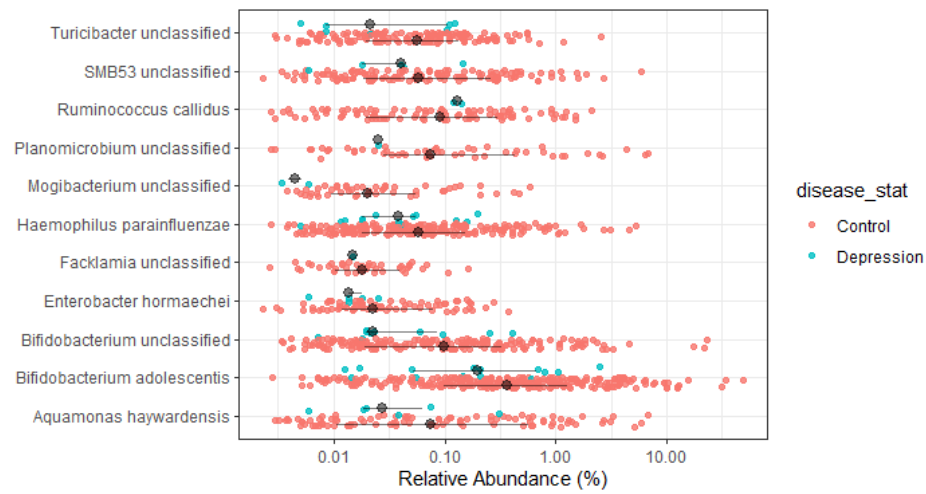


Figure 4.9. Welch's t.test (rarefied) results.

4.2.6. MaAsLin2

MaAsLin2 gave two statistically significant genera after BH correction, showing an increase of abundance in the MDD samples which were *Fusobacterium* and *Odoribacter*.

4.2.7. Summary of DAA Results

Table 4.1 summarizes the potential biomarkers found by DAA methods and their abundances in MDD patients. Table 4.2 shows the common features found by different DAA methods and counts them.

Table 4.1. The potential biomarkers identified by applying DAA methods.

Method	Biomarkers identified for MDD	Remark for MDD	Gram-stain	Statistical significance
DESeq2	Fusobacterium,	Increased	GN	9.362842e-03
	Enterobacter radicincitans,	Decreased	GN	1.684791e-05
	Comamonas,	Decreased	GN	1.948962e-02
	Bacillus,	Decreased	GP	1.356054e-02
	Prevotella copri,	Decreased	GN	1.684791e-05
	Pseudomonas viridiflava,	Decreased	GN	1.356054e-02
	Bacteroidetes eggerthii,	Decreased	GN	1.356052e-02
	Bifidobacterium,	Decreased	GP	4.165873e-04
	Lactococcus,	Decreased	GP	4.855522e-02
	Anaerococcus,	Decreased	GP	7.832889e-03
	Bifidobacterium adolescentis,	Decreased	GP	1.365348e-03
	Bacteroides ovatus,	Decreased	GN	2.887609e-02
	Acinetobacter,	Decreased	GN	4.855522e-02
	Dialister,	Decreased	GN	2.363039e-02
	Haemophilus parainfluenzae	Decreased	GN	2.603814e-02
LEfSe	Ruminococcus,	Increased	GP	0.04359561
	Blautia producta,	Increased	GP	0.0112983
	Adlercreutzia,	Increased	GP	0.0188570
	Anaerotruncus,	Increased	GP	0.0134515
	Coprobacillus,	Decreased	GP	0.0212046
	Bifidobacterium adolescentis	Decreased	GP	0.0358800
Wilcoxon (rare)	Eubacterium dolichum,	Increased	GP	0.01577876
	Adlercreutzia,	Increased	GP	0.01668246
	Anaerotruncus,	Decreased	GP	0.04353992
	Prevotella melaninogenica,	Decreased	GN	0.01575956
	Coprobacillus,	Decreased	GP	0.02133292
	Bifidobacterium adolescentis	Decreased	GP	0.0061819145
Welch's t.test (rare)	Ruminococcus callidus,	Increased	GP	0.005063958
	Turicibacter,	Decreased	GP	0.04035049
	SMB53,	Decreased	GP	0.005063958
	Planomicrobium,	Decreased	GP	0.04210176
	Mogibacterium,	Decreased	GP	0.01397642
	Haemophilus parainfluenzae,	Decreased	GN	0.005063958
	Facklamia,	Decreased	GP	0.04296341
	Enterobacter hormaechei,	Decreased	GN	0.01397642
	Bifidobacterium,	Decreased	GP	0.040350493
	Bifidobacterium adolescentis,	Decreased	GP	0.005063958
	Aquamonas haywardensis,	Decreased	GN	0.013957642
ALDEx2	-	-	-	-
MaAsLin2	Fusobacterium	Increased	GN	0.001449658
	Odoribacter	Increased	GN	0.015795218

Table 4.2. Biomarkers identified by all types of differential abundance analysis methods, the arrows represent the decrease/increase of the abundance of biomarker in MDD patients.

Bacterial Biomarker for MDD	Gram stain	DESeq2	LefSe	Wilcoxon	Welch's t.test	MaAsLin2	Count
Acinetobacter	Gram-Negative	↓					1
Adlercreutzia	Gram-Positive		↑	↑			2
Anaerococcus	Gram-Positive	↓					1
Anaerotruncus	Gram-Positive		↑	↓			2
Aquamonas haywardensis	Gram-Negative				↓		1
Bacillus	Gram-Positive	↓					1
Bacteroides ovatus	Gram-Negative	↓					1
Bacteroidetes eggerthii	Gram-Negative	↓					1
Bifidobacterium adolescentis	Gram-Positive	↓	↓	↓	↓		4
Bifidobacterium	Gram-Positive	↓			↓		2
Blautia producta	Gram-Positive		↑				1
Comamonas	Gram-Negative	↓					1
Coprobacillus	Gram-Positive		↓	↓			2
Dialister	Gram-Negative	↓					1
Enterobacter hormaechei	Gram-Negative				↓		1
Enterobacter radincitans	Gram-Negative	↓					1
Eubacterium dolichum	Gram-Positive			↑			1
Facklamia	Gram-Positive				↓		1
Fusobacterium	Gram-Negative	↑				↑	2
Haemophilus parainfluenzae	Gram-Negative	↓			↓		2
Lactococcus	Gram-Positive	↓					1
Mogibacterium	Gram-Positive				↓		1
Odoribacter	Gram-Negative					↑	1
Planomicrobium	Gram-Positive				↓		1
Prevotella copri	Gram-Negative	↓					1
Prevotella melaninogenica	Gram-Negative			↓			1
Pseudomonas viridiflava	Gram-Negative	↓					1
Ruminococcus callidus	Gram-Positive				↑		1
Ruminococcus	Gram-Positive		↑				1
SMB53	Gram-Positive				↓		1
Turicibacter	Gram-Positive				↓		1

4.3. Machine Learning Models on MDD Data

4.3.1. Pilot Experiments

Before discarding samples from the Control group to balance the dataset, several approaches were tested to observe the behavior of models. 4 data augmentation methods were employed which were raw, undersampled, oversampled, and ratioed (2:1) models. After the dataset was split into training and test set using 70/30 ratio, the training set was split into validation and test sets using LOOCV and hyperparameters were tuned using GridSearchCV. After the tuning process was done, best model was retrieved in each case and the test set was used to fit the model to obtain results as given in the following tables.

Table 4.3 shows the results for raw models (no data augmentation). It can be clearly seen from the accuracy results and f1-scores that each model shows a bias towards the Control group, leaving us with inconclusive results as the RF, DT, and XGB models failed to identify any MDD samples. It is clear that this approach is inadmissible for this dataset as the severe imbalance causes overfitting and fails to differentiate MDD samples from the Control group.

Table 4.3. Results of all raw ML models during pilot experiments.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC	PR-AUC
LR	85.3%	11%	8.3%	14.3%	51.5%	6.4%
RF	94%	0%	0%	0%	44.1%	8.5%
DT	90%	0%	0%	0%	41.7%	6%
XGB	94%	0%	0%	0%	41.2%	5.5%

Table 4.4 shows the results for undersampled models using NearMiss method. Unlike the raw models, this time, there are no inconclusive results but they are still very low for the MDD group, leaving us with PR-AUC and Precision values of less than 10% for each model. Low Precision values with an average of around 50% of Recall

values shows us that the all models failed to capture almost half of the actual MDD samples from the test set and misclassified a big amount of Control samples as MDD samples. Since we are interested in finding potential biomarkers linked to MDD, we want to achieve a higher Precision value for our models with a low amount of False Positives.

Table 4.4. Results of all undersampled ML models during pilot experiments.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC	PR-AUC
LR	23%	10%	5.4%	71%	56.4%	8%
RF	46%	11.2%	6.3%	57.1%	50.5%	7.2%
DT	51%	10%	5%	43%	45.6%	5.5%
XGB	52%	12%	7%	57.1%	50.3%	6.8%

Table 4.5 shows us the results for oversampled models using SMOTE method. The accuracy scores are higher than 70% with very low f1-scores, indicating a case of overfitting for the Control group. XGB model failed to identify any MDD samples, whereas the rest of the models resulted in low Precision and Recall scores. This shows us that the models failed to identify MDD samples and classified them as the majority group, causing an increase in the number of False Negatives which explains the decrease in Recall scores. Low Precision scores, which are a result of high False Positives, hints us that these models failed to get relevant information during the training phase, being unable to differentiate sample groups from each other and resulting in a mismatch of classification.

Table 4.6 shows us that the results for ratioed models using both NearMiss and SMOTE methods with a ratio of 2:1 for groups. The first oversampling aims to increase the number of minority samples so that the bias from the majority group can be moved to the minority and the undersampling discards samples from the majority group to remove the bias and possibly solve the imbalance problem.

Table 4.5. Results of all oversampled ML models during pilot experiments.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC	PR-AUC
LR	70%	15%	9%	43%	54.5%	14.3%
RF	77%	7%	5%	14%	41.2%	6%
DT	71%	6%	3%	14%	50.3%	6.3%
XGB	83%	0%	0%	0%	43%	5.6%

The issue in this case is that both synthetic data is added and possibly useful data is removed from the dataset. This approach failed for the RF and XGB models, with a very poor f1-score, Precision and Recall values, indicating a high number of False Negatives and False Positives which means the models mismatched the groups of the test samples, being unable to differentiate the samples. LR and DT models, however, showed better Recall scores than their counterparts. This difference between 2 to 2 model groups made this method unreliable for the future despite being somewhat promising.

Table 4.6. Results of all ratioed ML models during pilot experiments.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC	PR-AUC
LR	54%	16%	9%	71%	51.4%	6.2%
RF	67%	10%	6%	29%	51.9%	6.9%
DT	60%	21%	12%	86%	71.8%	10.2%
XGB	60%	8%	5%	29%	53.6%	7.1%

Overall, raw and oversampled methods resulted in inconclusive results which made them ineligible for our actual modeling. When we compared undersampled and ratioed methods, the overall behavior of models could be observed as a trend in the undersampled case unlike the ratioed method where LR and DT models performed differently than RF and XGB models. This difference was less dramatic in the undersampled version and the misclassification of the models could easily be explained by their low accuracy results. Furthermore, instead of adding synthetic samples to our

dataset, making a sacrifice and discarding samples from the majority group seemed like a better idea. In order to observe the effect of the NearMiss method for under-sampling, we compared it with the RandomUnderSampler method with 2 seeds. The chosen seeds were the default value (a random value chosen by the algorithm) and 100. The comparison of 3 methods for all models can be seen in the Table 4.7.

Table 4.7. Comparison of all models using undersampling methods NearMiss, RandomUnderSampler (random_state="default"), and RandomUnderSampler (random_state=100) respectively.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC	PR-AUC
LR_NearMiss	23%	10%	5.4%	71%	56.4%	8%
LR_RandomUnderSampler (default)	41%	15%	8%	86%	58.5%	8.2%
LR_RandomUnderSampler (100)	48%	12%	7%	57%	55%	6.7%
RF_NearMiss	46%	11.2%	6.3%	57.1%	50.5%	7.2%
RF_RandomUnderSampler (default)	44%	8%	5%	43%	41.7%	5.8%
RF_RandomUnderSampler (100)	48%	12%	7%	57%	56.6%	8.7%
DT_NearMiss	51%	10%	5%	43%	45.6%	5.5%
DT_RandomUnderSampler (default)	50%	9%	5%	43%	52%	6.5%
DT_RandomUnderSampler (100)	39%	14%	8%	86%	60.7%	7.6%
XGB_NearMiss	52%	12%	7%	57.1%	50.3%	6.8%
XGB_RandomUnderSampler (default)	51%	10%	5%	43%	49.3%	20.1%
XGB_RandomUnderSampler (100)	47%	9%	5%	43%	55.6%	7.3%

Table 4.7 shows us that for the RandomUnderSampler method can change the results of the models, sometimes better and sometimes worse than the NearMiss method. As the RandomUnderSampler method randomly selects samples from the majority group, this randomness can alter the results and the models become more unreliable as the seed changes. NearMiss, on the other hand, uses heuristics to select samples from the majority group closest to the minority group (we can also call it NearMiss-1 in this case since it is the default version of this method). It is a collection of undersampling methods based on the kNN approach where samples are selected from the majority group with minimum average distance to three closest minority group samples. The main issue with undersampling, regardless of the method that is chosen, is that potentially useful data is deleted from the majority group which can be critical to fitting

a more robust decision boundary for the models. In comparison to NearMiss, RandomUnderSampler does this process randomly without detecting and preserving samples which could be more informative than the others. Thus, leaving us with the better option of NearMiss method. Finally, it is important to mention that since we are not adding any synthetic sample to our dataset like the oversampling method, there is no potential data leakage during the training process. Losing potentially informative samples that could help the models decide on the decision boundaries between the groups is a tradeoff at this point and NearMiss is a better approach to handle the decision boundary as it does not rely on randomness.

4.3.2. Comparison of All Models

After the pilot experiments were finalized, we have decided to undersample our initial dataset using the NearMiss method, resulting in a cohort of 46 samples in total. Afterwards, the dataset was split into training and test sets using 70/30 ratio. Then the training set was split into validation and test sets using LOOCV and hyperparameters were tuned. As the imbalance problem was solved, there was a significant improvement in the results for all of the models.

Table 4.8. Results of ML models with all features.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC	PR-AUC
LR	79%	82%	70%	100%	95.9%	96.2%
RF	79%	73%	100%	57%	100%	100%
DT	86%	88%	78%	100%	95.9%	93.7%
XGB	79%	73%	100%	57%	93.9%	93.8%

Table 4.8 shows us that all models did very well with the dataset. RF and XGB models with 100% Precision whereas LR and DT models with 100% Recall. Lower Recall values for the first 2 models indicate the increase in False Negatives which means that the models are missing out on actual MDD samples. As a tradeoff, it is better to miss some actual MDD samples than to misclassify Control samples as false

MDD samples like the other models because we are interested in biomarkers related to MDD.

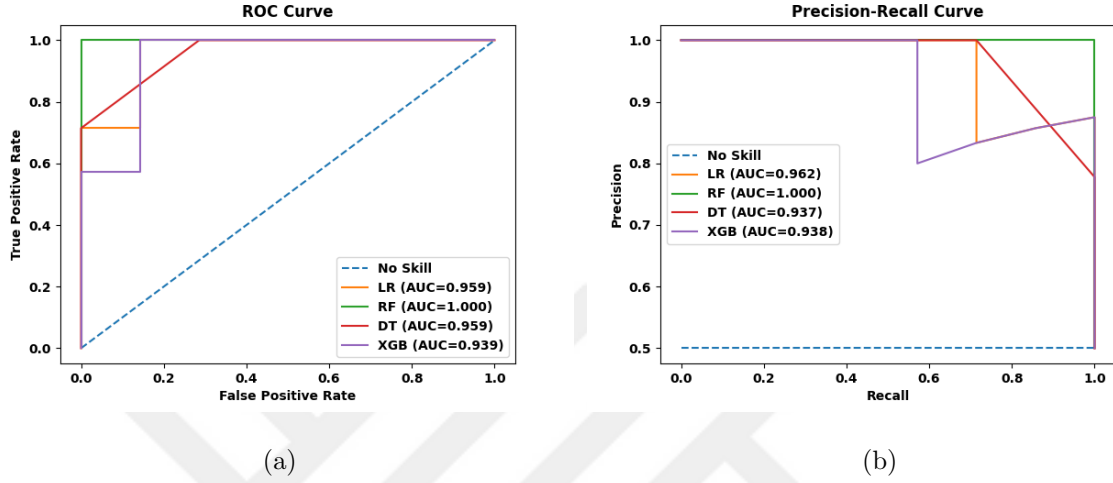


Figure 4.10. Areas under (a) Receiver operating characteristic (ROC-AUC) and (b) Precision-Recall (PR-AUC) curves for models using raw datasets.

Before finding potential biomarkers, as the dataset had too many features/OTUs, we also compared the same models with DAA filtered version where we only kept the features found from the DAA results as given in Table 4.2.

Table 4.9. Results of ML models with DAA filtered features.

Model	Accuracy	F1-score	Precision	Recall	ROC-AUC	PR-AUC
LR	79%	80%	75%	86%	83.7%	90.8%
RF	64%	44%	100%	29%	95.9%	96.8%
DT	50%	0%	0%	0%	64.3%	60.7%
XGB	71%	60%	100%	43%	81.6%	86.4%

Table 4.9 shows us that when the features from DAA results were used, DT performed the worst, being unable to classify MDD samples in the test set. LR performed the best among the models but its lack of interpretability was the problem for further

investigation. RF and XGB models performed well with 100% Precision scores despite their low Recall scores. This decrease in performance could indicate that DAA filtered features might have some missing features that could be used to improve the decision boundaries for classifying the samples but this needs further investigation. The final hyperparameter values after tuning were given in Table 4.10 for model reproducibility.

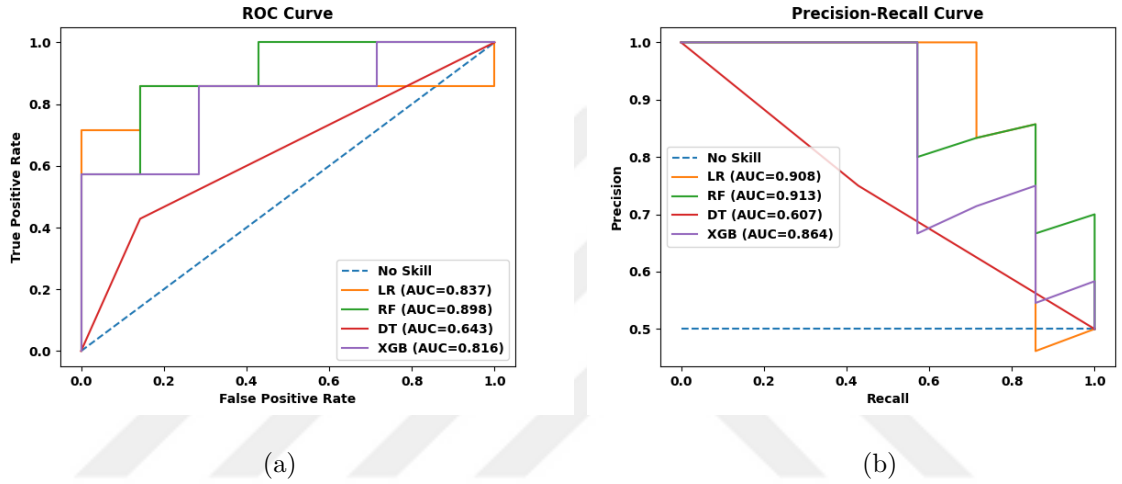


Figure 4.11. Areas under (a) Receiver operating characteristic (ROC-AUC) and (b) Precision-Recall (PR-AUC) curves for models using DAA filtered datasets.

4.3.3. Comparison of Best Two Models

Moving on with the 2 best models (RF and XGB) of our interest, we have gathered the most important features from trained models respectively. Despite LR's good results, it was not chosen as the model interpretability became an issue. These features were then compared between two datasets (raw and DAA filtered) per model and the average importance values (Gini importance for RF and F-score for XGB) were sorted to show potential biomarkers per model. Raw RF found 47 important features, DAA filtered RF found 25 important features and 12 of those features were present in both versions. For the case of XGB, raw XGB found 11 important features, DAA filtered XGB found 10 important features and 4 of those features were present in both versions.

Table 4.10. Tuned hyperparameter values retrieved from both raw and DAA filtered models.

	Raw model hyperparameters	DAA model hyperparameters
LR	penalty='l2' C=0.001 solver='liblinear'	penalty='l2' C=0.001 solver='liblinear'
RF	n_estimators=25 max_depth=3 max_leaf_nodes=3 max_features='sqrt' criterion='gini'	n_estimators=50 max_depth=3 max_leaf_nodes=6 max_features='sqrt' criterion='gini'
DT	max_depth=3 min_samples_leaf=5 criterion='gini'	max_depth=2 min_samples_leaf=5 criterion='gini'
XGB	max_depth=6 min_child_weight=1	max_depth=6 min_child_weight=1

4.3.4. Average of Most Important Features of Random Forest

By their average importance values (Figure 4.12), RF model's top 10 important features were *Ruminococcus*, *Odoribacter*, *Dorea*, *Bacteroides*, *Coprobacillus*, *Prevotella copri*, *Eubacterium dolichum*, *Blautia producta*, *Adlercreutzia*, and *Oscillospira*.

4.3.5. Average of Most Important Features of XGBoost

By their average importance values (Figure 4.13), XGB model's top 10 important features were *Odoribacter*, *Ruminococcus*, *Adlercreutzia*, *Bifidobacterium adolescentis*, *Blautia*, *Bacteroides*, *Parabacteroides distasonis*, *Planomicrobium*, *Oscillospira*, and *Dorea*.

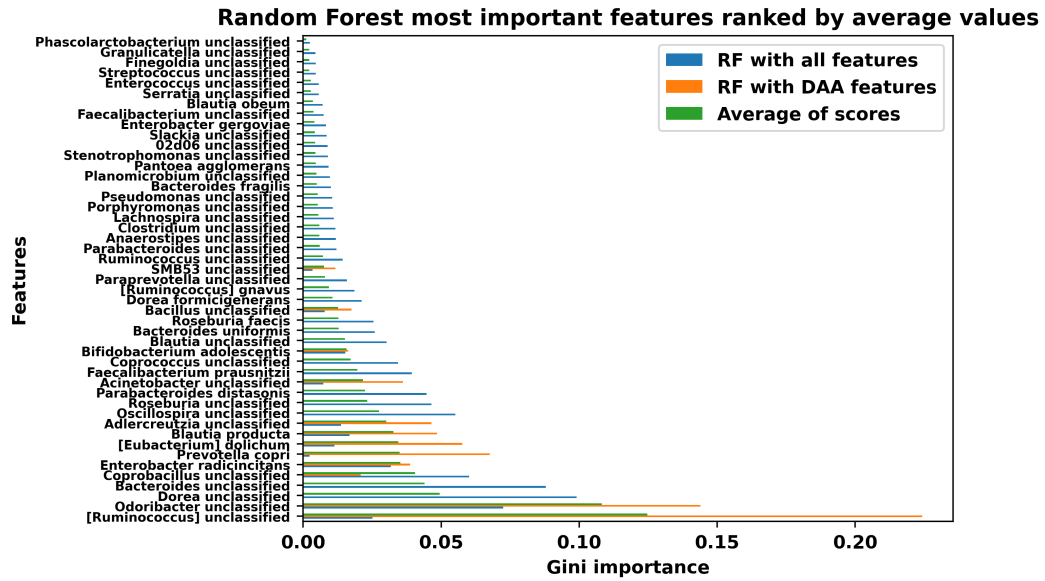


Figure 4.12. Most important features of Random Forest model ranked by their average gini importance values.

4.3.6. SHAP Values of Raw Random Forest Model

As the feature importance values do not give any information about the features' impact on which class they belong to, SHAP values were calculated for each model per dataset for comparison. Each dot on the beeswarm graph in Figure 4.14 represents a sample from the dataset and their color represents their value/abundance. In general, almost all features show a similar pattern where lower values of them have negative SHAP values, meaning when they are higher in abundance, the model is more likely to predict that the sample is from the MDD group. The reverse case is only seen for *Enterobacter radicincitans* where the positive values are clustered, meaning that instances of a higher amount of this feature (red points) have higher negative values which suggests that high abundance of this feature has a stronger negative impact (likelihood of being classified as Control) on class prediction than the positive impact of low abundance of class prediction.

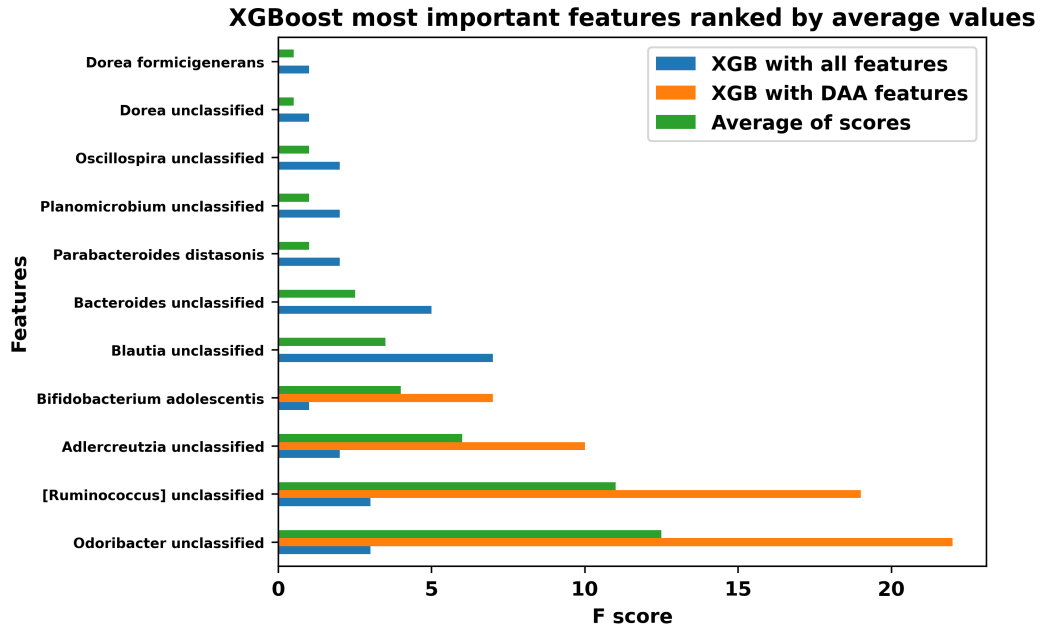


Figure 4.13. Most important features of XGBoost model ranked by their average F-score values.

When we look at the order of SHAP values of features and how they correspond to the respective model's feature importance values by order, we can observe that *Dorea*, *Odoribacter*, *Bacteroides*, *Oscillospira*, and *Coprobacillus* are high in SHAP values whereas *Blautia producta* and *Eubacterium dolichum* seem to be lower in SHAP values, having less impact on model output despite being important features. The remaining most important features from the average of RF models were not found to have impact on raw RF model output.

4.3.7. SHAP Values of DAA Filtered Random Forest Model

For DAA filtered RF model (Figure 4.15), the genus *Bifidobacterium* with different strains of species (*adolescentis*, *bifidum*, and *animalis*) showed different patterns. *Bifidobacterium adolescentis* looked even for both classes but high value samples extended towards MDD. *Bifidobacterium bifidum* and *Bifidobacterium animalis*, on the other hand, showed that increase in abundance impacts the result more negatively than

a small positive outcome, meaning they result in having more effect on being classified as Control. The features on the bottom half of the figure showed dense but low clusters of samples with a few samples on the opposite side which could be outliers.

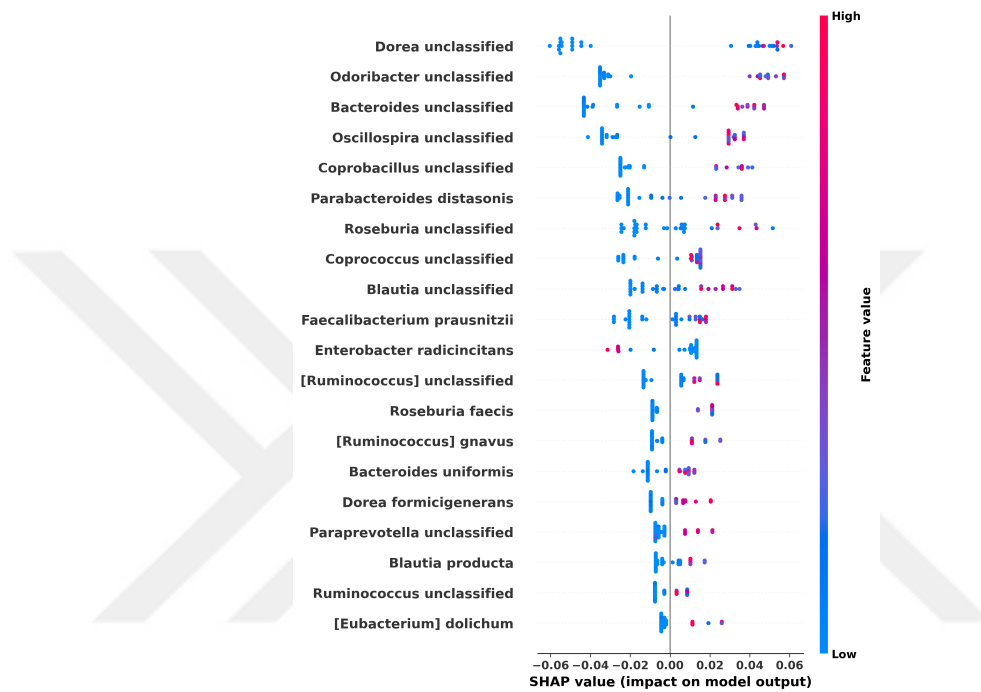


Figure 4.14. SHAP values of raw Random Forest model.

When we look at the order of SHAP values of features and how they correspond to the respective model's feature importance values by order, we can observe that *Bacteroides* and *Adlercreutzia* seem to be lower in SHAP values, having less impact on model output despite being important features. The remaining most important features from the average of RF models were not found to have impact on DAA filtered RF model output.

4.3.8. Potential Biomarkers Retrieved from Random Forest Models

For common features in both RF models, we did not observe majority of them in both of the SHAP figures. The remaining features showed different values for model

impact. This was most likely due to the fact that RF picked up other features as important and they might indeed have an impact on the classification result in terms of their SHAP values.

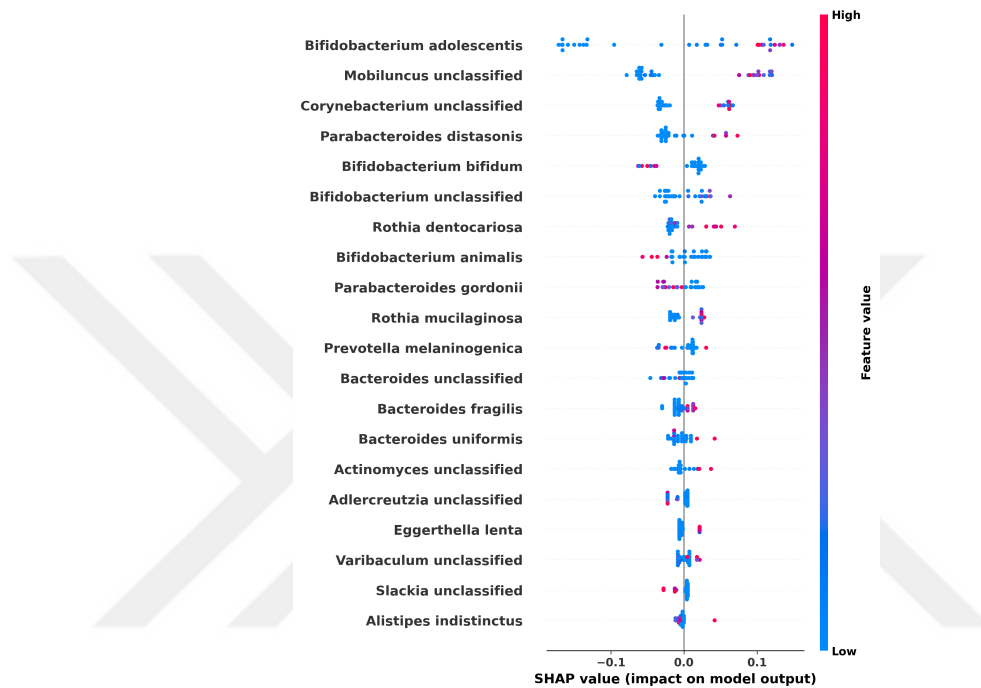


Figure 4.15. SHAP values of DAA filtered Random Forest model.

When we individually observed these figures for those common features who had any impact on the results, we can observe that species *Odoribacter unclassified*, *Coprobacillus unclassified*, *Ruminococcus unclassified*, *Blautia producta*, and *Eubacterium dolichum* increased and species *Enterobacter radicincitans*, *Adlercreutzia unclassified* and *Acinetobacter unclassified* decreased in MDD patients. *Bifidobacterium adolescentis* showed similar patterns for both groups with few abundant samples extending towards the MDD group, indicating that it might be a feature that increases in MDD patients but further investigation is required to solve the uncertainty around the low value samples being dispersed on both sides. This uncertainty might be a direct result of undersampling which might have eliminated more distinctive samples from the cohort and also due to the values of tuned hyperparameters.

4.3.9. SHAP Values of Raw XGBoost Model

For raw XGB model (Figure 4.16), the features *Planomicrobium unclassified* and *Bifidobacterium adolescentis* seemed to result in the classification of Control group when high in abundance and showed a stronger negative impact on the results as the lower values were clustered with low SHAP values. The rest of the features showed a pattern where they all resulted in the classification of MDD group when they were high in abundance. For the bottom half of the group, the features seemed to have no impact on the results.

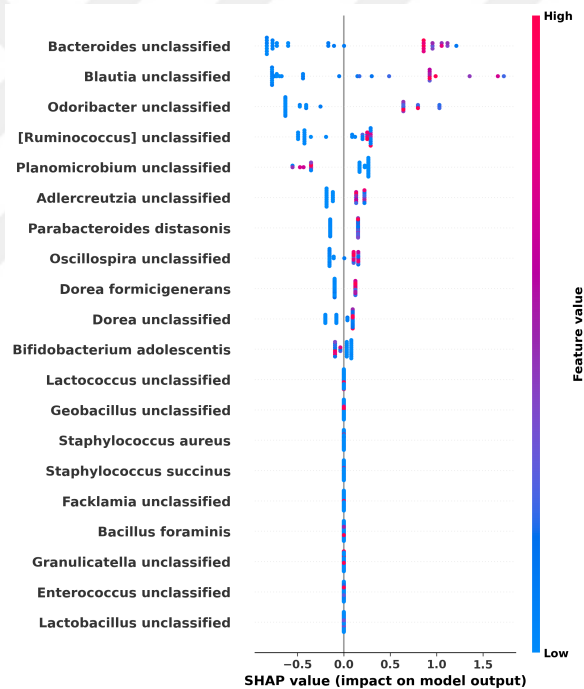


Figure 4.16. SHAP values of raw XGBoost model.

When we look at the order of SHAP values of features and how they correspond to the respective model's feature importance values by order, we can observe that *Odoribacter*, *Ruminococcus*, *Adlercreutzia*, *Coprobacillus*, *Prevotella copri*, and *Blautia producta* showed similar patterns with their feature importance values. The remaining most important features from the average of XGB models were not found to have

impact on raw XGB model output.

4.3.10. SHAP Values of DAA Filtered XGBoost Model

For DAA filtered XGB model (Figure 4.17), *Bifidobacterium adolescentis*, *Enterobacter hormaechei*, *Pseudomonas viridiflava*, *Acinetobacter unclassified*, and *Bifidobacterium unclassified* seemed to result in negative SHAP values (in favor of Control group) when they were high in abundance. The rest of the features showed the opposite pattern and the bottom half of the features showed no impact on the results.

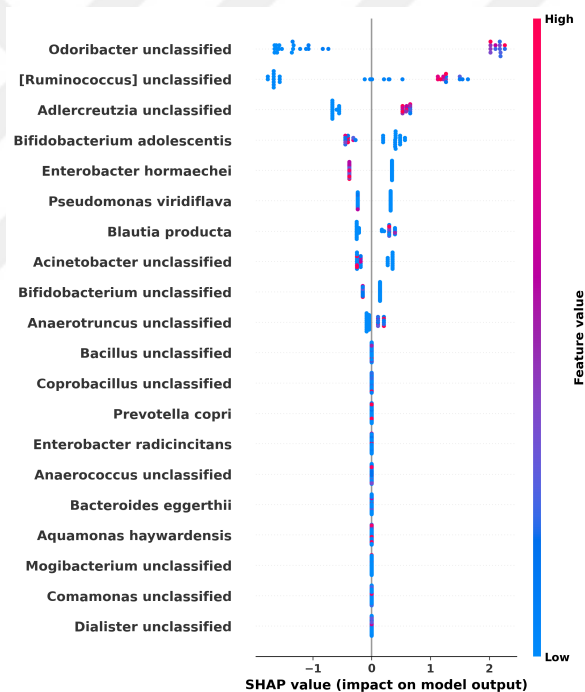


Figure 4.17. SHAP values of DAA filtered XGBoost model.

When we look at the order of SHAP values of features and how they correspond to the respective model's feature importance values by order, we can observe that *Bacteroides*, *Blautia*, *Odoribacter*, *Ruminococcus*, *Planomicrobium*, *Adlercreutzia*, *Bifidobacterium adolescentis*, *Parabacteroides distasonis*, *Oscillospira*, and *Dorea* showed similar patterns with their feature importance values. The remaining most important

features from the average of XGB models were not found to have impact on DAA filtered XGB model output.

4.3.11. Potential Biomarkers Retrieved from XGBoost Models

For common features in both XGB models, upon observing both SHAP figures, it can be seen that species *Odoribacter unclassified*, *Ruminococcus unclassified*, and *Adlercreutzia unclassified* were increased and *Bifidobacterium adolescentis* was decreased in MDD patients.

4.3.12. Overall Potential Biomarkers

Overall, when we compare the common features and their impact for both models, we can come to the conclusion that species *Odoribacter unclassified*, *Ruminococcus unclassified* and *Adlercreutzia unclassified* were increased in MDD patients whereas *Bifidobacterium adolescentis* was decreased but needs further investigation for a more precise result.

5. DISCUSSION

In this study, we present a bioinformatic data analysis on a publicly available 16S rRNA AGP dataset related to the fecal microbiota of MDD patients. This work's main goal is to statistically analyze the structure of the dataset and then identify the features that can be used to differentiate MDD patients. In the dataset, there was a total of 384 samples (23 MDD and 361 HC) and the only statistically significant sample characteristic was found to be the age of the samples. Firstly, low abundant taxa that were not prevalent in the 10% of the samples were filtered and then diversity analysis was performed. A significant difference among the groups was not found in either the alpha or the beta diversity indexes.

Several DAA methods found differing results when it came to finding a potential biomarker for MDD. It should be noted that different methods do not always pick up the same taxons as statistically significant as the choice of analysis and the normalization method can affect the false discovery rate [32], hence why they differ in results. Furthermore, just because a potential biomarker is found significant by multiple DAA methods, this does not mean that it could not be a false positive due to the identical noise being picked up by those methods [43], which requires further methods to prove the significance of that feature.

To further investigate any potential biomarkers, we built ML models that could accurately predict MDD patients. Considering the different nature of ML models, their hyperparameters, the features of the dataset and model complexity in general, we observed different results for different models and features. Before any modelling, we undersampled the dataset for better results. In order to decide on this decision, we first did pilot experiments with the original dataset using 4 different methods. The dataset was split first and then the data was augmented during cross validation in order to tune the hyperparameters. We have observed that undersampled and ratioed models performed the best but instead of relying on synthetic data, we did a tradeoff and

chose undersampling method. After that, we also compared undersampling methods NearMiss and RandomUnderSampler to decide on which is more reliable and observed that randomness alters the results which causes the decision boundaries for the models to change in significant ways. As a result, NearMiss method which uses heuristics to discard samples was chosen to move forward.

After the undersampling of the cohort as a first step, the dataset was then split into training and test sets and same model architectures were used to train the models without any data augmentation during the hyperparameter tuning as the groups were balanced. Same procedure was done with a DAA filtered version of the dataset where only features found from DAA methods were used during training process. After training raw and DAA filtered versions of all models, we chose the two best performing models for comparison which were Random Forest (RF) and XGBoost (XGB) and compared the raw dataset with DAA filtered dataset in terms of features. LR also performed really well but unlike RF and XGB, it lacked strength in model interpretability. DT, on the other hand, gave poor results. The average feature importance values of these models were plotted and SHAP values were investigated for each case to observe the impact of features on model output.

The overall consensus after observing SHAP values was that unclassified species of *Odoribacter*, *Ruminococcus* and *Adlercreutzia* were increased whereas species *Bifidobacterium adolescentis* was decreased in MDD patients. Even though it needs further investigation due to our choice of modeling, *Bifidobacterium adolescentis* should be considered a potential biomarker because it was also found statistically significant by most amount of DAA methods and again showed a decrease of abundance in MDD patients.

The genus *Bifidobacterium* is a gram-positive acetate and lactate producing bacteria from the phylum *Actinobacteria* which plays a cross-feeding role in butyrate production by carbohydrate fermentation and it is one of the most used and effective type of probiotics and prebiotics. The abundance of *B. adolescentis* in the colon microbiota varies by age, showing a decrease as the host ages. The genus *Bifidobacterium*

takes a role in the development of infant immunity and has benefits such as antibacterial features against pathogens, improving intestinal barrier, helping host immune system, and anti-inflammatory properties [56]. It is also responsible for producing gamma-aminobutyric acid (GABA) and vitamin syntheses such as riboflavin, niacin and folate [22]. Several reviews and studies also found similar results, indicating a decrease in genus *Bifidobacterium* in MDD patients [22], [57,58], although there is one study that found the opposite [15].

For the genera *Ruminococcus*, *Odoribacter* and *Adlercreutzia*, there are inconsistent results in the literature. The genus *Ruminococcus* from the phylum *Firmicutes* is a gram-positive anaerobic bacteria and despite our findings, several studies have found that it was declined in MDD patients [8], [14], [19], [57–59] whereas one study found that it was increased in MDD patients [60]. *Ruminococcus* species are involved in the breakdown of complex carbohydrates and dietary fibers. *Ruminococcus* breakdown cellulose (with the formation of methane), accumulates a reserve iodophilic polymer of glucose in the cytoplasm. They are more abundant in diets rich in plant-based fibers and can be influenced by the type and amount of fiber consumed. A high-fiber diet tends to increase the abundance of *Ruminococcus*. Some species of *Ruminococcus* are associated with Crohn’s disease and with inflammatory bowel disease.

The genus *Odoribacter* from the phylum *Bacteroidetes* is a gram-negative atypical pathogenic bacteria. One study showed that it declined in MDD patients [61] whereas other studies found the opposite results [62,63], in favor of our findings. This genus is often associated with a diet high in fiber. Fiber-rich foods, such as fruits, vegetables, and whole grains, can promote the growth of *Odoribacter* species. They play a role in fermenting dietary fibers and producing short-chain fatty acids (SCFAs), which are beneficial for gut health. The microbiota’s gram-negative bacterial abundance may be elevated by a Western diet and may serve as a possible marker for MDD [2].

The genus *Adlercreutzia* from the phylum *Actinomycetota* is gram-positive bacteria. There are not enough findings about this genus in the literature regarding its

relationship to MDD compared to our findings except for a study where it was found that it declined in MDD patients [64] whereas another study found the opposite increased [65]. This genus is less well-studied but is generally associated with a diet rich in fiber and prebiotics. High-fiber diets and certain fermented foods can support the growth of *Adlercreutzia*. The relationship between diet and the abundances of specific gut bacteria like *Odoribacter*, *Ruminococcus*, and *Adlercreutzia* is an area of research to be further explored. Overall, a diet rich in dietary fiber and prebiotics tends to support the growth of beneficial gut bacteria. Conversely, a diet high in processed foods and low in fiber can reduce their abundance. The specific effects can vary based on individual differences, such as baseline gut microbiota composition and genetic factors.

As mentioned before, due to the heterogeneous nature of MDD, external factors like sex, diet and stress play a role in the pathophysiology of MDD. When it comes to considering the differences in sex, MDD was observed two times more in women than men and showed differences in the abundance of certain bacterial compositions. For instance, one study showed that men had a higher abundance of *Bacteroides* and *Prevotella* compared to women [3]. One reason behind the impact of sex in terms of gut microbial diversity and abundance could be the bidirectional interaction between the gut microbiota and estrogen levels in the human body. The metabolism of estrogen in the enterohepatic circulation is one of the examples behind this interaction. Furthermore, several experiments on rodents showed that bilateral ovariectomy may cause dysbiosis in the gut, increasing the abundance of *Clostridium*, further proving the relationship between the gut microbiome and the sex hormones, specifically estrogen [66]. Despite the differences between MDD and Control groups caused by sex, it seems that diet and antibiotic use influence the gut microbiota the most [1]. Apart from individual bacteria, *Firmicutes*/*Bacteroidetes* (F/B) ratio has been proposed as a potential biomarker, as observations show a decrease in *Firmicutes* and an increase in *Bacteroidetes* in MDD patients, it has been found that F/B ratio was higher in women compared to men [1], [4], [67].

The importance of diet and its effects on bacterial diversity in the gut microbiome have gained attention as changes in dietary habits and weight due to MDD's impact on one's appetite is one of the observed symptoms [15], [68]. Whether the microbial composition of MDD patients is a direct result of dietary habits or related to a dysfunctional glucose metabolism is unclear but the latter is implicated in the etiology of MDD. This relationship between gut microbiota diversity and metabolic pathways is not surprising as it has been revealed that the consumption of carbohydrates has increased in MDD patients with respect to their counterparts [16].

Even though scientific findings regarding the relationship between diet and mental health have not been entirely consistent, differences between Eastern and Western diets may still affect the gut microbiome composition and diversity. Correlations have been found between lower observations of MDD and relatively healthier dietary patterns from Norway, Japan and Mediterranean regions, which have well-balanced amounts of essential nutrients such as vegetables, fruits, nuts, grains, legumes and seafood [21]. A plant-based diet, specifically, has been found to improve depressive symptoms [7]. Western diet, on the other hand, is known for its lower amount of fruits and vegetables, and higher amounts of refined, fried and processed meals with high-fat dairy products which are linked to MDD as they increase inflammation in the body [21].

Short-chain fatty acids (SCFAs) play a key role in gastrointestinal health as their decrease in abundance and metabolic production was found to be one of the major characteristics of gut dysbiosis in various disorders including MDD [69]. They are beneficial bacterial metabolites produced by gut microbes when indigestible dietary fibers such as starch, products of starch degradation that were not absorbed before, oat, wheat bran, cellulose, and pectin [70] are fermented in the large intestine and the most abundant ones found in the colon come in three major forms, i.e., acetate, propionate, and butyrate with an approximate molar proportion of 60:20:20 respectively [69], [71, 72]. SCFAs have neuroactive properties [69] and they can stimulate the CNS [4]. They target endocrine cells, the immune system and neurons. They likely mediate gut and brain interactions through multiple pathways, such as immune signals, hormones and

the vagus nerve [4].

SCFAs can also stimulate the release the synthesis of tryptophan hydroxylase (5-HT) also known as serotonin in the gut from intestinal enterochromaffin cells [2], [4]. It is released from neurons in the enteric nervous system to modulate motility and appears to be an important mediator of the gut-brain axis [4]. Serotonin is produced and secreted in multiple places in the body but more than 90% of it is present in the enterochromaffin cells inside the gastrointestinal tract and gut lumen is responsible for the release upon stimulation in order to regulate the gut motility and perception. An increase in the secretion of serotonin in the gastrointestinal tract, however, may reduce the tryptophan levels in the blood circulation causing a reduction of serotonin synthesis in the brain [73].

They are the main energy source of colonocytes. Butyrate, specifically, is used by colonocytes to generate energy to increase the epithelial oxygen consumption, maintaining an anaerobic environment in the gut to prevent the colonization of aerobic pathogens and regulate the barrier function of colonic epithelium [71], [74]. Butyrate also affects the glucose homeostasis, lipid metabolism and gut inflammation. It exerts anti-inflammatory, anti-oxidant, anti-obesity and metabolic regulation effects [72]. Butyrate plays a key role as a metabolite because several animal studies have shown that butyrate-producing bacteria are beneficial when treating depressive symptoms as they also increase brain-derived neurotrophic factors (BDNF). BDNF are important because they play a vital role in the survival, maintenance, differentiation, and synaptic plasticity of nerve cells in the brain and dysfunction of neurogenesis and neuroplasticity in the levels of BDNF have been proposed to be one of the reasons for MDD [73].

Even though the majority of the butyrate is synthesized from carbohydrate metabolism by butyrate-producing bacteria via butyryl-CoA, acetate-producing bacteria can generate butyrate from acetate via acetyl-CoA, allowing a cross-feeding relationship between the two bacteria types. One example would be the acetate-producing genus *Bifidobacterium* which has antidepressant-like properties as it can regulate sero-

tonin and BDNF levels in the brain [45], [69].

The correlation between SCFAs and the abundances of specific gut bacteria like *Odoribacter*, *Ruminococcus*, and *Adlercreutzia* is also significant because these bacteria play key roles in SCFA production. *Ruminococcus* species are significant contributors to SCFA production, particularly acetate and butyrate. They break down complex carbohydrates into simpler forms that can be fermented by other bacteria into SCFAs. A positive correlation exists between *Ruminococcus* abundance and SCFA levels, especially in diets rich in resistant starches and fibers. These bacteria are integral in initiating the fermentation process that leads to SCFA production. *Odoribacter* species are also known producers of SCFAs, particularly butyrate. An increase in *Odoribacter* abundance is often correlated with higher levels of SCFAs, especially butyrate. This relationship is commonly observed in individuals consuming a high-fiber diet, as fiber is the primary substrate for SCFA production. *Adlercreutzia* is less well-known for SCFA production directly, it contributes to the overall microbial ecosystem that supports SCFA production. *Adlercreutzia* may influence the production of SCFAs indirectly by interacting with other bacteria and participating in the complex microbial networks of the gut. In short, these bacteria *Odoribacter*, *Ruminococcus*, and *Adlercreutzia* either directly produce SCFAs or support a microbial environment conducive to SCFA production, particularly when the diet is rich in dietary fibers. Mixed findings in literature; changes in abundance (either increase or decrease) are associated with MDD, potentially impacting SCFA production and gut health.

It is noteworthy to mention that results can vary depending on various factors such as the chosen filtering criteria, scaling methods, and methods of analysis. The dataset that was used in our study was retrieved from Qiita and it was preprocessed using an older version of QIIME and GreenGenes. Furthermore, AGP data needs a bloom analysis to eliminate bacteria that were higher in abundance due to blooming during the sample collection process. Considering the original size of the sample collection, the initial preprocessing step took in different places in different times so a batch correction step is required (or at least must be done to observe whether the variance

was due to biological or technical reasons). Moreover, the high dimensional structure of microbiome data with high amount of zeros made it harder for machine learning modeling to obtain good results which is why undersampling was necessary to obtain good results. This step, however, resulted in loss of data which might explain some of the inconsistencies in the results but still not enough to explain all of them, which is one of the remaining problems that scientists are trying to tackle. A better metadata with a machine-readable vocabulary could be used to implement more covariates for statistical analyses and machine learning modeling as external factors also play a role in understanding microbiota.



6. CONCLUSION

Major Depressive Disorder (MDD) has been increasingly linked to gut microbiota composition, as the gut-brain axis plays a crucial role in mental health. This study analyzed the AGP data consisting of 361 control patients and 23 MDD patients. The cohort did not show any statistical significance in terms of cohort characteristics or diversity indexes except for the patients' age. After deciding on which data augmentation method to use moving forward during our pilot experiments (we undersampled the cohort to 46 samples in total), using the raw dataset and dataset with only the features from our differential abundance analysis, we have built various machine learning models and by using our best two interpretable models which were Random Forest and XGBoost, we have found various important potential biomarkers by observing their SHAP values. Random Forest found taxa *Ruminococcus*, *Odoribacter*, *Dorea*, *Bacteroides*, *Coprobacillus*, *Prevotella copri*, *Eubacterium dolichum*, *Blautia producta*, *Adlercreutzia*, and *Oscillospira* whereas XGBoost found taxa *Odoribacter*, *Ruminococcus*, *Adlercreutzia*, *Bifidobacterium adolescentis*, *Blautia*, *Bacteroides*, *Parabacteroides distasonis*, *Planomicrobium*, *Oscillospira*, and *Dorea* to be important. By checking their intersection, we have concluded that despite some inconsistent results in the literature, species *Bifidobacterium adolescentis* and genera *Odoribacter*, *Ruminococcus*, and *Adlercreutzia* could be potential biomarkers for MDD where our models showed that *B. adolescentis* decreased and the rest increased in MDD patients' gut. Alterations in these bacteria may reflect broader changes in the gut microbiota associated with MDD, affecting gut health, inflammation, and the gut-brain axis.

Research on *Odoribacter* with MDD is still limited. However, some studies suggest that lower/higher abundances of SCFA-producing bacteria, including *Odoribacter*, may be associated with depressive symptoms. This is because SCFAs like butyrate have anti-inflammatory properties and are involved in maintaining gut barrier integrity, both of which can influence brain health. The relationship between *Ruminococcus* and MDD is complex, as some studies have found an increase in certain *Ruminococcus* species in

individuals with MDD, while others have found a decrease. The specific species of *Ruminococcus* might be important, as some are beneficial, while others could be associated with negative health outcomes when dysregulated. *Adlercreutzia* is less studied in the context of MDD, but there is evidence to suggest that its abundance might be altered in individuals with depressive disorders. This genus is involved in the metabolism of phytoestrogens and could influence hormonal pathways linked to mood regulation. A disruption in *Adlercreutzia* could, therefore, potentially impact mental health. More research is needed to fully understand these relationships and their implications for MDD.



7. RECOMMENDATIONS

As mentioned in the final part of the Discussion section, AGP data requires a bloom analysis to eliminate any bacteria that bloomed during the sample collection phase. Assuming that this is done to the raw data, the bioinformatics pipeline should be applied using the newest version of the tool QIIME (or an alternative method, like the tool “mothur”) and the reference database GreenGenes. Then, a batch correction procedure must be applied or it should be taken as a covariate for statistical analyses and machine learning modeling. During the machine learning modeling process, meta-data should be utilized to better understand the model relationship between features and classification results since correlations are expected in a microbiome environment. Instead of focusing on a single dataset, multiple datasets can be analyzed together where each dataset is treated as the batch that was categorized into for comparison. This could also result in a large, possibly balanced dataset so that a data augmentation method will not be required. Gut microbiome analysis can be very subjective and analyses and modeling should be reproducible for other datasets.

REFERENCES

1. Medina-Rodríguez, E. M., A. A. Cruz, J. C. D. Abreu and E. Beurel, “Stress, Inflammation, Microbiome and Depression”, *Pharmacology, Biochemistry and Behavior*, Vol. 227-228, p. 173561, 2023.
2. Barandouzi, Z. A., A. Starkweather, W. A. Henderson, A. Gyamfi and X. Cong, “Altered Composition of Gut Microbiota in Depression: A Systematic Review.”, *Frontiers in Psychiatry*, Vol. 11, p. 541, 2020.
3. Hu, X., Y. Li, J. Wu, H. Zhang, Y. Huang, X. Tan, W. Li, X. Zhou, P. Xie, O. I. Olasunkanmi, J. Zhou, Z. Sun, M. Liu, G. Zhang, J. Yang, P. Zheng and P. Xie, “Changes of Gut Microbiota Reflect the Severity of Major Depressive Disorder: A Cross Sectional Study”, *Translational Psychiatry*, Vol. 13, No. 1, p. 137, 2023.
4. Modesto-Lowe, V., M. Chaplin and D. Sgambato, “Major Depressive Disorder and the Gut Microbiome: What Is the Link?”, *General Psychiatry*, Vol. 36, No. 1, p. e100973, 2023.
5. Kunugi, H., “Gut Microbiota and Pathophysiology of Depressive Disorder.”, *Annals of Nutrition and Metabolism*, Vol. 77, No. 2, Suppl., pp. 11–20, 2021.
6. Chang, L., Y. Wei and K. Hashimoto, “Brain-Gut-Microbiota Axis in Depression: A Historical Overview and Future Directions.”, *Brain Research Bulletin*, Vol. 182, pp. 44–56, 2022.
7. Dong, Z., X. Shen, Y. Hao, J. Li, H. Xu, L. Yin and W. Kuang, “Gut Microbiome: A Potential Indicator for Predicting Treatment Outcomes in Major Depressive Disorder”, *Frontiers in Neuroscience*, Vol. 16, p. 813075, 2022.
8. Chen, M., C.-R. Xie, Y.-Z. Shi, T.-C. Tang and H. Zheng, “Gut Microbiota and Major Depressive Disorder: A Bidirectional Mendelian Randomization”, *Journal*

of *Affective Disorders*, Vol. 316, pp. 187–193, 2022.

9. Han, K., Q. Xie, L. Liu, X. Wu, L. He, Y. Shi, R. Zhang, G. He, Z. Dong and T. Yu, “Different Roles of Microbiota and Genetics in the Prediction of Treatment Response in Major Depressive Disorder”, *Journal of Psychiatric Research*, Vol. 161, pp. 402–411, 2023.
10. Vellingiri, B., S. Y. Aishwarya, S. B. Jancy, G. S. Abhishek, H. W. S. Babu, P. Vijayakumar, A. Narayanasamy, S. Mariappan, R. Sangeetha, A. V. Gopalakrishnan, R. Parthasarathi and M. Iyer, “An Anxious Relationship Between Autism Spectrum Disorder and Gut Microbiota: A Tangled Chemistry?”, *Journal of Clinical Neuroscience*, Vol. 99, pp. 169–189, 2022.
11. Angelova, I. Y., A. S. Kovtun, O. V. Averina, T. A. Koshenko and V. N. Danilenko, “Unveiling the Connection Between Microbiota and Depressive Disorder Through Machine Learning”, *International Journal of Molecular Sciences*, Vol. 24, No. 22, p. 16459, 2023.
12. Chervenkov, T., “Genus-Level Analysis of Gut Microbiota in Children with Autism Spectrum Disorder: A Mini Review”, *Children*, Vol. 10, No. 7, p. 1103, 2023.
13. Agarwala, S., B. Naik, B. Naik and N. B. Ramachandra, “Mucosa-Associated Specific Bacterial Species Disrupt the Intestinal Epithelial Barrier in the Autism Phenome.”, *Brain, behavior, immunity - health*, Vol. 15, p. 100269, 2021.
14. Gao, M., J. Wang, P. Liu, H. Tu, R. Zhang, Y. Zhang, N. Sun and K. Zhang, “Gut Microbiota Composition in Depressive Disorder: A Systematic Review, Meta-Analysis, and Meta-Regression”, *Translational Psychiatry*, Vol. 13, No. 1, p. 379, 2023.
15. Knudsen, J. K., C. Bundgaard-Nielsen, S. Hjerrild, R. E. Nielsen, P. Leutscher and S. Sørensen, “Gut Microbiota Variations in Patients Diagnosed with Major

- Depressive Disorder-A Systematic Review”, *Brain and behavior*, Vol. 11, No. 7, p. e02177, 2021.
16. Chung, Y.-C. E., H.-C. Chen, H. L. Chou, I. Chen, M.-S. Lee, L.-C. Chuang, Y.-W. Liu, M.-L. Lu, C.-H. Chen, C.-S. Wu, M.-C. Huang, S. Liao, Y. Ni, M. Lai, W.-L. Shih and P.-H. Kuo, “Exploration of Microbiota Targets for Major Depressive Disorder and Mood Related Traits.”, *Journal of Psychiatric Research*, Vol. 111, pp. 74–82, 2019.
 17. Irum, N., T. Afzal, M. M. A. Faraz, Z. Aslam and F. Rasheed, “The Role of Gut Microbiota in Depression: An Analysis of the Gut-Brain Axis”, *Frontiers in Behavioral Neuroscience*, Vol. 17, p. 1185522, 2023.
 18. Yang, Z., J. Li, X. Gui, X. Shi, Z. Bao, H. Han and M. D. Li, “Updated Review of Research on the Gut Microbiota and Their Relation to Depression in Animals and Human Beings.”, *Molecular Psychiatry*, Vol. 25, No. 11, pp. 2759–2772, 2020.
 19. Chen, M.-M., P. Wang, X. Xie, Z. Nie, S.-X. Xu, N. Zhang, S. Wang, L. Yao and Z. Liu, “Young Adults with Major Depression Show Altered Microbiome”, *Neuroscience*, Vol. 522, pp. 23–32, 2023.
 20. Marx, W., B. W. Penninx, M. Solmi, T. A. Furukawa, J. Firth, A. F. Carvalho and M. Berk, “Major Depressive Disorder”, *Nature Reviews Disease Primers*, Vol. 9, No. 1, p. 44, 2023.
 21. Kumar, A., J. Pramanik, N. Goyal, D. Chauhan, B. S. Sivamaruthi, B. G. Prajapati and C. Chaiyasut, “Gut Microbiota in Anxiety and Depression: Unveiling the Relationships and Management Options”, *Pharmaceuticals*, Vol. 16, No. 4, p. 565, 2023.
 22. Schrod, C., A. Mahavni, G. P. McNamara, M. D. Tallman, B. T. Brugger, L. Schwarz and A. Bhattacharyya, “The Gut Microbiome and Depression: A Re-

- view”, *Nutritional Neuroscience*, Vol. 26, No. 10, pp. 953–959, 2023.
23. Papoutsoglou, G., S. Tarazona, M. B. Lopes, T. Klammersteiner, E. Ibrahim, J. Eckenberger, P. Novielli, A. Tonda, A. Simeon, R. Shigdel, S. Béreux, G. Vitali, S. Tangaro, L. Lahti, A. Temko, M. Claesson and M. Berland, “Machine Learning Approaches in Microbiome Research: Challenges and Best Practices”, *Frontiers in Microbiology*, Vol. 14, p. 1261889, 2023.
 24. Mayeux, R., “Biomarkers: Potential Uses and Limitations.”, *Neurorx*, Vol. 1, No. 2, pp. 182–188, 2004.
 25. Cheung, S. G., A. R. Goldenthal, A. Uhlemann, J. J. Mann, J. M. Miller and M. E. Sublette, “Systematic Review of Gut Microbiota and Major Depression.”, *Frontiers in Psychiatry*, Vol. 10, p. 34, 2019.
 26. Chen, J.-J., P. Zheng, Y.-Y. Liu, X.-G. Zhong, H.-Y. Wang, Y.-J. Guo and P. Xie, “Sex Differences in Gut Microbiota in Patients with Major Depressive Disorder.”, *Neuropsychiatric Disease and Treatment*, Vol. 14, pp. 647–655, 2018.
 27. Martin, S. E., C. S. Kraft, T. R. Ziegler, E. C. Millson, L. Rishishwar and G. S. Martin, “The Role of Diet on the Gut Microbiome, Mood and Happiness”, MedRxiv, 2023.03.18.23287442, 2023.
 28. O’Neill, S., M. Minehan, C. R. Knight-Agarwal and D. B. Pyne, “Alterations in Gut Microbiota Caused by Major Depressive Disorder or a Low FODMAP Diet and Where They Overlap”, *Frontiers in Nutrition*, Vol. 10, p. 1303405, 2024.
 29. Marcos-Zambrano, L. J., K. Karaduzovic-Hadziabdic, T. L. Turukalo, P. Przyamus, V. Trajkovik, O. Aasmets, M. Berland, A. Gruca, J. Hasic, K. Hron, T. Klammersteiner, M. Kolev, L. Lahti, M. B. Lopes, V. Moreno, I. Naskinova, E. Org, I. Paciência, G. Papoutsoglou, R. Shigdel, B. Stres, B. Vilne, M. Yousef, E. Zdravevski, I. Tsamardinos, E. C. de Santa Pau, M. J. Claesson, I. Moreno-

- Indias and J. Truu, “Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment.”, *Frontiers in Microbiology*, Vol. 12, p. 634511, 2021.
30. Weiss, S., Z. Z. Xu, S. D. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. A. Lozupone, J. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, E. R. Hyde and R. Knight, “Normalization and Microbial Differential Abundance Strategies Depend Upon Data Characteristics”, *Microbiome*, Vol. 5, No. 1, p. 27, 2017.
 31. Qian, X., T. Chen, T. Chen, Y.-P. Xu, L. Chen, F.-X. Sun, M. Lu, M. Lu and Y.-X. Liu, “A Guide to Human Microbiome Research: Study Design, Sample Collection, and Bioinformatics Analysis”, *Chinese Medical Journal*, Vol. 133, No. 15, pp. 1844–1855, 2020.
 32. Gloor, G. B., M. P. Nixon and J. D. Silverman, “Normalizations are not What You Think; Explicit Scale Simulation in ALDEx2”, *BioRxiv*, 2023.10.21.563431v1, 2023.
 33. McMurdie, P. J. and S. Holmes, “Waste Not, Want Not: Why Rarefying Microbiome Data is Inadmissible.”, *PLOS Computational Biology*, Vol. 10, No. 4, p. e1003531, 2014.
 34. Li, P., H. Luo, B. Ji and J. Nielsen, “Machine Learning for Data Integration in Human Gut Microbiome”, *Microbial Cell Factories*, Vol. 21, No. 1, p. 241, 2022.
 35. Medina, R. H., S. Kutuzova, K. N. Nielsen, J. Johansen, L. H. Hansen, M. Nielsen and S. Rasmussen, “Machine Learning and Deep Learning Applications in Microbiome Research”, *ISME Communications*, Vol. 2, No. 1, p. 98, 2022.
 36. Lundberg, S. M. and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions”, *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, edited by Von Luxburg, U., I. Guyon, S. Bengio,

- H. Wallach, R. Fergus, R. Garnett, S. V. N. Vishwanathan, Vol. 30, Long Beach, USA, p. 4768-4777, 2017.
37. Chang, C.-C., T.-C. Liu, C.-J. Lu, H.-C. Chiu and W.-N. Lin, “Machine Learning Strategy for Identifying Altered Gut Microbiomes for Diagnostic Screening in Myasthenia Gravis”, *Frontiers in Microbiology*, Vol. 14, p. 1227300, 2023.
 38. Davis, J. and M. Goadrich, “The Relationship Between Precision-Recall and ROC Curves”, *Proceedings of the 23rd International Conference on Machine Learning*, New York, USA, p. 233–240, 2006.
 39. McDonald, D., E. R. Hyde, J. W. Debelius, J. W. Debelius, J. T. Morton, A. Gonzalez, G. Ackermann, A. A. Aksenov, B. Behsaz, C. Brennan, Y. Chen, L. D. Goldasich, P. C. Dorrestein, R. R. Dunn, A. K. Fahimipour, M. Embree, J. Gaffney, J. A. Gilbert, G. Gogul, J. L. Green, P. Hugenholtz, G. Humphrey, C. Huttenhower, M. A. Jackson, S. Janssen, D. V. Jeste, L. Natarajan, L. Jiang, S. T. Kelley, D. Knight, T. Kościółek, J. Ladau, J. Leach, C. Marotz, D. Meleshko, A. V. Melnik, J. L. Metcalf, H. Mohimani, E. Montassier, J. A. Navas-Molina, T. T. Nguyen, S. D. Peddada, P. A. Pevzner, K. S. Pollard, G. Rahnavard, A. Robbins-Pianka, N. Sangwan, J. Shorenstein, L. Smarr, S. J. Song, T. D. Spector, A. D. Swafford, V. G. Thackray, L. R. Thompson, A. Tripathi, Y. Vázquez-Baeza, A. Vrbanc, P. E. Wischmeyer, E. Wolfe, Q. Zhu and R. Knight, “American Gut: an Open Platform for Citizen-Science Microbiome Research”, *mSystems*, Vol. 3, No. 3, pp. e00031–18, 2018.
 40. Edgar, R. C., “Updating the 97% Identity Threshold for 16S Ribosomal RNA OTUs”, *Bioinformatics*, Vol. 34, No. 14, pp. 2371–2375, 2017.
 41. Oksanen, J., G. L. Simpson, F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. O’Hara, P. Solymos, M. H. H. Stevens, E. Szoecs, H. Wagner, M. Barbour, M. Bedward, B. Bolker, D. Borcard, G. Carvalho, M. Chirico, M. De Caceres, S. Durand, H. B. A. Evangelista, R. FitzJohn, M. Friendly, B. Furneaux, G. Han-

- nigan, M. O. Hill, L. Lahti, D. McGlinn, M.-H. Ouellette, E. Ribeiro Cunha, T. Smith, A. Stier, C. J. Ter Braak and J. Weedon, “Vegan: Community Ecology Package”, <https://CRAN.R-project.org/package=vegan>, 2024, accessed on December 19, 2024.
42. Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn and J. J. Egozcue, “Microbiome Datasets Are Compositional: And This Is Not Optional.”, *Frontiers in Microbiology*, Vol. 8, p. 2224, 2017.
 43. Nearing, J. T., G. M. Douglas, M. G. Hayes, J. MacDonald, D. K. Desai, N. Allward, C. M. A. Jones, R. J. Wright, A. S. Dhanani, A. M. Comeau and M. G. I. Langille, “Microbiome Differential Abundance Methods Produce Different Results Across 38 Datasets”, *Nature Communications*, Vol. 13, No. 1, p. 342, 2022.
 44. Anderson, M. J., *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, New Jersey, 2014.
 45. Yang, L. and J. Chen, “Benchmarking Differential Abundance Analysis Methods for Correlated Microbiome Sequencing Data”, *Briefings in Bioinformatics*, Vol. 24, No. 1, p. bbac607, 2022.
 46. Pan, A. Y., “Statistical Analysis of Microbiome Data: The Challenge of Sparsity”, *Current Opinion in Endocrine and Metabolic Research*, Vol. 19, pp. 35–40, 2021.
 47. Love, M. I., W. Huber and S. Anders, “Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2”, *Genome Biology*, Vol. 15, No. 12, p. 550, 2014.
 48. Segata, N., J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett and C. Huttenhower, “Metagenomic Biomarker Discovery and Explanation”, *Genome Biology*, Vol. 12, No. 6, p. R60, 2011.
 49. Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hol-

- lister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn and C. F. Weber, “Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities”, *Applied and Environmental Microbiology*, Vol. 75, No. 23, pp. 7537–7541, 2009.
50. Fernandes, A. D., J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell and G. B. Gloor, “Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-seq, 16S rRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis”, *Microbiome*, Vol. 2, No. 1, p. 15, 2014.
 51. Mallick, H., A. Rahnavard, L. J. McIver, S. Ma, Y. Zhang, L. H. Nguyen, T. L. Tickle, G. Weingart, B. Ren, E. H. Schwager, S. Chatterjee, K. N. Thompson, J. E. Wilkinson, A. Subramanian, Y. Lu, L. Waldron, J. N. Paulson, E. A. Franzosa, H. C. Bravo and C. Huttenhower, “Multivariable Association Discovery in Population-Scale Meta-Omics Studies.”, *PLOS Computational Biology*, Vol. 17, No. 11, p. e1009442, 2021.
 52. Quillet, J.-C., M. Siani-Rose, R. McKee, B. Goldstein, M. Taylor and I. Kurek, “A Machine Learning Approach for Understanding the Metabolomics Response of Children with Autism Spectrum Disorder to Medical Cannabis Treatment”, *Scientific Reports*, Vol. 13, No. 1, p. 13022, 2023.
 53. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-Learn: Machine Learning in Python”, *Journal of Machine Learning Research*, Vol. 12, No. Oct, pp. 2825–2830, 2011.
 54. Lemaitre, G., F. Nogueira and C. K. Aridas, “Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning”, *Journal of Machine Learning Research*, Vol. 18, No. 17, pp. 1–5, 2017.

55. Chawla, N. V., K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique”, *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002.
56. Arruda, H. S., M. V. Geraldi, M. F. Cedran, J. L. Bicas, M. R. Marostica Junior and G. M. Pastore, *Bioactive Food Components Activity in Mechanistic Approach*, Academic Press, Massachusetts, 2022.
57. Sanada, K., S. Nakajima, S. Kurokawa, A. Barceló-Soler, D. Ikuse, A. Hirata, A. Yoshizawa, Y. Tomizawa, M. Salas-Valero, D. M. Blumberger, Y. Noda, M. Mimura, A. Iwanami and T. Kishimoto, “Gut Microbiota and Major Depressive Disorder: A Systematic Review and Meta-Analysis.”, *Journal of Affective Disorders*, Vol. 266, pp. 1–13, 2020.
58. Liang, S., X. Wu, X. Hu, T. Wang and F. Jin, “Recognizing Depression from the Microbiota–Gut–Brain Axis”, *International Journal of Molecular Sciences*, Vol. 19, No. 6, p. 1592, 2018.
59. Maes, M., A. Vasupanrajit, K. Jirakran, P. Klomkiew, P. Chanchaem, C. Tunvirachaisakul and S. Payungporn, “Exploration of the Gut Microbiome in Thai Patients with Major Depressive Disorder Shows a Specific Bacterial Profile with Depletion of the Ruminococcus Genus as a Putative Biomarker”, *Cells*, Vol. 12, No. 9, p. 1240, 2023.
60. Zhao, H., K. Jin, C. Jiang, F. Pan, J. Wu, H. Luan, Z. Zhao, J. Chen, T. Mou, Z. Wang, J. Lu, S. Lu, S. Hu, Y. Xu and M. Huang, “A Pilot Exploration of Multi-Omics Research of Gut Microbiome in Major Depressive Disorders”, *Translational Psychiatry*, Vol. 12, No. 1, p. 8, 2022.
61. Shen, Y., X. Yang, G. Li, J. Gao and Y. Liang, “The Change of Gut Microbiota in MDD Patients Under SSRIs Treatment.”, *Scientific Reports*, Vol. 11, No. 1, p. 14918, 2021.

62. Liu, P., M. Gao, Z. Liu, Y. Zhang, H. Tu, L. Lei, P. Wu, A. Zhang, C. Yang, G. Li, N. Sun and K. Zhang, "Gut Microbiome Composition Linked to Inflammatory Factors and Cognitive Functions in First-Episode, Drug-Naive Major Depressive Disorder Patients", *Frontiers in Neuroscience*, Vol. 15, p. 800764, 2022.
63. Zhong, Q., J.-J. Chen, Y. Wang, W.-H. Shao, C.-J. Zhou and P. Xie, "Differential Gut Microbiota Compositions Related With the Severity of Major Depressive Disorder", *Frontiers in Cellular and Infection Microbiology*, Vol. 12, p. 907239, 2022.
64. Ghannoum, M. A., M. Ford, R. A. Bonomo, A. Gamal and T. S. McCormick, "A Microbiome-Driven Approach to Combating Depression During the COVID-19 Pandemic.", *Frontiers in Nutrition*, Vol. 8, p. 672390, 2021.
65. Liu, P., Z. Liu, J. Wang, J. Wang, M. Gao, Y. Zhang, C. Yang, A. Zhang, G. zhi Li, X. Li, S. Liu, L. Liu, N. Sun and K. Zhang, "Immunoregulatory Role of the Gut Microbiota in Inflammatory Depression", *Nature Communications*, Vol. 15, No. 1, p. 3003, 2024.
66. Shobeiri, P., A. Kalantari, A. L. Teixeira and N. Rezaei, "Shedding Light on Biological Sex Differences and Microbiota-Gut-Brain Axis: A Comprehensive Review of Its Roles in Neuropsychiatric Disorders.", *Biology of Sex Differences*, Vol. 13, No. 1, p. 12, 2022.
67. Zhang, Q., Y. Yun, H. An, W. Zhao, T. Ma, Z. Wang and F. Yang, "Gut Microbiome and Daytime Function in Chinese Patients with Major Depressive Disorder", *Journal of Psychosomatic Research*, Vol. 157, p. 110787, 2022.
68. Groen, R. N., N. C. de Clercq, M. Nieuwdorp, H. J. R. Hoenders and A. K. Groen, "Gut Microbiota, Metabolism and Psychopathology: A Critical Review and Novel Perspectives", *Critical Reviews in Clinical Laboratory Sciences*, Vol. 55, No. 4, pp. 283–293, 2018.

69. Cheng, Y., J. Liu and Z. Ling, “Short-Chain Fatty Acids-Producing Probiotics: A Novel Source of Psychobiotics.”, *Critical Reviews in Food Science and Nutrition*, Vol. 62, No. 28, pp. 7929–7959, 2022.
70. Notting, F., W. Pirovano, W. Sybesma and R. Kort, “The Butyrate-Producing and Sporeforming Bacterial Genus *Coprococcus* as a Potential Biomarker for Neurological Disorders”, *Gut Microbiome*, Vol. 4, p. e16, 2023.
71. Singh, V. K., G. Lee, H. Son, H. Koh, E. S. Kim, T. Unno and J. Shin, “Butyrate Producers, “The Sentinel of Gut”: Their Intestinal Significance with and Beyond Butyrate, and Prospective Use as Microbial Therapeutics”, *Frontiers in Microbiology*, Vol. 13, p. 1103836, 2023.
72. Amiri, P., S. A. Hosseini, S. Ghaffari, H. Tutunchi, S. Ghaffari, E. Mosharkesh, S. Asghari and N. Roshanravan, “Role of Butyrate, a Gut Microbiota Derived Metabolite, in Cardiovascular Diseases: A Comprehensive Narrative Review”, *Frontiers in Pharmacology*, Vol. 12, p. 837509, 2022.
73. Suda, K. and K. Matsuda, “How Microbes Affect Depression: Underlying Mechanisms via the Gut–Brain Axis and the Modulating Role of Probiotics”, *International Journal of Molecular Sciences*, Vol. 23, No. 3, p. 1172, 2022.
74. Zhu, L.-B., Y.-C. Zhang, H.-H. Huang and J. Lin, “Prospects for Clinical Applications of Butyrate-Producing Bacteria”, *World Journal of Clinical Pediatrics*, Vol. 10, No. 5, pp. 84–92, 2021.