

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



AI-Driven Metagenomic Analysis for Microbial Biomarker Discovery in Metabolic Syndrome

by

Sahar Safdar

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Health and Life Sciences

Department of Bioinformatics and Biosciences

2026

Copyright © 2026 by Sahar Safdar

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.



CERTIFICATE OF APPROVAL

AI-Driven Metagenomic Analysis for Microbial Biomarker Discovery in Metabolic Syndrome

by

Sahar Safdar

(MBS251003)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Babar Jamal Bacha	NUMS, Rawalpindi
(b)	Internal Examiner	Dr. Sami Ullah Jan	CUST, Islamabad

Prof. Dr. Syeda Marriam Bakhtiar

Thesis Supervisor

August, 2026

Prof. Dr. Syeda Marriam Bakhtiar
Head
Dept. of Bioinfo. & Biosciences
August, 2026

Prof. Dr. Sahar Fazal
Dean
Faculty of Health & Life Sciences
August, 2026

Author's Declaration

I, **Sahar Safdar** hereby state that my MS thesis titled “**AI-Driven Metagenomic Analysis for Microbial Biomarker Discovery in Metabolic Syndrome**” is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.



(Sahar Safdar)

Registration No: MBS251003

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled “**AI-Driven Metagenomic Analysis for Microbial Biomarker Discovery in Metabolic Syndrome**” is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.



(Sanar Sardar)

Registration No: MBS251003

Acknowledgement

First and foremost, I am deeply grateful to **Allah Almighty** for His countless blessings, mercy, and guidance throughout my academic journey. I also send Darood-o-Salam upon the **Holy Prophet Muhammad** , whose life remains the greatest source of guidance, patience, and inspiration.

I would like to express my sincere gratitude to my respected supervisor, **Prof. Dr. Syeda Marriam Bakhtiar**, for her valuable guidance, continuous support, encouragement, and mentorship throughout this research work. Her expertise, patience, and constructive feedback helped me complete this thesis successfully.

I am also thankful to the Department of Bioinformatics and Biosciences, Capital University of Science and Technology, Islamabad, for providing academic support and a learning environment during my MS studies. I would like to acknowledge my teachers for their guidance, knowledge, and encouragement throughout my academic journey.

My deepest gratitude goes to my beloved parents, **Abu** and **Ammi**, for their endless prayers, love, sacrifices, and moral support. Their encouragement has always been my greatest strength. I am also thankful to my brothers, **Salman** and **Usman**, my sister, **Sania**, my teachers, especially **Ma'am Shah Gull** and my friends, especially **Saher**, for their constant support, motivation, and encouragement.

Finally, I would like to thank everyone who supported me directly or indirectly during the completion of this thesis. Their kindness, prayers, and encouragement will always be remembered with gratitude.

(Sahar Safdar)

Abstract

Metabolic diseases are increasingly associated with gut microbiome dysbiosis; however, integrated approaches combining taxonomic profiling, ecological interpretation, statistical testing, and explainable artificial intelligence remain limited. This study applied a comprehensive metagenomic workflow to investigate gut microbial signatures across Normal, Obese, and Disease groups. Raw shotgun metagenomic reads were quality assessed, trimmed, and taxonomically classified using Kaiju, followed by multi-rank taxonomic composition analysis from phylum to species level followed by diversity analysis. Differentially abundant taxa were identified using the Mann–Whitney U test, while Random Forest classification and SHAP-based explainable AI were used to prioritize disease-associated microbial biomarkers. Microbial co-occurrence network analysis were further performed to interpret metabolic and ecological changes. Taxonomic profiling identified a diverse gut microbial community across multiple ranks, including 31 phyla, 52 classes, 114 orders, 203 families, 388 genera, and 535 species. Disease samples showed marked taxonomic shifts, reduced alpha diversity, and clear Bray–Curtis separation from Normal and Obese groups. PERMANOVA confirmed significant community-level differences across all taxonomic ranks using Bray–Curtis distances. Differential abundance analysis revealed extensive disease-associated taxa, particularly at genus and species levels, with stronger separation in Normal vs Disease and Obese vs Disease comparisons than in Normal vs Obese. Random Forest models showed strong classification performance for disease-related comparisons, with high ROC-AUC values, while Normal vs Obese showed weaker discrimination. SHAP-based explainable AI highlighted candidate species-level biomarkers, including *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *Dorea longicatena*, *Faecalibacterium prausnitzii*, *Phocaeicola vulgatus*, and *Morganella morganii*. Co-occurrence network analysis indicated reduced microbial network complexity in Disease samples. Overall, this study demonstrates that integrating metagenomics, statistical testing, machine learning, explainable AI, functional annotation, and network ecology provides a robust framework for identifying disease-associated gut microbiome signatures and candidate non-invasive microbial biomarkers.

Keywords: Metabolic syndrome; gut microbiome; metagenomics; microbial biomarkers; machine learning; explainable artificial intelligence; taxonomic profiling; dysbiosis.

Contents

Author's Declaration	iii
Plagiarism Undertaking	iv
Acknowledgement	v
Abstract	vi
List of Figures	xi
List of Tables	xv
Abbreviations	xvi
1 Introduction	1
1.1 Background	1
1.1.1 Pathophysiology of MetS	2
1.1.2 Causes and Effects of MetS	3
1.1.3 Development of MetS	4
1.1.4 Diagnostic Criteria of MetS	5
1.1.5 Global Disease Burden of MetS	7
1.2 Human Gut Microbiota	7
1.2.1 Human Gut Microbiota Dysbiosis	8
1.3 Problem Statement	10
1.4 Research Question	10
1.5 Hypothesis	11
1.6 Aims and Objectives	11
1.7 Scope of Study	11
1.8 Impact on Society	13
2 Literature Review	14
2.1 Metabolic Syndrome: Clinical and Pathophysiological Overview . .	14
2.2 Gut Microbiome and Host Metabolic Regulation	15
2.3 Metagenomic Approaches for Microbiome Analysis	16
2.4 Artificial Intelligence in Microbiome-Based Biomarker Discovery . .	17

2.5	Recent Advances in AI-Driven Metagenomic Research	18
2.6	Explainable Artificial Intelligence and Precision Medicine Potential of Microbiome-Derived Biomarkers	20
2.7	Research Gaps in Existing Literature	20
2.8	Proposed Study Rationale	22
3	Methodology	23
3.1	Study Design	23
3.2	Metagenomics Data Retrieval	24
3.3	SRA Run Selection and Metadata Preparation	24
3.4	Raw Sequence Data Download	25
3.5	Raw Read Quality Assessment using FastQC	25
3.6	Adapter Removal and Quality Trimming using Trimmomatic	25
3.7	Post-Trimming Quality Assessment	26
3.8	Metagenomics Assembly	26
3.9	Contig Binning	26
3.10	Genome Bin Assessment using CheckM	27
3.11	Reads Level Taxonomic Classification Using Kaiju	27
3.12	Taxonomic Composition Analysis	27
3.13	Wilcoxon Rank-Sum Test	28
3.14	Alpha Diversity Analysis	28
3.15	Beta Diversity Analysis	29
3.16	Principal Coordinate Analysis	29
3.17	PERMANOVA Analysis	29
3.18	Differential Abundance Analysis	30
3.19	Machine Learning Based Classification using XGBoost and Ex- plainable AI	30
3.20	Candidate Biomarker Prioritization	31
3.21	Performance Matrices Evaluation	31
3.22	Co-occurrence Analysis	32
3.23	Data Visualization	33
3.24	Computational Environment	33
4	Results	34
4.1	FastQC Report	34
4.2	Adapter Removal and Quality Trimming using Trimmomatic	35
4.3	Metagenomics Assembly Summary	35
4.4	Contig Binning Performance	36
4.5	Genome Bin Assessment using CheckM	37
4.6	Taxonomic Composition Analysis	38
4.7	Wilcoxon Rank-Sum Test	45
4.7.1	Normal vs Obese Comparison	45
4.7.2	Normal vs Disease Comparison	47
4.7.3	Obese vs Disease Comparison	49
4.8	Alpha Diversity Analysis	51

4.9	Beta Diversity Analysis	54
4.10	PERMANOVA Analysis	55
4.11	Machine Learning Based Classification using XGBoost and Ex- plainable AI for Candidate Biomarker Proritization	57
4.11.1	Random Forest Species-level Biomarker Classification Results	57
4.11.1.1	Normal vs Disease Comparison	58
4.11.1.2	Normal vs Obese Comparison	58
4.11.1.3	Obese vs Disease Comparison	60
4.11.2	XAI-based Disease Biomarker Analysis	63
4.11.2.1	Normal vs Disease Comparison	64
4.11.2.2	Obese vs Disease Comparison	67
4.12	Differential Abundance Analysis	69
4.13	Performance Matrices Evaluation	71
4.14	Cooccurrence Analysis	73
5	Discussion	77
6	Conclusion and Future Work	86
	Bibliography	88

List of Figures

1.1	Core components of metabolic syndrome [1].	2
1.2	Causes and consequences of metabolic syndrome [4].	4
1.3	Age-related development and progression of metabolic syndrome [7]	5
1.4	Global burden of metabolic syndrome [10].	7
1.5	Functional roles of the human gut microbiota [11].	8
1.6	Human gut microbiota dysbiosis and associated systemic disorders [12].	9
1.7	Scope of the proposed AI-driven metagenomic framework for biomarker discovery in metabolic syndrome.	12
1.8	Potential societal impact of AI-driven metagenomic biomarker dis- covery in metabolic syndrome.	13
3.1	Methodological framework for microbiome analysis.	23
4.1	General FastQC summary of SRR782588.	34
4.2	Summary of adapter removal and read-quality trimming using Trim- momatic.	35
4.3	Metagenomic assembly statistics of assembled contigs.	36
4.4	Summary of metagenomic contig binning performance across se- lected samples.	37
4.5	CheckM-based genome completeness assessment of recovered metaS- PAdes assemblies, showing single-copy marker distribution, missing markers, heterogeneity, and contamination levels across assemblies 86–97.	38
4.6	Species-level taxonomic composition and summary of statistically significant taxa across metadata groups. Stacked bar plots show the relative abundance of dominant species across Normal, Obese, and Disease groups, while the summary plot shows the number of significant taxa detected at different taxonomic levels using a threshold of $p < 0.05$	45

4.7	Mann–Whitney U test-based differential abundance analysis between the Normal and Obese groups across multiple taxonomic levels. Panel present the significantly different taxa identified at the phylum, family, genus, and species levels, respectively. Red boxplots represent the Normal group, whereas blue boxplots represent the Obese group. Each point corresponds to the relative abundance of a taxon in an individual sample. The central line within each box represents the median, the box indicates the interquartile range, and the whiskers illustrate the variability of the abundance values.	46
4.8	Mann–Whitney U test-based differential taxa analysis between Normal and Disease groups. Boxplots show significantly different taxa between Normal and Disease samples across the phylum, family, genus, and species levels. Each point represents an individual sample, boxes show group-wise abundance distributions, and asterisks indicate statistically significant differences at $p < 0.05$. The figures demonstrate the progression of disease-associated microbial differences from broad to specific taxonomic levels.	49
4.9	Mann–Whitney U test-based differential taxa analysis between Obese and Disease groups. Boxplots show significantly different taxa between Obese and Disease samples across the phylum, family, genus, and species levels. Each point represents an individual sample, boxes show group-wise abundance distributions, and asterisks indicate statistically significant differences at $p < 0.05$. The figures demonstrate the progression of disease-associated microbial differences from broad to specific taxonomic levels.	51
4.10	Bray–Curtis distance heatmap: Pairwise species-level dissimilarity matrix across samples, where darker colors represent lower dissimilarity and yellow colors represent higher dissimilarity.	55
4.11	Bray–Curtis PCoA plot: Principal coordinate analysis based on species-level Bray–Curtis distances, showing clear separation of Disease samples from Normal and Obese samples along PCoA1.	56
4.12	PERMANOVA-based comparison of Bray–Curtis beta diversity across taxonomic ranks. The plot shows PERMANOVA test statistics for phylum, class, order, family, genus, and species levels across Disease, Obese, and Normal groups. All taxonomic ranks showed significant group-wise compositional differences ($p = 0.001$) based on 999 permutations, with the strongest separation observed at the phylum level followed by class, order, family, genus, and species levels.	57
4.13	Normal vs Disease comparison using species-level Random Forest biomarker classification. The confusion matrix shows classification performance, while the ROC curve represents model discrimination ability.	59

4.14	Normal vs Disease comparison using species-level Random Forest biomarker classification. The heatmap shows the abundance pattern of top species-level biomarkers across samples, while the feature importance plot highlights the most influential microbial taxa contributing to classification.	60
4.15	Normal versus Obese comparison at the species level using Random Forest biomarker classification. The confusion matrix shows misclassification between the two groups, while the ROC curve indicates poor discriminatory performance with an AUC below 0.5.	61
4.16	Normal vs Obese Comparison Specie Level Random Forest Biomarker Classification Heatmap and Feature Importance.	62
4.17	Obese vs Disease Comparison Specie Level Random Forest Biomarker Classification Confusion Matrix and ROC Curve.	62
4.18	Obese vs Disease Comparison Specie Level Random Forest Biomarker Classification Heatmap and Feature Importance.	63
4.19	XAI-based disease-associated species biomarker analysis for Normal vs Disease comparisons Confusion Matrix and ROC Curve.	65
4.20	XAI-based disease-associated species biomarker analysis for Normal vs Disease comparisons Feature Importance and SHAP plot	66
4.21	XAI-based disease-associated species biomarker analysis for Normal vs Disease comparisons Confusion Matrix and ROC Curve.	68
4.22	XAI-based disease-associated species biomarker analysis for Obese versus Disease comparison, including biomarker heatmap, feature importance, and SHAP summary plot.	69
4.23	Integrated XAI and Mann–Whitney U test-based species-level biomarker bubble plots. The plots show candidate biomarker species identified by combining statistical differential abundance testing with SHAP-based model importance. The x-axis represents $\log_2$ fold-change, the y-axis represents $-\log_{10}(p\text{-value})$, point color indicates the enriched group, and bubble size represents mean absolute SHAP value.	70
4.24	Heatmap of AI model evaluation metrics. Comparative performance summary across genus- and species-level classification models, showing high metric values for disease-related comparisons and lower values for Normal versus Obese models.	71
4.25	Presence–Absence heatmap of the top 40 species across normal, obese, and disease groups.	73
4.26	Disease-group microbial co-occurrence network. The network shows a single positive association between <i>Bacteroides</i> and <i>Phocaeicola</i> , with 2 nodes, 1 edge, 1 positive connection, and no negative edges.	74
4.27	Normal-group microbial co-occurrence network. The network displays genus-level microbial associations in Normal samples, showing 16 nodes and 16 retained edges, including 11 positive and 5 negative relationships.	75

4.28	Obese-group microbial co-occurrence network. The network shows genus-level microbial co-occurrence associations in Obese samples, containing 7 nodes and 10 positive edges, with no negative associations detected.	76
5.1	Graphical abstract of the integrated metagenomic and explainable AI framework for disease-associated gut microbiome biomarker discovery.	84

List of Tables

1.1	Diagnostic criteria for metabolic syndrome according to different clinical definitions [5, 8, 9].	6
2.1	Summary of recent studies using AI, machine learning, and metagenomics for microbiome-based biomarker discovery.	19
4.1	Top ten phyla ranked according to their mean relative abundance across all study samples.	39
4.2	Top ten microbial classes ranked according to their mean relative abundance across all study samples.	40
4.3	Top ten microbial orders ranked according to their mean relative abundance across all study samples.	40
4.4	Top ten microbial families ranked according to their mean relative abundance across all study samples.	41
4.5	Top ten microbial genera ranked according to their mean relative abundance across all study samples.	42
4.6	Top ten microbial species ranked according to their mean relative abundance across all study samples.	43
4.7	Top species-level biomarkers identified using SHAP importance, statistical significance, FDR correction, and $\log_2$ fold-change analysis between Normal and Disease groups.	66
4.8	Top species-level biomarkers identified using SHAP importance, statistical significance, FDR correction, and $\log_2$ fold-change analysis between Obese and Disease groups.	67

Abbreviations

MetS	Metabolic Syndrome
NCEP–ATP	National Cholesterol Education Program–Adult Treatment Panel
IDF	International Diabetes Federation
AHA/NHLBI	American Heart Association and National Heart, Lung and Blood Institute
HDL	High-density lipoprotein

Chapter 1

Introduction

1.1 Background

Metabolic syndrome is a complex, multifactorial syndrome and a group of metabolic abnormalities occurring in the same person. As indicated in the figure [1.1](#), the main components of metabolic syndrome are visceral obesity, insulin resistance, hypertension, high triglycerides and low HDL cholesterol. These abnormalities are not independent of each other, but rather are biologically linked, and frequently caused by a similar underlying pathology, namely an excess of visceral fat accumulation, an impaired insulin signaling, a chronic inflammatory status, and metabolic stresses due to lifestyle factors [\[1\]](#).

The metabolic syndrome as a grouping of health conditions, a crucial point to comprehend the disease. The person might not look like they are very sick, but the combination of abdominal obesity, high blood pressure, high cholesterol and insulin resistance will slowly raise the likelihood of serious cardiometabolic conditions. Visceral obesity is important because fat around the abdominal organs is more metabolically damaging than fat under the skin.

Palpable release of free fatty acids, inflammatory cytokines and adipokines by visceral adipose tissue, which disrupt insulin action and lipid metabolism. This is one of the key components of metabolic syndrome. Another factor that is shown in the data is insulin resistance. Normally, insulin facilitates glucose uptake into tissues (e.g., skeletal muscle, adipose tissue), and inhibits glucose production in

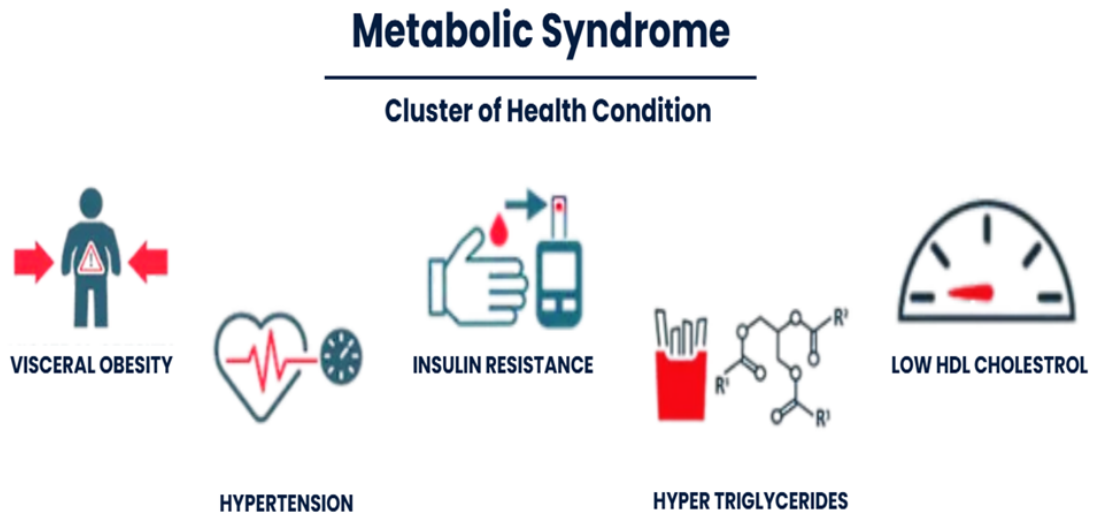


FIGURE 1.1: Core components of metabolic syndrome [1].

the liver. In insulin resistance, however, these tissues become insensitive to the effects of insulin. This leaves the blood glucose higher, the pancreas trying to catch up by producing more insulin, and the blood glucose level being continually elevated by this extra production of insulin, leading to metabolic stress. This hyper-insulinemic state and abnormal glucose tolerance may lead to Type 2 Diabetes Mellitus. Thus, insulin resistance can serve as a mediating factor between obesity, hyper-glycemia, dys-lipidaemia and cardiovascular risk [1, 2].

1.1.1 Pathophysiology of MetS

The pathophysiology gives a biological level explanation of the development of metabolic syndrome. Mechanisms are insulin resistance, adipose tissue dysfunction, visceral obesity, and chronic inflammation. The adipocytes in people with excess visceral fat become large and non-functional. These dysfunctional adipocytes release more free fatty acids to the bloodstream. These fatty acids are absorbed by the liver and converted into triglycerides which may cause hypertriglyceridemia and fatty liver. Meanwhile, at the same time inflammatory mediators like tumor necrosis factor- α , interleukins and other cytokines disrupt the insulin signaling pathways, further adding to insulin resistance. Adipose tissue dysfunction also plays a role in the regulation of adipokines. Adipose tissue secretes hormones

in a healthy metabolic state that control appetite, insulin sensitivity, inflammation, and lipid metabolism. In metabolic syndrome, however, these protective adipokines, like adiponectin, are decreased while these pro-inflammatory mediators are heightened. Low levels of adiponectin are linked to poor fatty acid oxidation, decreased insulin sensitivity and increased vascular inflammation. This results in a systemic inflammatory condition and is why metabolic syndrome is referred to as a "disease of chronic low-grade inflammation" [2], [3].

1.1.2 Causes and Effects of MetS

Metabolic syndrome is a multifactorial disorder, as explained in the causes and consequences figure 1.2, listed causes are lack of physical exercise, consuming more calories than burning; unhealthy diet with high carbohydrate and animal fat; sedentary lifestyle; alcohol consumption; stress; less than ideal quality sleep or sleep pattern; and shift work. These are very relevant in today's societies where many people are becoming more and more urban, eat processed meals, don't exercise much and have irregular sleeping habits. Physical inactivity lowers the level of energy expenditure and leads to weight gain. It also reduces the uptake of glucose into the skeletal muscle causing insulin resistance. Over-consumption of calories, particularly of refined carbohydrates and saturated fats leads to increased fat storage and dyslipidemia. Obesity and vascular dysfunction are also worsened by sedentary lifestyle. Chronic stress also triggers activation of the neuroendocrine pathways such as an increase of cortisol, which can lead to increased appetite, central fat deposits, and glucose levels [4, 5]. Shift work and poor sleep destroys 4 key factors for good health: circadian rhythm, hormonal balance, appetite/satiety control and insulin sensitivity. All of these lifestyle factors contribute to the pattern of a person's metabolism allowing metabolic syndrome to develop. The consequences shown in the figure include diabetes, fatty liver disease, kidney injury, heart failure, obstructive sleep apnea, polycystic ovary syndrome, dyslipidemia, high heart rate, uric acid elevation, and inflammation. These effects show there is a relationship between metabolic syndrome and more than just cosmetic and weight problems. A systemic disease that has significant clinical implications. Ob-

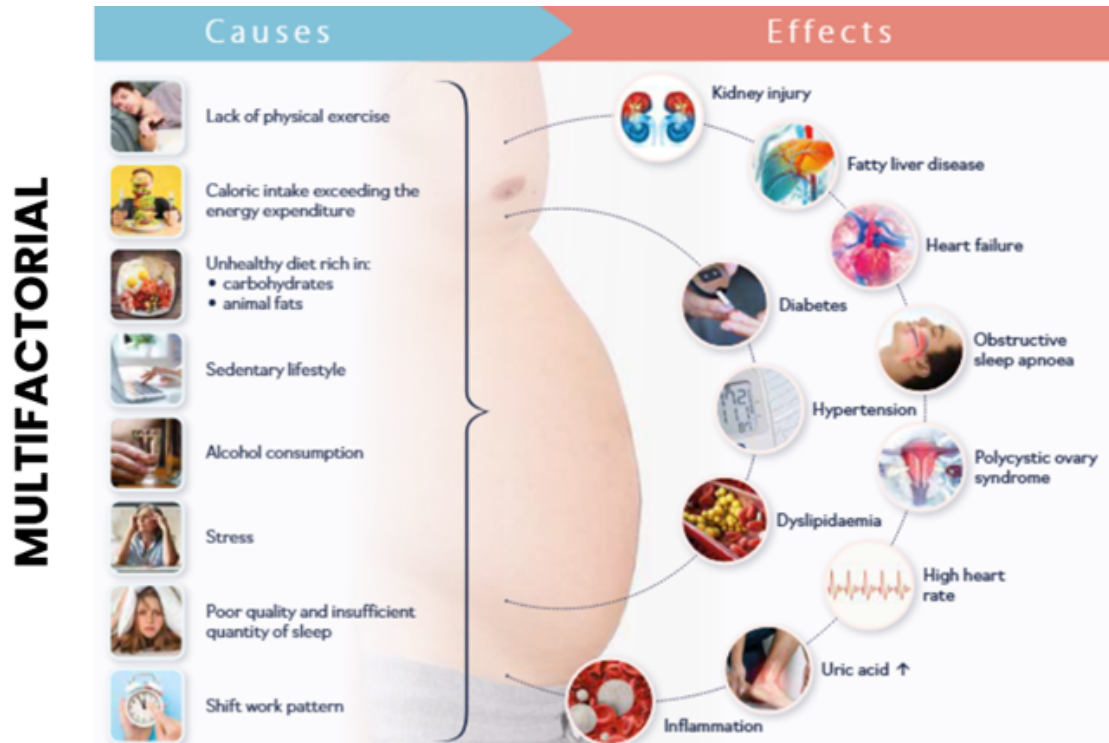


FIGURE 1.2: Causes and consequences of metabolic syndrome [4].

structive sleep apnea may increase insulin resistance and hypertension by inducing intermittent hypoxia and activation of sympathetic nerves, for instance. Women with PCOS have a very close relation with insulin resistance and metabolic abnormalities. Hyperuricemia can be caused by renal defects and higher metabolic stress. These conditions are associated with metabolic syndrome and make it even more complicated [5, 6].

1.1.3 Development of MetS

Metabolic syndrome is a condition which will evolve, as indicated by the figure 1.3. Often starts with normal body weight and then slowly moves into overweight and obesity and eventually to multiple metabolic issues. Cardiometabolic risk also gets worse with age. This is a gradual process that may take years to develop as a result of chronic unhealthy eating, decreased amount of physical activity, sitting around too much, stress and poor sleep habits. It is observed that early risk factors like overweight and obesity can eventually result in pre-diabetes, high blood pressure, dyslipidemia, chronic inflammation, polycystic ovary syndrome, atherosclerosis,

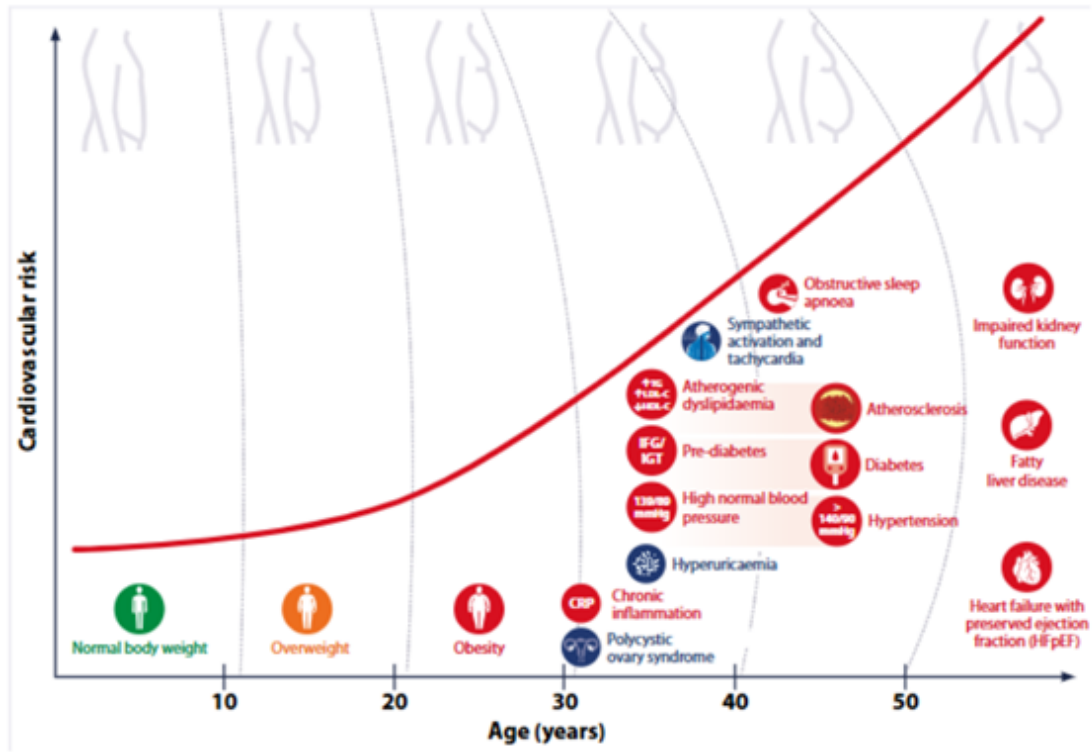


FIGURE 1.3: Age-related development and progression of metabolic syndrome [7]

diabetes, liver disease, kidney impairment and heart failure. Clinically important in this progression is that metabolic syndrome serves as an opportunity for early intervention. When people are detected during the central obesity, insulin resistance or mild dyslipidemia stage, treatment and lifestyle changes can help to prevent the onset of diabetes and cardiovascular complications. As the disease progresses, however, if it goes undiagnosed, the risk will slowly build and can build up until it causes irreversible organ damage [2, 7].

1.1.4 Diagnostic Criteria of MetS

The diagnostic criteria shown in table 1.1 indicates that metabolic syndrome is diagnosed by clinical parameters, which can be measured, such as waist circumference, HDL cholesterol, triglycerides, fasting glucose, and blood pressure. These parameters are used is due to the fact that they are the major metabolic disturbances associated with the syndrome. Central adiposity is associated with an increase in waist circumference. Low HDL concentration is a sign of inadequate lipid transport and poor cardiovascular protective function. A raised triglyceride

level suggests abnormal lipid metabolism that is usually associated with high levels of fat in the liver and insulin resistance. Increased blood glucose and blood pressure levels are markers of glucose dyshomeostasis and vascular dysfunction and cardiovascular burden. While the definitions vary slightly among NCEP-ATP III, IDF, AHA/NHLBI, and ATP III, they all share the same basic guideline: metabolic syndrome should be diagnosed when several cardiometabolic risk factors are present.

TABLE 1.1: Diagnostic criteria for metabolic syndrome according to different clinical definitions [5, 8, 9].

Criterion	NCEP-ATP III ¹	IDF ²	AHA/NHLBI ³	ATP III ⁴
Waist circumference (cm)				
Males	≥ 102	≥ 94	≥ 94	≥ 102
Females	≥ 88	≥ 80	≥ 80	≥ 88
HDL cholesterol (mg/dL)				
Males	< 40	< 40	< 40	< 40
Females	< 50	< 50	< 50	< 50
Triglycerides (mg/dL)	≥ 150	≥ 150	≥ 150	≥ 150
Fasting glucose (mg/dL)	≥ 100	≥ 100	≥ 100	≥ 100
Blood pressure (mmHg)	$\geq 130/ \geq 85$	$\geq 130/ \geq 85$	$\geq 130/ \geq 85$	$\geq 130/ \geq 85$

¹NCEP-ATP = National Cholesterol Education Program-Adult Treatment Panel; ²IDF = International Diabetes Federation; ³AHA/NHLBI = American Heart Association and National Heart, Lung, and Blood Institute; ⁴HDL = high-density lipoprotein. A: Three of five criteria required. B: Central adiposity required; two of the subsequent four criteria required.

Waist circumference cut-offs, for instance, vary between organizations particularly between the sexes but all identify central obesity as a key diagnostic characteristic. Likewise, affected persons have low normal HDL cholesterol levels, and their triglyceride, blood glucose and blood pressure levels are elevated, respectively higher than 150 mg/dL, higher than 100 mg/dL, and higher than 130/85 mmHg. These are only the 5 criteria that define metabolic syndrome but there is a much wider underlying biology [5, 8, 9].

1.1.5 Global Disease Burden of MetS

The global burden number corroborates the public health significance of metabolic syndrome as shown in figure 1.4. Based on the data given, it is estimated that, in 2023, an estimated 1.54 billion adults globally had metabolic syndrome. The prevalence was presented to be 31.0% of females and 25.7% of males with estimated numbers of prevalent cases being 846 million females and 692 million males.

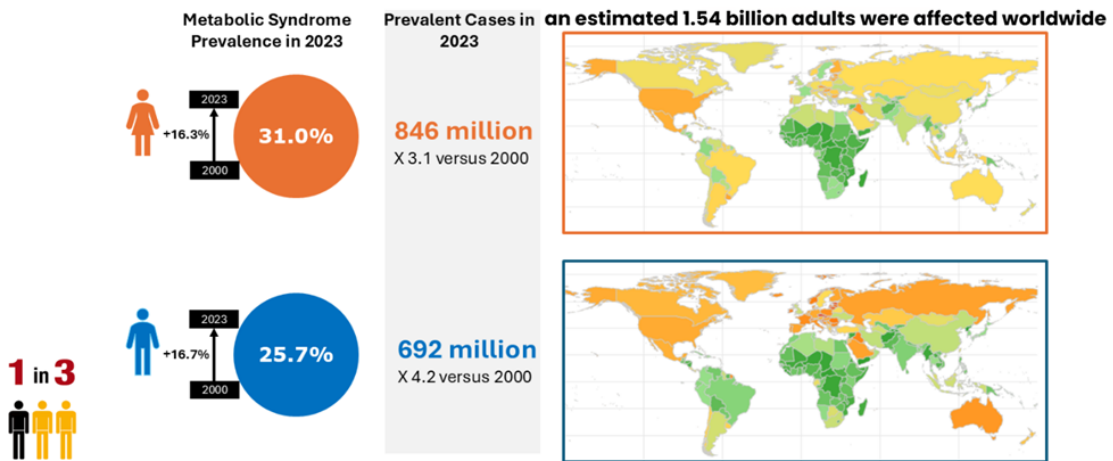


FIGURE 1.4: Global burden of metabolic syndrome [10].

This suggests that metabolic syndrome is not an uncommon disease, and that many adults in the world suffer from it. The increase over 2000 also indicates that the load is increasing significantly. The increased prevalence in females might be attributed to body fat distribution, hormonal differences, hormonal changes with menopause, lifestyle differences, and physical activity patterns. But there is a high disease burden in both males and females. The global maps also indicate that metabolic syndrome is common to many regions of the world, not just in high income countries. This is also relevant as low- and middle-income countries are also undergoing a rapid transition in their diet and lifestyle that could raise the risk of metabolic diseases [10].

1.2 Human Gut Microbiota

Microbiota or microbial community of the human gut represents a complex ecosystem of microorganisms residing in the gut and helping to maintain human health

as shown in figure 1.5. Gut microbiota plays a role in metabolism of therapeutic compounds, modulation of insulin-related pathways, angiogenesis, fat storage, resisting pathogen colonization, protection of intestinal epithelium, production of vitamins and amino acids, and breaking down food compounds as shown in the figure. The functions demonstrate that the gut microbiota is a metabolically active system and not a mere microbial community.

In the context of metabolic syndrome its function in fat storage, modulation of insulin system, control of nutrients, control of inflammation and protection of the epithelial barrier is particularly relevant, as any disturbance of these functions can be a contributing factor to obesity, insulin resistance, dyslipidemia and chronic low-grade inflammation [11].

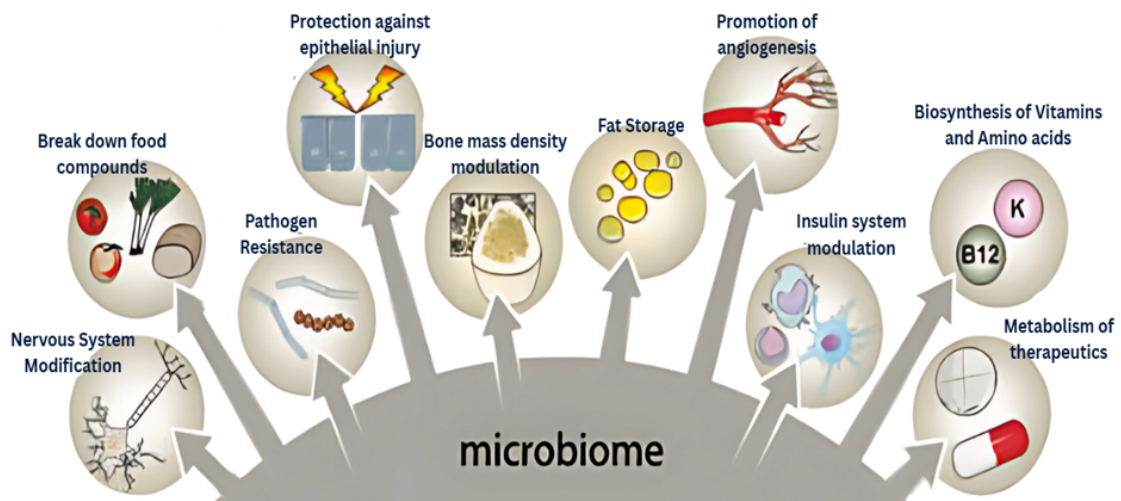


FIGURE 1.5: Functional roles of the human gut microbiota [11].

1.2.1 Human Gut Microbiota Dysbiosis

Human gut microbiota dysbiosis is a decrease in beneficial microbes, an increase in harmful microbes, a change in diversity and a disruption of microbial functions as shown in figure 1.6. Human gut microbiota dysbiosis refers to an imbalance in the composition, diversity and functional activity of the intestinal microbial community. It is generally characterized by a reduction in beneficial commensal microorganisms, an increase in potentially pathogenic microbes and alterations in microbial metabolite production.

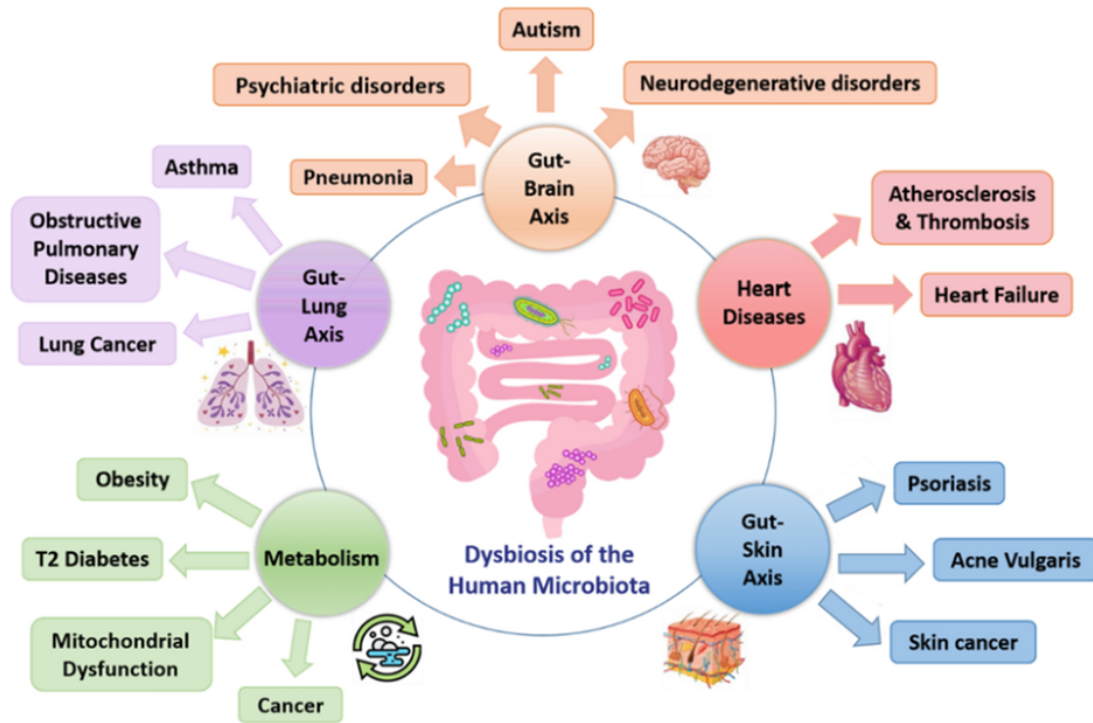


FIGURE 1.6: Human gut microbiota dysbiosis and associated systemic disorders [12].

One of the major consequences of dysbiosis is disruption of the intestinal epithelial barrier. Increased intestinal permeability allows bacterial components, particularly lipopolysaccharides, to enter the systemic circulation. These components activate immune and inflammatory pathways, resulting in persistent low-grade inflammation. Chronic inflammation can interfere with insulin signalling, reduce glucose uptake and promote the development of insulin resistance. It can also alter lipid metabolism and facilitate excessive fat accumulation, particularly within visceral adipose tissue.

Dysbiosis may also modify the production of important microbial metabolites, including short-chain fatty acids, secondary bile acids, branched-chain amino acids and trimethylamine-derived compounds. Short-chain fatty acids normally contribute to intestinal barrier integrity, immune regulation, glucose homeostasis and energy metabolism. Therefore, a reduction in beneficial metabolite-producing bacteria may impair metabolic regulation. Conversely, excessive production of certain microbial metabolites may contribute to inflammation, dyslipidaemia, vascular dysfunction and cardiovascular disease.

The gut–brain axis involves bidirectional communication between the intestinal microbiota and the central nervous system through neural, endocrine, immune and metabolic pathways. Alterations in this communication have been associated with neurological and psychiatric disorders. Similarly, the gut–lung axis may influence respiratory immune responses and has been associated with asthma, pneumonia and obstructive pulmonary disorders. The gut–skin axis connects intestinal microbial changes with systemic inflammation and immune dysfunction, potentially contributing to conditions such as psoriasis and acne.

Therefore, gut microbiota dysbiosis may act as an important biological link between obesity, insulin resistance, dyslipidaemia, hypertension and systemic inflammation. These interconnected mechanisms support the investigation of microbial taxa and microbial functions as potential biomarkers for the early detection, classification and progression of metabolic syndrome [12].

1.3 Problem Statement

Metabolic syndrome is a rapidly increasing public health problem characterized by obesity, insulin resistance, dyslipidemia, hypertension, and chronic inflammation. Traditional risk factors like fasting glucose, triglyceride, HDL cholesterol, waist circumference and blood pressure are clinically relevant, but these primarily identify metabolic syndrome once it has been established. The existing microbiome studies associated with metabolic syndrome tend to be descriptive and consist only of taxonomic profiling, rather than a focus on accurate prediction, functional interpretation, explainable discovery of biomarkers, and personalized medicine. Thus, an integrated AI-powered metagenomic framework is needed to process shotgun metagenomic data, discover disease-associated microbial features, forecast metabolic syndrome status, and elucidate the key microbial biomarkers for the diagnosis of every disease.

1.4 Research Question

How can explainable AI-driven metagenomic analysis be used to develop robust, interpretable, and clinically relevant microbial biomarker models for metabolic

syndrome prediction and personalized risk stratification?

1.5 Hypothesis

Shotgun metagenomic profiles of the human gut microbiome can have unique microbial taxonomic and functional signatures to distinguish between people with metabolic syndrome and healthy controls. In addition, the incorporation of machine learning, deep learning, and explainable artificial intelligence will enhance the prediction accuracy of the disease and uncover the microbial markers that are easy to understand and are associated with metabolic syndrome.

1.6 Aims and Objectives

The aim of this study is to develop an AI-driven metagenomic framework for the discovery of gut microbiome-derived biomarkers associated with metabolic syndrome. The objectives include:

- i. To analyze shotgun metagenomic data from metabolic syndrome and control groups for taxonomic and functional profiling of the gut microbiome.
- ii. To predict microbial biomarkers in normal, obese, and disease conditions using machine learning and deep learning techniques.
- iii. To analyze key microbial biomarkers and their clinical relevance in metabolic syndrome.

1.7 Scope of Study

This study concentrates on the computational aspects of shotgun metagenomic data to find gut microbiome signatures related to metabolic syndrome as shown in figure [1.7](#).

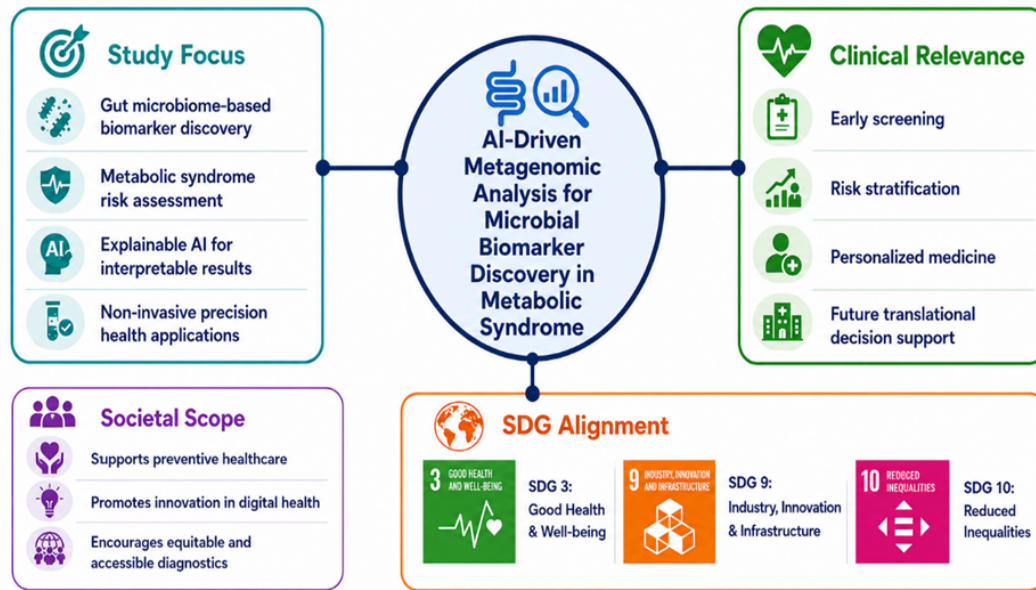


FIGURE 1.7: Scope of the proposed AI-driven metagenomic framework for biomarker discovery in metabolic syndrome.

Features encompass data preprocessing, quality control, host-read removal, taxonomic profiling, functional profiling, AI based disease prediction, and explainable biomarker discovery. Pathways, genes, and microbial taxa associated with obesity, insulin resistance, inflammation, dyslipidemia and metabolic dysregulation will be explored. This study has a clinical objective of helping identify non-invasive microbial markers that could enhance the early screening, risk stratification and personalized treatment of metabolic syndrome (MetS). The proposed framework can be useful to fill the gap between microbiome study and clinical decision making by generating interpretable microbiome biomarkers rather than just microbial profiles. Moreover, the study is linked to the pertinent Sustainable Development Goals such as SDG 3: Good Health and Well-being, which is related to early detection and prevention of chronic metabolic complications. It could also support SDG 9: Industry, Innovation and Infrastructure in terms of AI-based biomedical innovation, and SDG 10: Reduce Inequalities in its application of non-invasive, data-driven approaches that may provide preventative healthcare in the future that is accessible to all.

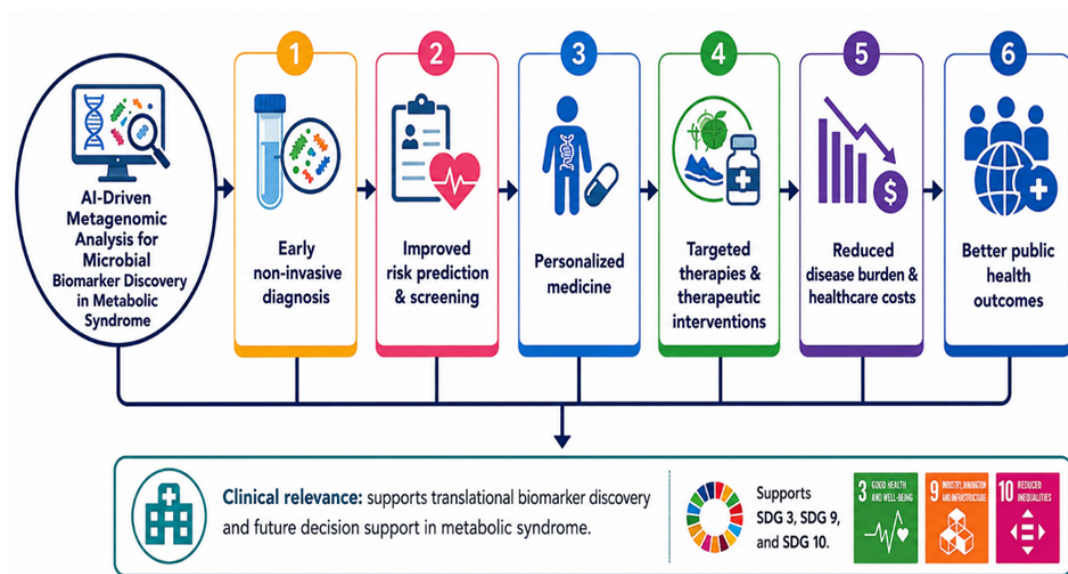


FIGURE 1.8: Potential societal impact of AI-driven metagenomic biomarker discovery in metabolic syndrome.

1.8 Impact on Society

This research contribution may help society by providing support for the early, non-invasive and personalized detection of metabolic syndrome. The discovery of these biomarkers can help lower the burden of disease if metabolic syndrome is a risk factor for other conditions, such as type 2 diabetes, cardiovascular disease, fatty liver disease, kidney dysfunction, and other chronic conditions. Clinical applications of this work are likely to be the use of microbial biomarkers to aid in disease risk prediction, early detection and planning personalized interventions. The findings of the study could be more easily interpretable and useful for future translational studies because they can be linked to microbial features with explainable AI. Public health applications of microbiome-based biomarker discovery could be precision medicine, lifestyle-based prevention and targeted therapeutics. This can help achieve better health results, lower long-term health care expenditures, and better individual and population health care management of metabolic diseases as shown in figure 1.8.

Chapter 2

Literature Review

2.1 Metabolic Syndrome: Clinical and Pathophysiological Overview

Metabolic syndrome is a cardiometabolic disorder that occurs when central obesity, insulin resistance, hypertension, dyslipidemia and impaired glucose regulation coexist. It is not a single disease, but a collection of linked metabolic abnormalities that all contribute to an increased risk of developing type 2 diabetes mellitus, cardiovascular disease, non-alcoholic fatty liver disease, chronic kidney disease, and other chronic diseases. The following is a combination of all of these key clinical features of metabolic syndrome that are generally considered to be part of the definition as it is used in clinical practice: increased waist circumference, increased blood pressure, increased fasting blood glucose, increased triglycerides, and decreased HDL, though the waist circumference may differ based on ethnicity and population-specific risk profiles [1, 8].

The progressive and systemic nature of metabolic syndrome is a key clinical significance. People who have metabolic syndrome may experience metabolic dysfunction over time, starting with excess adiposity and insulin resistance, and then moving into diabetes, vascular dysfunction, and organ-specific dysfunction. Insulin resistance is a pivotal part of this process, leading to the reduction of glucose uptake in skeletal muscle and fat, and hepatic glucose production. In addition to

insulin resistance, visceral adiposity is associated with metabolic syndrome via the causes of enhanced release of free fatty acids, imbalance of adipokines, oxidative stress and chronic low-grade inflammation [7].

The role of the visceral fat as just a repository for excess energy is now viewed as incorrect, and the visceral fat has been shown to be an active endocrine and immunological organ. Enlarged adipocytes and infiltrating immune cells produce inflammatory cytokines and adipokines that disrupt insulin signaling and drive systemic inflammation, which is a hallmark of obesity. This inflammatory setting is associated with dyslipidemia, endothelial dysfunction, hypertension, hepatic lipid accumulation and vascular injury.

Metabolic syndrome thus must be viewed as a "multi-organ disease" with patterns of synchronized dysfunction involving adipose tissue, liver, skeletal muscle, pancreas, vascular endothelial cells and immune cells. This is the basis of the strong association of metabolic syndrome with cardiometabolic morbidity and mortality, which is intrinsically linked to the above pathophysiology [5, 6].

2.2 Gut Microbiome and Host Metabolic Regulation

The human gut microbiota has become an important player in host metabolism and immune system. It is a complex ecosystem made up of various microorganisms, such as bacteria, archaea, fungi, viruses and others, primarily those that inhabit the gastrointestinal system.

The gut microbiota plays a role in nutrient metabolism, xenobiotic and drug metabolism, maintenance of gut mucosal barrier integrity, immune modulation and resistance against pathogenic organisms. It also helps in fermentation of indigestible dietary carbohydrates to produce short-chain fatty acids (SCFAs) like acetate, propionate, and butyrate that affect the health of the epithelium, inflammatory signaling, glucose metabolism, lipid regulation, and pathways that affect appetite [11].

A healthy gut microbiome regulates the production of the gut mucus, integrity

of tight junctions, renewal of epithelial cells, and immune responses. This barrier function is of significant importance with respect to metabolic syndrome as it can lead to microbial products from the gut, like lipopolysaccharides, being released into systemic circulation. It can be termed metabolic endotoxemia, which can promote chronic low-level inflammation and dysregulate insulin sensitivity in metabolic tissues. This mechanism can cause gut microbial imbalance to play a role in development of insulin resistance, obesity-related inflammation, dyslipidemia and impaired glucose regulation which are core characteristics of metabolic syndrome [13].

Gut Dysbiosis is an imbalance of gut microbiota composition, diversity, stability or function. Dysbiosis can include loss of beneficial organisms, proliferation of opportunistic organisms, decreased diversity of organisms, changes in the production of microbial metabolites, and changes in host–microbe communication. In recent years, dysbiosis has been associated with various systemic disease axes such as gut–brain axis, gut–lung axis, gut–heart axis, gut–skin axis, and metabolism-related disorders. This systemic impact underscores a gut microbiome’s impact on immune regulation, cardiometabolic function, neurological signaling, and inflammatory disease processes [12].

2.3 Metagenomic Approaches for Microbiome Analysis

Previous investigations of the microbiome have used 16S rRNA gene sequencing to define the composition of bacteria in the microbiome. While 16S sequencing can provide some information on the broad taxonomic patterns, it lacks species-level resolution and functional interpretation. Shotgun metagenomic sequencing is more holistic, since it sequences the entire microbial DNA from a sample, enabling the analysis of microbial taxa, genes, metabolic pathways and functional potential. This is especially important in metabolic syndrome research where disease-associated functions of microbes could be as important as their abundance. For instance, species of different microbes can carry out the same metabolism, so pathway-level analysis can uncover disease signatures that are biologically infor-

mative and that may not be captured by taxonomy-based methods [14, 15].

Although considerable advances have been made, there are still many microbiome studies of metabolic syndrome that are descriptive in nature. They often describe variation between the microbial community found in healthy and diseased subjects, but do not necessarily link these findings to predictive, reproducible, or clinically meaningful biomarkers. Microbiome datasets also have a large dimension, are sparse, compositional, and have significant inter-individual variation. These features pose difficulties for biomarker discovery with traditional statistical methods [12]. Thus, the need for more sophisticated computational approaches to generate disease relevant patterns within complex metagenomic data.

2.4 Artificial Intelligence in Microbiome-Based Biomarker Discovery

AI is a promising tool for analysis of complex microbiome data. Nonlinear relationships and interactions between microbial taxa, genes, pathways, and clinical variables can be detected by machine learning models. Typical algorithms applied to microbiome-based disease prediction are random forest, support vector machine, logistic regression, gradient boosting and neural network-based methods. Such models can be used to identify high dimensional microbial features that can be used to classify disease and control groups, and can be used to aid identification of microbial signatures associated with metabolic syndrome. Deep learning techniques also have great potential to further enhance pattern recognition by learning hierarchical representation of complex data sets automatically, but their effectiveness will rely heavily on the size of the data sample, the quality of the dataset and the design of the validation. AI has shifted the focus in microbiome research from descriptive profiling to predictive modelling. But just predicting is not enough for clinical translation [16, 17]. There is a need to validate AI models through scientifically sound techniques like cross-validation, independent testing, external validation cohorts, and comprehensive metrics like sensitivity, specificity, accuracy, F1-score, and the area under the receiver operating characteristic curve. If the model is not validated, it may work well on one data set but not on an-

other because people have different diets, different geographical locations, different ages, are taking different medications, using different sequencing protocols, and are using different bioinformatics pipelines. Therefore, building clinically relevant microbiome biomarkers is a complex task that is demanding both strong predictive performance and consistency between datasets [18].

2.5 Recent Advances in AI-Driven Metagenomic Research

Recent advancements in metagenomics and artificial intelligence have significantly transformed microbiome research, particularly in disease prediction and biomarker discovery. The integration of machine learning and deep learning techniques with high-throughput sequencing data has enabled researchers to uncover complex microbial patterns associated with various diseases, including metabolic syndrome [19]. Deep learning models, in particular, have shown superior performance in handling high-dimensional and nonlinear microbiome data, facilitating improved classification accuracy and functional interpretation. Moreover, recent studies emphasize that AI-driven microbiome analysis can enhance precision medicine by enabling personalized diagnostics and therapeutic strategies. Table 2.1 demonstrates the recent advancement in AI driven metagenomics approach [19, 20]. One of the drawbacks of many AI-based biomedical models is that they are black-box models. The complex machine learning and deep learning algorithms could give the correct answer, but not necessarily provide a clear indication of what features are the most important for the prediction. The lack of interpretability is especially troublesome in biomarker discovery where the clinician and the researcher must draw inferences about the reason for the relevance of a microbial feature to disease [18, 21]. Overall, these studies demonstrate the growing importance of combining metagenomic data with interpretable AI approaches for reliable microbial biomarker discovery. However, broader validation, standardized workflows, and clinically representative datasets remain necessary before routine medical application. These developments indicate that AI-assisted metagenomic analysis can improve the identification of complex microbial patterns that may remain unde-

tected through conventional statistical methods.

TABLE 2.1: Summary of recent studies using AI, machine learning, and metagenomics for microbiome-based biomarker discovery.

Study	Method used	Key findings	Limitation
Przymus et al., 2025	Deep learning (CNN, autoencoders)	Deep learning effectively extracts patterns from high-dimensional microbiome data.	Requires large datasets and has a risk of overfitting [19].
Tripathi et al., 2025	AI/ML in microbiome	AI enables disease prediction and biomarker discovery.	Data heterogeneity and lack of standardization [20].
Zhou et al., 2025	AI and metagenomics	Integration enhances precision medicine applications.	Clinical translation remains challenging [16].
Tegegne et al., 2025	Metagenomics and clinical application	The gut microbiome plays an important role in disease diagnosis.	Limited integration of computational and clinical approaches [13].
Wang et al., 2026	AI in microbiome	AI-driven models improve the interpretation of complex microbiome data.	Requires interdisciplinary expertise and computational resources [17].
Fu et al., 2026	Metagenomic atlas study	Identified microbiome patterns associated with disease and aging.	Limited focus on individual disease phenotypes [22].
Novielli et al., 2024	Explainable AI and gut microbiome data	Explainable AI identified microbial taxonomic markers contributing to disease classification at the individual-subject level.	Focused on colorectal cancer and requires broader validation in metabolic disorders [21].
Dakal et al., 2025	AI and multi-omics integration	Advanced computational tools can integrate metagenomics, metabolomics, metaproteomics, and other omics for biomarker discovery.	Multi-omics integration requires substantial computational resources and standardized datasets [23].
Li et al., 2025	Machine learning for microbiome-based disease prediction	Provided best-practice guidance for developing microbiome-based disease-prediction models and improving their reliability.	Model performance may decline across populations because of differences in diet, geography, and sequencing methods [24].
Gonçalves et al., 2025	Metadata integration and ML	Host metadata can improve microbiome-based disease prediction and reduce misleading microbial associations.	Requires complete and well-curated metadata, which are often unavailable in public datasets [25].
Cai et al., 2025	Interpretable ML and dietary gut microbiota index	Identified associations between a dietary gut microbiota index and components of metabolic syndrome using interpretable ML.	Used a dietary index as a microbiome proxy rather than direct shotgun metagenomic profiling [26].
Romano et al., 2025	ML-based meta-analysis of gut microbiome studies	Large-scale microbiome meta-analysis showed that ML can classify disease cases and identify reproducible microbial features.	Focused on Parkinson's disease; therefore, its relevance to metabolic syndrome is primarily methodological [27].
Chang et al., 2024	Explainable ML and gut microbiota/metabolites	Explainable AI identified important microbial and metabolite biomarkers for disease diagnosis.	A small sample size limited the generalizability of the findings [28].
Tegegne and Savidge, 2025	Gut microbiome metagenomics and clinical precision medicine	Highlighted the clinical potential of gut metagenomics for diagnosis, risk stratification, and precision medicine.	Clinical translation still requires standardization, external validation, and regulatory acceptance [13].

The use of explainable AI further strengthens biomarker discovery by clarifying the contribution of individual microbial taxa to model predictions. Nevertheless, differences in sequencing platforms, preprocessing pipelines, population characteristics, and analytical methods can reduce reproducibility across studies [19, 20].

Such improvements may support the translation of microbiome-derived biomarkers into reliable tools for disease screening and personalized healthcare.

2.6 Explainable Artificial Intelligence and Precision Medicine Potential of Microbiome-Derived Biomarkers

Prioritizing candidate microbial biomarkers can be achieved through the use of explainable AI techniques, including SHAP, LIME, permutation importance, feature attribution, and model-specific importance scores. These approaches enable scientists to see if microbial characteristics that are biologically plausible, or confounding patterns, are responsible for the model predictions. Regarding the research of metabolic syndrome, XAI is especially relevant, as the challenge is not only to predict whether a patient has metabolic syndrome but also to discover potentially interpretable microbial signatures related to obesity, insulin resistance, dyslipidaemia, inflammation and cardiometabolic risk. This is where XAI becomes crucial to shift from computational prediction to the discovery of biomarkers for translation [28]. Combining the technologies of shotgun metagenomics, AI-based prediction, and explainable AI is highly relevant to precision medicine. The goal of precision medicine is to identify people based on their biological risk profile and to facilitate tailored prevention or treatment measures. The gut microbiome metagenomics is further discussed as a significant player in clinical practice as it can facilitate better diagnostics, risk-stratification, therapeutic development, and patient-specific decision-making. Microbiome-derived biomarkers can provide an early signal in metabolic syndrome for those at higher risk for cardiometabolic disease. They can also assist with individual dietary guidance, lifestyle adjustments, probiotics and prebiotics, and tracking therapeutic outcomes [29, 30].

2.7 Research Gaps in Existing Literature

Despite significant advancements in metagenomics and artificial intelligence-driven microbiome research, several critical gaps remain that limit their effective

application in clinical settings. One of the major limitations is the inadequate integration of deep learning approaches with explainable artificial intelligence (XAI), which restricts the interpretability of predictive models and reduces their acceptance in healthcare decision-making [31].

Additionally, most existing studies employ generalized models and lack disease-specific optimization, particularly for complex conditions such as metabolic syndrome, thereby limiting their predictive accuracy and clinical relevance. Furthermore, although numerous studies have identified microbial associations with disease, there is insufficient translation of these findings into actionable precision medicine frameworks, such as personalized diagnostics and targeted therapeutic strategies [17].

Another key challenge lies in the intrinsic nature of metagenomic data, which is high-dimensional, sparse, and compositional, making it difficult for conventional and even some advanced machine learning methods to effectively capture underlying biological patterns. Finally, many AI-based models lack robust clinical validation and biological interpretability, creating a gap between computational predictions and real-world medical applications.

Addressing these limitations is essential for developing reliable, interpretable, and clinically applicable frameworks for microbiome-based disease prediction and precision medicine [19]. But there are some areas where the existing literature is lacking.

Many of the studies study the taxonomic differences, but they do not adequately perform functional profiling, since metabolic syndrome is a functional metabolic disorder. Secondly, numerous prediction models lack independent or external validation, which makes them less reliable in clinical practice. Third, the biological meaning of the black box AI models is not always easy to identify, as they often lack the required interpretation.

Fourth, very few studies combine feature extraction for metagenomic data, AI disease prediction, explainable AI and clinical relevance in an integrated way. The lack of these points underlines the necessity for an integrated, AI-based metagenomic strategy to discover biomarkers in metabolic syndrome [26].

2.8 Proposed Study Rationale

The proposed study falls in this gap. The study can be used to investigate the taxonomic composition of the microbial communities, as well as the changes in metabolic pathways linked with metabolic syndrome, from shotgun metagenomic data. It can use machine learning and deep learning models to evaluate if microbial characteristics can predict disease status. It integrates explanatory AI to help determine the microbial taxa, genes or pathways that are most predictive of the model. This integrated approach is a step beyond descriptive analysis of the microbiome and towards predictive, interpretable and clinically relevant biomarker discovery. This approach has clinical relevance because it is potentially useful for the early non-invasive screening and risk stratification. The microbiome differences in stool are appealing because they are non-invasive and might actually represent metabolic changes that precede clinical complications. Once confirmed, these markers may be incorporated into future decision support tools for screening, individualized intervention, and future disease monitoring and/or treatment response. This is especially relevant in metabolic syndrome where early treatment can help to lower the chances of developing diabetes, cardiovascular disease, fatty liver disease, and renal complications. To conclude, metabolic syndrome is a major cardiometabolic condition that is caused by insulin resistance, visceral adiposity, inflammation, dyslipidemia, and systemic metabolic abnormalities. The gut microbiome has a pivotal role in nutrient processing, immune function, maintenance of the epithelial barrier, production of microbial metabolites, and metabolic signaling in the host. Metabolic endotoxemia, energy harvest, short-chain fatty acid production, bile acid dysregulation and chronic inflammation are mechanisms by which dysbiosis might contribute to metabolic syndrome. Shotgun metagenomics is an efficient way to enrich for taxonomic and functional signatures of microbes and AI is a powerful tool for predictive modelling analysis of complex microbiome data. Explainable AI helps in that process by enabling the interpretable and biologically meaningful discovery of biomarkers. These methods will create a solid basis for clinically relevant, non-invasive microbial biomarkers for metabolic syndrome.

Chapter 3

Methodology

3.1 Study Design

The aim of this study was to explore the changes in gut microbial community and to find the microbial non-invasive markers associated with metabolic syndrome-related disease conditions in obesity. Shotgun metagenomic sequencing data from publicly available libraries in the NCBI database were downloaded.

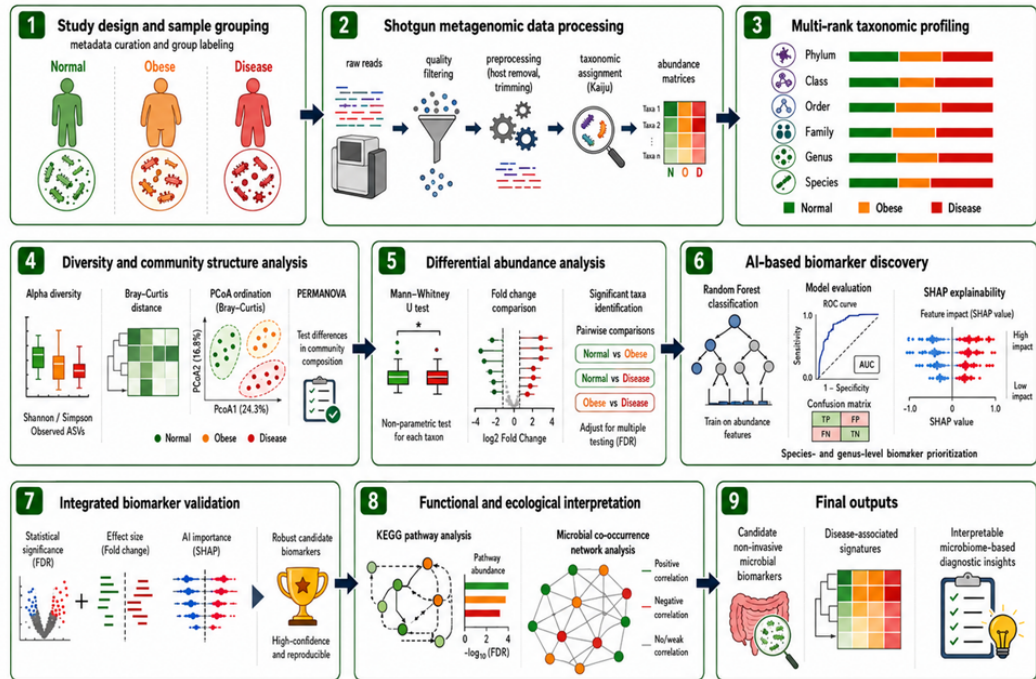


FIGURE 3.1: Methodological framework for microbiome analysis.

Three large clinical groups (Normal, Obese and Disease) were compared. Raw data

retrieval, QC, adapter and low-quality read trimming, metagenomic assembly, contig binning, genome quality assessment, taxonomy classification, abundance matrix construction, diversity analysis, differential abundance testing, machine-learning classification, explainable artificial intelligence interpretation, and final biomarker prioritisation. The whole study design is shown in figure 3.1.

3.2 Metagenomics Data Retrieval

Three NCBI BioProject records were retrieved for human gut microbiota, obesity and metabolic syndrome, respectively, to which shotgun metagenomic sequencing data were available. Data sets were chosen due to the presence of human gut or fecal metagenomic data associated with changes in microbial community associated with metabolic disease. The first dataset was downloaded from NCBI BioProject PRJNA429968 “Human gut microbiota and metabolic syndrome in Seniors.” The project can be accessed from: <https://www.ncbi.nlm.nih.gov/bioproject/429968>.

The second data set was acquired from NCBI BioProject – PRJNA417691 – “Mexican women with obesity and metabolic syndrome Raw sequence reads. This study examined the changes in gut microbiota in women with obesity and metabolic syndrome.

The project can be found at: <https://www.ncbi.nlm.nih.gov/bioproject/417691>. These BioProjects enabled sequencing data to be used for comparative analysis of gut microbial profiles between the Normal, Obese and Disease groups.

3.3 SRA Run Selection and Metadata Preparation

The accessions that were part of the selected BioProjects were retrieved from NCBI SRA. The final three groups were standardized from the various labels that were used, for instance, normal was grouped with Normal, obese with Obese, and disease with Disease. Samples labeled as “nonobese” or “control” were allocated to the Normal group as was done in the original description of the samples.

3.4 Raw Sequence Data Download

The raw sequencing reads were retrieved from the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) via the command line tools provided with Linux operating systems. The SRA Toolkit was used to retrieve sequencing data for the selected SRR accessions and convert them into FASTQ format.

3.5 Raw Read Quality Assessment using FastQC

Before further processing, the quality of the raw FASTQ files was evaluated using FastQC. Each sample was assessed overall for sequencing quality using FastQC to check for possible technical issues such as adapter contamination, low quality bases, abnormal GC content, sequence duplication and overrepresented sequences.

3.6 Adapter Removal and Quality Trimming using Trimmomatic

The adapter sequences were removed and the sequences were trimmed for quality using Trimmomatic. This step was necessary to eliminate Illumina adapter sequences, low quality bases and reads that were extremely short and may have impacted downstream assembly and classification. Unpaired reads were obtained by Trimmomatic. Following quality trimming, reads that became shorter than the minimum acceptable sequence length were excluded from subsequent analysis. This step prevented highly fragmented or uninformative sequences from contributing to the metagenomic assembly.

Only the high-quality paired reads were retained for the subsequent metagenomic assembly and classification procedures. The removal of sequencing artefacts and low-quality regions therefore increased the overall reliability of the processed dataset and reduced the possibility of false taxonomic assignments. Unpaired reads produced during trimming were not included in the downstream paired-end analysis [32, 33].

3.7 Post-Trimming Quality Assessment

After adapter removal and quality trimming, the processed sequencing reads were reassessed using FastQC. This second quality-control evaluation was performed to confirm that adapter contamination, low-quality nucleotide positions, and excessively short reads had been successfully removed. The post-trimming FastQC reports were compared with the initial quality reports to evaluate the improvement achieved during preprocessing. MultiQC was subsequently used to combine the individual FastQC outputs into a consolidated quality-control report. This enabled the quality metrics of all samples to be examined collectively and facilitated the identification of samples with unusual quality characteristics. The combined report provided a consistent overview of preprocessing performance across the complete dataset. Samples that satisfied the required quality criteria after trimming were retained for metagenomic assembly, whereas any sample showing persistent technical abnormalities would require further inspection before inclusion in the downstream analysis [32, 33].

3.8 Metagenomics Assembly

A metagenomic assembler was used to assemble high quality trimmed reads into contigs. MetaSPAdes was employed for metagenomic assembly since it has been developed for metagenomic assembly of complex microbial communities, and can accommodate the heterogeneous abundance patterns typically seen in shotgun metagenomic datasets. Assembler generated contigs, scaffolds and assembly graph files [34, 35].

3.9 Contig Binning

Metagenome-assembled genomes were reconstructed from assembled contigs by binning them into genome bins using MetaBat. Binning was done based on the composition of contig sequences and based on coverage related information. The aim of this step was to group contigs originating from the same microbial genome [36, 37].

3.10 Genome Bin Assessment using CheckM

CheckM was used to evaluate genome bins reconstructed. CheckM is a pipeline that assesses metagenome-assembled genomes by lineage-specific marker genes and estimates the completeness and contamination of genome assemblies. CheckM produced estimates of completeness, contamination, strain heterogeneity, of marker lineage and marker gene recovery for each of the bins. Completeness and contamination of the bins were used to determine more reliable bins [38].

3.11 Reads Level Taxonomic Classification Using Kaiju

Taxonomic profiling at the read level was done with Kaiju. Kaiju categorises metagenomic reads by similarity of proteins and this enables the detection of microbial taxa with high sensitivity in complex metagenomic samples. Taxonomic profiles were created for each sample using kaiju. Raw Kaiju output included read-level classification results and tabular summary files included abundance information at various taxonomic levels.

Downstream abundance-based analysis was conducted on the tabular outputs. Taxonomic summaries were obtained at the phylum, class, order, family, genus and species level [39]. Overall, 145 samples had Kaiju-based taxonomic profiles. The number taxa identified varied by taxonomic level, with the largest taxonomic levels (phylum) having fewer taxa, and the smallest taxonomic level (species-level profiling) having the most taxa identified.

3.12 Taxonomic Composition Analysis

Taxonomic structure of the microbial community was compared between the Normal, Obese and Disease groups to understand the structure of microbes. Taxonomic composition plots were prepared at various ranks by using relative abundance matrix. The most abundant taxa were displayed separately from each other and lower abundance taxa were combined in an “Others” category for readability.

Composition's analysis was carried out at phylum, family, genus and species level. The abundance pattern was analyzed on both the sample and group level. Sample-wise plots indicated variation between samples while group-wise plots indicated the average taxonomic structure of each clinical group. These visualizations gave a general overview of the microbial community changes between obesity and disease status.

3.13 Wilcoxon Rank-Sum Test

Wilcoxon rank sum test (Mann–Whitney U test) was used to compare alpha diversity values between each pair of samples. This is a non-parametric test due to the fact that abundance and diversity data of microbiomes are often not normally distributed [40]. Pairwise group comparisons were made between the Normal and Disease group, the Obese and Disease group, and the Normal and Obese group. The p value was used for every alpha diversity index to assess the statistical significance in the differences between groups. The P value of <0.05 was considered statistically significant. The significance levels were as follows: Not significant ($p > 0.05$); significant ($p < 0.05$); highly significant ($p < 0.01$); very highly significant ($p < 0.001$); and extremely significant ($p < 0.0001$).

3.14 Alpha Diversity Analysis

Alpha diversity analysis is used to assess the diversity, richness, dominance and evenness of the microbial community within a sample. Taxonomic abundance matrices were used to compute the alpha diversity indices at various ranks. Observed taxa, Shannon diversity index, Simpson diversity index, inverse Simpson index, Pielou evenness, dominance index and Chao1 richness estimate were used as diversity metrics [41, 42]. Number of detected taxa was observed in each sample. The Shannon diversity indicated both richness and evenness. Both Simpson and inverse Simpson indices were used to measure dominance and diversity structure. The evenness Pielou was used to assess the evenness with which the taxa were distributed among the samples, whereas the Chao1 was used to estimate the richness of the samples, taking into account rare taxa. The average number of species in a

given sample, or group-wise alpha diversity, was compared between the Normal, Obese and Disease samples [42]. The distribution of diversity values was displayed using violin plots, boxplots and sample-level points.

3.15 Beta Diversity Analysis

The differences among the microbial community composition of the samples were evaluated using beta diversity analysis. Relative abundance profiles were used for calculating Bray–Curtis dissimilarity. Bray–Curtis dissimilarity was chosen as it is commonly applied in ecological and microbiome studies and takes into account abundance-based differences between microbial communities. Genus and species level Bray–Curtis distance matrices were produced. These matrices were employed for ordination analysis and statistical analysis of the difference between microbial community groups [41].

3.16 Principal Coordinate Analysis

Normal, Obese and Disease groups were plotted using PCA with Bray–Curtis dissimilarity matrix. Samples were colored by clinical group and each point in the PCoA plot was a sample. The first two principal coordinate axes were plotted to give an overview of the largest sources of variation in the microbial community structure. The emphasis on genus and species level was due to the fact that these taxonomic levels offer biologically meaningful resolution for microbial biomarker discovery [43].

3.17 PERMANOVA Analysis

Microbial community composition was statistically tested for significant difference among groups by performing a permutational multivariate analysis of variance (PERMANOVA). Bray–Curtis distance matrices were analysed using PERMANOVA with 999 permutations. A p-value < 0.05 was deemed statistically significant. Finally, Bray–Curtis PERMANOVA indicated statistically significant group level separation at the genus and species level with $p = 0.001$. [44].

3.18 Differential Abundance Analysis

Differential abundance analysis was carried out to determine taxa that were significantly different between groups. Wilcoxon rank-sum test/Mann–Whitney U test was used for taxa-wise comparisons. The pairwise comparisons were Normal vs. Disease, Obese vs. Disease, and Normal vs. Obese. The primary biological comparison was given a special focus: that of Normal versus Disease and Obese versus Disease. The mean abundance, median abundance, test statistic, p-value, false-discovery rate (FDR) adjusted p-value, log2 fold-change, detection frequency, and category of significance were computed for each taxon. The presence or absence of an enrichment or depletion of a taxon in the Disease group was determined by calculating a log2 fold change between the Disease and Reference groups. Taxa which were statistically significant ($p < 0.05$) were identified. Multiple testing correction was made using Benjamini–Hochberg procedure. Differentially abundant taxa were also assessed for their biological relevance and the consistency with the machine learning-based biomarker results [45–47].

3.19 Machine Learning Based Classification using XGBoost and Explainable AI

For disease biomarker discovery, XGBoost classification, which is effective on structured biological data, was used to capture the complex nonlinear relationships among microbial features. The genus- and species-level abundance matrices were used to train the XGBoost models. If a comparison was made for disease, then the Disease group was considered the positive class. This enabled the interpretation of models to be directly associated with prediction of disease. Positive model contributions were features that helped classify Disease and negative model contributions were features that help classify the non-disease group. The performance of models was assessed based on ROC-AUC, confusion matrix, classification report, and CV AUC [48, 49]. Models that performed well on the disease data were chosen to analyze via XAI. XGBoost models were used to interpret using SHAP analysis and to determine the most influential microbial taxa in disease

classification. SHAP values were calculated for each microbial feature in each sample. This study, positive SHAP values were considered as taxa contributing to Disease prediction, and negative SHAP values were considered as taxa contributing away from Disease prediction. Taxa were ranked according to their overall importance in the classification model based on mean absolute SHAP values. For primary disease-associated comparisons (Normal vs Disease, Obese vs Disease), SHAP analysis was carried out. The resulting SHAP values were then used to determine taxa that have strong predictive importance. SHAP importance was not the only criterion used for biomarker final selection, however, but was combined with abundance direction and statistical significance [50, 51].

3.20 Candidate Biomarker Prioritization

Explainable AI results were combined with differential abundance to carry out final biomarker prioritization. A taxon was selected as a biomarker for a disease if it had: high SHAP importance, disease-enriched abundance, statistically significant Wilcoxon/Mann–Whitney p-value, log2 fold-change sign consistent across all samples, and acceptable detection frequency across all samples. The final integrated table listed for each candidate biomarker, taxonomic rank, taxon name, comparison group, mean absolute SHAP value, mean SHAP value, disease enrichment status, Wilcoxon/Mann–Whitney p-value, FDR p-value, mean abundance in each group, median abundance in each group, log2 fold-change, detection frequency, and significance category. By combining these approaches, the risk of relying solely on the importance of these biomarkers was minimized. Final microbial biomarkers were thus validated both statistically and by machine learning interpretability.

3.21 Performance Matrices Evaluation

The machine-learning models were tested with common classification indicators based on the confusion matrix. The samples were split into two classes for each binary comparison, with the Disease group used as the positive class for disease associated comparisons. The class labels were then compared against the model

predictions to get the number of True Positives, True Negatives, False Positives and False Negatives.

True positive was defined as Disease samples that were correctly classified as Disease and true negative was defined as non-Disease samples correctly classified as non-Disease. After the samples were classified, the number of samples that were classified as Disease when they were actually non-Disease were referred to as False Positive and those classified as non-Disease when they were actually Disease were referred to as False Negative.

The overall percentage of samples correctly classified was used as the measure of accuracy. The percentage of the Disease samples that were predicted as Disease was measured [52, 53]. For recall, or sensitivity, the percentage of the actual Disease samples correctly identified by the model will be measured. Specificity was used for measuring how many times non-Disease were correctly predicted as non-Disease. The F1 score was derived as the harmonic mean of precision and recall. This metric was chosen because it offers a fair assessment of the model performance, particularly if class distribution is not uniform. The ROC curve was created by varying the classification threshold and plotting the resulting sensitivity (true positive rate) and false positive rate. The area under the ROC curve was used to evaluate the overall discriminative ability of the model. A ROC-AUC close to 1 implied good classification performance while a close to 0.5 was a weak discrimination. Five-fold cross validation was used to test the stability of the model performance [60]. In this method, the data set was split into five parts. For each iteration four subsets were used for training, and one subset for testing. Abundance matrices were both genus and species level and were evaluated independently of each other. The primary binary comparisons were for Normal vs Disease, Obese vs Disease, and Normal vs Obese [54].

3.22 Co-occurrence Analysis

The top genera were selected in each group and pairwise correlation-based associations were computed to create group-wise co-occurrence networks. Each node was a genus and each edge retained the positive and/or negative association. Changes

in the complexity of microbial associations were assessed by comparing network measures (nodes, edges, average degree, density, positive and negative interactions) between groups [55].

3.23 Data Visualization

The final visualizations have been created with Python. Publication style figures were made to aid the interpretation of the taxonomic composition, microbial diversity, beta diversity, differential abundance, machine-learning performance, SHAP interpretation, and integrated biomarker prioritization. The figures were tabulated by analysis type and printed out in high resolution for presentation in the thesis. All plots were checked to ensure that they were well labeled, the legends were understandable, the groups were separated in an appropriate manner, and the formatting was consistent.

3.24 Computational Environment

The bioinformatics preprocessing steps were all done in Ubuntu/WSL. SRA Toolkit, FastQC, MultiQC, Trimmomatic, metaSPAdes, Binning tools, CheckM, GTDB-Tk, and Kaiju were the main tools used in the analysis. Analysis of downstream data has been carried out in Python using the following libraries: pandas, numpy, scipy, scikit-bio, scikit-learn, xgboost, shap, matplotlib, and seaborn. The data that was obtained from each workflow was stored in different folders: Raw Data, QC Reports, Trimmed Reads, Assembly Outputs, Genome Bins, Taxonomic Classification Files, Abundance Matrices, Statistical Results, Machine Learning Results, SHAP Results, and Final Figures. This organization provided the reproducibility and ensured that all intermediate and final outputs were kept for validation and reporting purposes.

Chapter 4

Results

4.1 FastQC Report

The representative SRA read file passed all major quality assessment modules from the FastQC quality assessment.

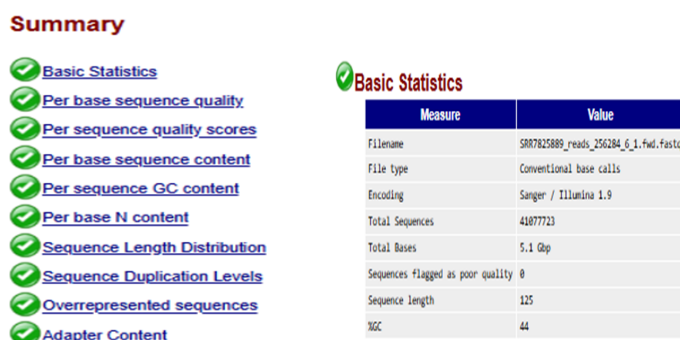


FIGURE 4.1: General FastQC summary of SRR782588.

It includes basic statistics, per-base sequence quality, per-sequence quality scores, per-base sequence content, GC content, N content, sequence length distribution, duplication levels, overrepresented sequences and adapter content. Basic statistics as shown in figure 4.1 demonstrated that a conventional base call FASTQ file was present and encoded in the Sanger/Illumina 1.9 format, that contained approximately 41.08 million reads, 5.1 Gbp of total bases, a fixed read length of 125 bp and 44% GC content, with none of the reads being flagged as poor quality. Per-base sequence quality was mostly in the high-quality range throughout the length of the read and the distribution of per-sequence quality scores was clustered at

high Phred scores.

The base composition exhibited relatively constant nucleotide proportions at each position and the GC distribution was unimodal as expected. There was essentially no per-base N content (little ambiguity of base calling) and a reasonable distribution of sequence lengths (125 bp) was confirmed.

4.2 Adapter Removal and Quality Trimming using Trimmomatic

To remove low-quality sequences and adapter-contaminated reads from selected SRA read datasets, they were processed with a trimmomatic-based pipeline before metagenomic analysis was performed as shown in figure 4.2.

TRIMMOMATIC RESULTS FOR SRR7825897_reads (object 256284/93/1)

Input Read Pairs	55899942	(100.0%)	
Both Surviving	50991383	(91.2%)	
Forward Only Surviving	3869909	(6.9%)	
Reverse Only Surviving	593379	(1.1%)	
Dropped	445271	(0.8%)	

FIGURE 4.2: Summary of adapter removal and read-quality trimming using Trimmomatic.

The proportion of reads retained after QC was relatively high for all samples displayed, suggesting that the raw sequencing data sets were appropriate for further processing by assembly and taxonomic profiling. A small percentage of discarded reads further suggests that only a small percentage of sequences did not meet trimming criteria.

4.3 Metagenomics Assembly Summary

A total of 43,921 contigs of about 45.67 Mbp in total were retrieved from the metagenomic assembly. No contigs exceeded 100,000 bp. The assembly showed

an N50 value of 24,310 bp, N75 of 7604 bp, L50 of 1092, and L75 of 12,453, with an average GC content of approximately 46.16%. The contig rank was plotted against the cumulative length curve which was used to show the distribution of sequence length of assembled sequences by contig as shown in figure 4.3. These assembly statistics indicate that the generated contigs provided sufficient sequence continuity for subsequent taxonomic and functional analyses. The observed GC content also reflects the diverse genomic composition of the microbial community.

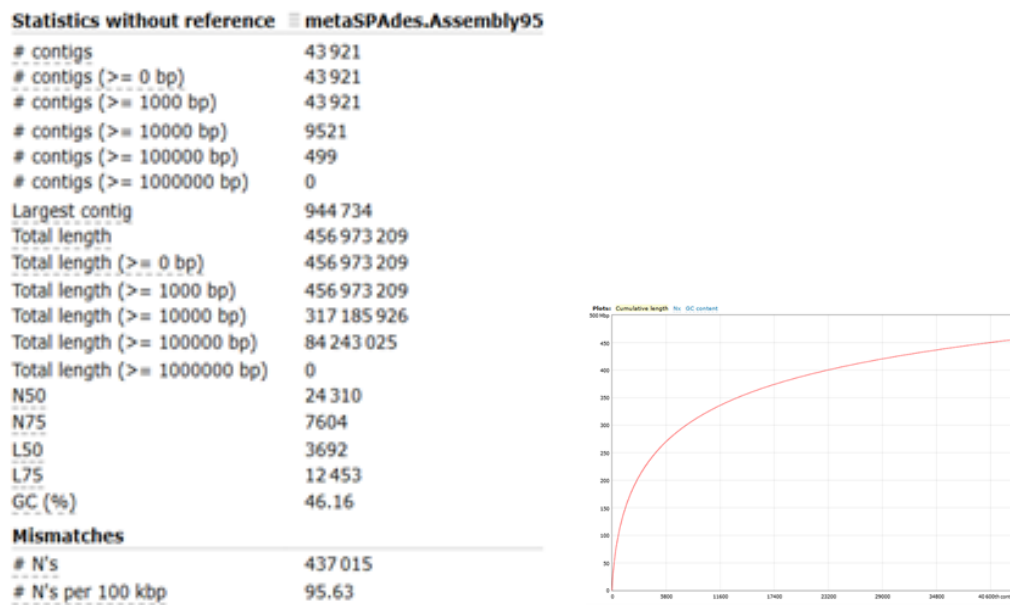


FIGURE 4.3: Metagenomic assembly statistics of assembled contigs.

4.4 Contig Binning Performance

The number of inputs contigs, binned contigs, and reconstructed genome bins were found to vary for selected metagenomic samples using contig binning analysis. The bar plot revealed that there was a high number of contigs assembled for each sample and that a reasonable number of these contigs were assembled and binned into the genome bins. Binning recovery ranged from 45% to nearly 70%, varied by sample. The number of reconstructed genome bins also varied among the datasets with some samples containing over 200 genome bins and others only a handful of reconstructed genome bins (less than 100) as shown in figure 4.4.

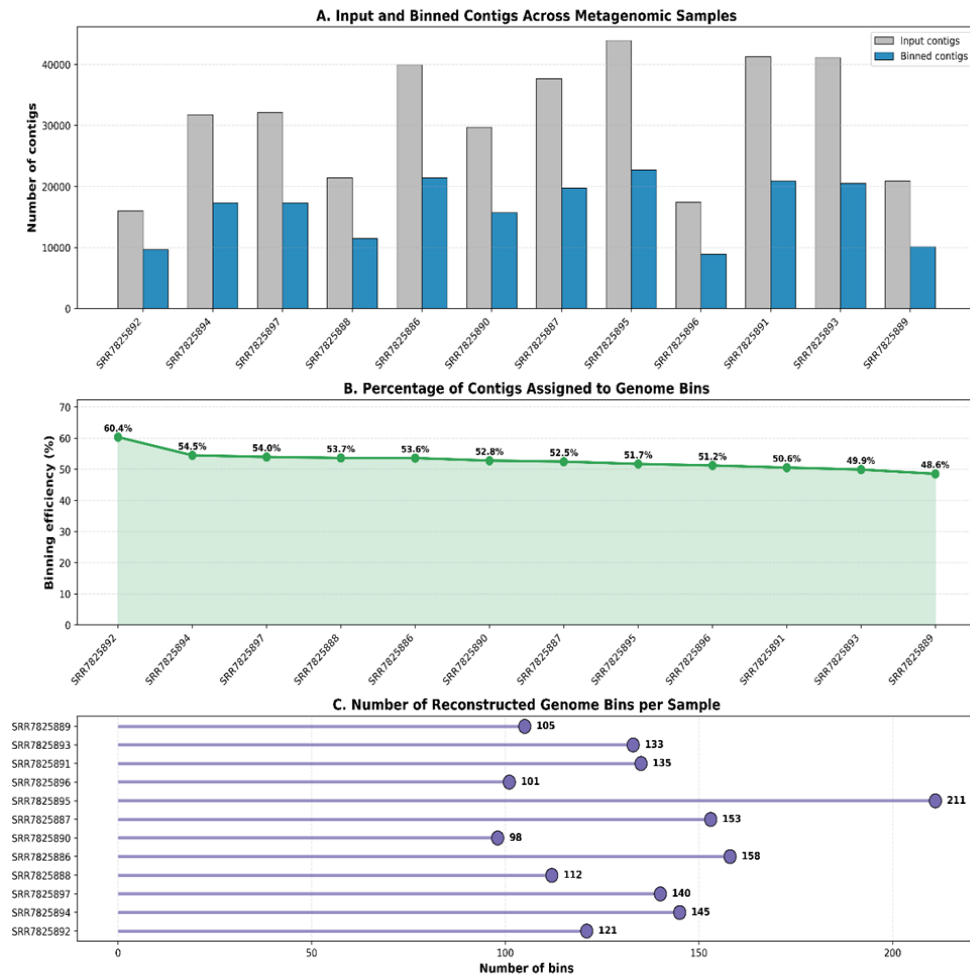


FIGURE 4.4: Summary of metagenomic contig binning performance across selected samples.

4.5 Genome Bin Assessment using CheckM

Selected metaSPAdes assemblies were analyzed based on CheckM for genome-bin quality assessment focusing on marker-copy distribution, missing markers, heterogeneity, and contamination levels. The heat map as shown in figure 4.5 revealed that the sampled bins had large amounts of multi-copy marker signals, and that there was good coverage within the contamination category for the assemblies presented. The number of missing marker signals was relatively small and the number of heterogeneity signals varied per sample. In total, the CheckM profile revealed that genome bins reconstructed from the sample needed quality filtering prior to being used for subsequent genome-resolved analysis as valid metagenome-assembled genomes.

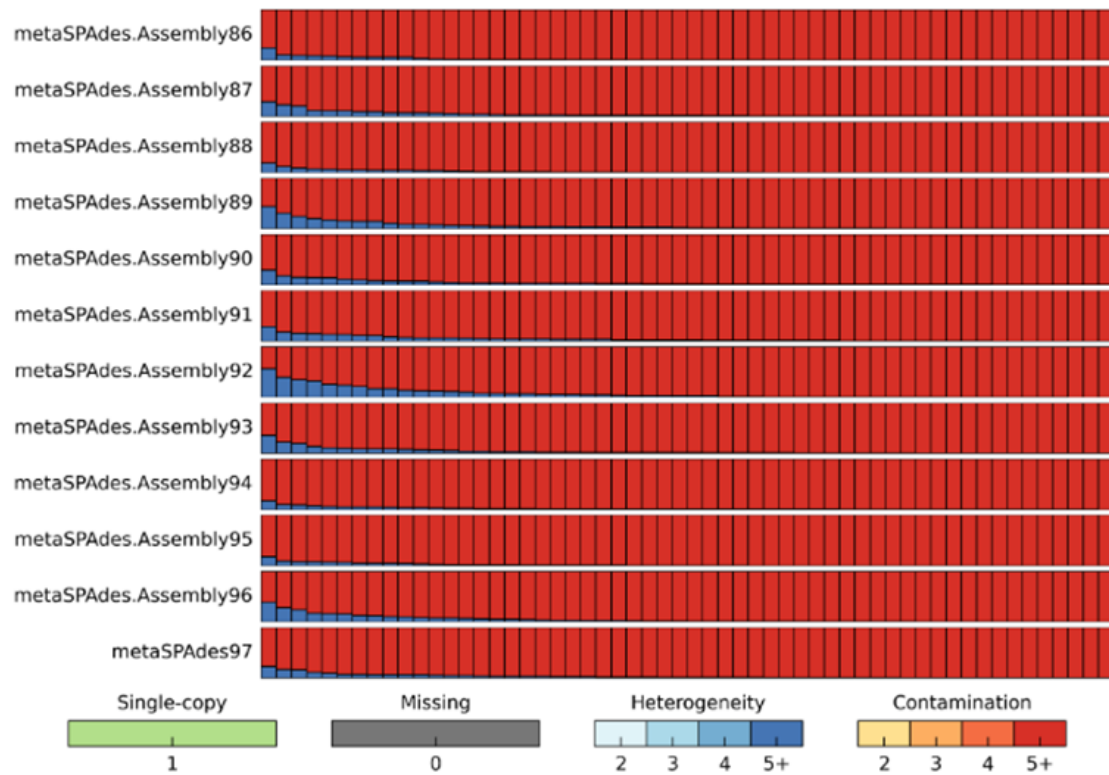


FIGURE 4.5: CheckM-based genome completeness assessment of recovered metaSPAdes assemblies, showing single-copy marker distribution, missing markers, heterogeneity, and contamination levels across assemblies 86–97.

4.6 Taxonomic Composition Analysis

Overall, the taxonomic composition analysis showed that the 145 metagenomic samples contained a diverse gut microbial community distributed across 31 phyla, 52 classes, 114 orders, 203 families, 388 genera, and 535 species. The community was dominated at higher taxonomic ranks by *Bacillota*, mainly represented by *Clostridia*, with major contributions from *Lachnospirales*, *Eubacteriales*, *Lachnospiraceae*, and *Oscillospiraceae* at subsequent ranks. At lower taxonomic levels, the profiles became increasingly diverse, with genera such as *Faecalibacterium*, *Blautia*, *Bacteroides*, *Bifidobacterium*, *Anaerostipes*, *Anaerobutyricum*, *Segatella*, and *Dorea* contributing substantially to the overall composition. Species-level profiling revealed the greatest variation among samples, with abundant taxa such as *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *[Ruminococcus] torques*, *Dorea longicatena*, *Segatella copri*, *Blautia obeum*, *Faecalibacterium prausnitzii*, and *Bifidobacterium longum*. The overall pattern shows that higher taxonomic ranks

were dominated by a few major bacterial groups, whereas genus- and species-level profiles revealed stronger sample-specific variation.

TABLE 4.1: Top ten phyla ranked according to their mean relative abundance across all study samples.

Rank	Phylum	Mean relative abundance (%)
1	<i>Bacillota</i>	57.511
2	<i>Bacteroidota</i>	13.611
3	<i>Actinomycetota</i>	5.558
4	<i>Pseudomonadota</i>	0.745
5	<i>Chloroflexota</i>	0.700
6	<i>Thermotogota</i>	0.435
7	<i>Verrucomicrobiota</i>	0.393
8	<i>Methanobacteriota</i>	0.111
9	<i>Thermodesulfobacteriota</i>	0.095
10	<i>Mycoplasmata</i>	0.075

At the phylum level as shown in table 4.1, the taxonomic profile showed that the microbial community was mainly dominated by *Bacillota*, which was detected in all 145 samples and represented the highest mean relative abundance across the dataset. *Bacteroidota* formed the second major phylum and was also widely distributed among samples, followed by *Actinomycetota*, *Pseudomonadota*, *Chloroflexota*, *Verrucomicrobiota*, and other low-abundance phyla. The stacked bar plot shows that most samples were characterized by a large *Bacillota* fraction, while the contribution of *Bacteroidota* and other phyla varied among samples. This indicates that the overall microbial community structure was dominated by a few major bacterial phyla, with additional minor phyla contributing to sample-specific variation. The dominance of *Bacillota* and *Bacteroidota* reflects the major contribution of these phyla to the overall gut microbial structure. Variations in the less abundant phyla may indicate individual- or condition-specific differences among the study samples.

TABLE 4.2: Top ten microbial classes ranked according to their mean relative abundance across all study samples.

Rank	Class	Mean relative abundance (%)
1	<i>Clostridia</i>	49.063
2	<i>Bacteroidia</i>	13.060
3	<i>Actinomycetes</i>	3.006
4	<i>Bacilli</i>	2.888
5	<i>Coriobacteriia</i>	2.393
6	<i>Erysipelotrichia</i>	1.983
7	<i>Negativicutes</i>	1.772
8	<i>Dehalococcoidia</i>	0.520
9	<i>Gammaproteobacteria</i>	0.502
10	<i>Thermotogae</i>	0.435

At the class level as shown in table 4.2, the dominant bacterial class was *Clostridia*, which accounted for the major portion of classified reads across nearly all samples. This pattern matches the strong dominance of *Bacillota* observed at the phylum level. *Bacteroidia* formed the second major class, corresponding to *Bacteroidota*, while *Actinomycetes*, *Bacilli*, *Coriobacteriia*, *Erysipelotrichia*, *Negativicutes*, *Verucomicrobiia*, *Gammaproteobacteria*, and *Thermotogae* were detected at lower relative abundances. The class-level plot shows that although *Clostridia* remained the most consistent and abundant group, several other classes contributed variably across individual samples, producing visible differences in community composition.

TABLE 4.3: Top ten microbial orders ranked according to their mean relative abundance across all study samples.

Rank	Order	Mean relative abundance (%)
1	<i>Lachnospirales</i>	20.602
2	<i>Eubacteriales</i>	16.639
3	<i>Bacteroidales</i>	13.059
4	<i>Lactobacillales</i>	2.689
5	<i>Bifidobacteriales</i>	2.687
6	<i>Erysipelotrichales</i>	1.983
7	<i>Coriobacteriales</i>	1.919
8	<i>Veillonellales</i>	1.134
9	<i>Peptostreptococcales</i>	0.736
10	<i>Dehalococcoidales</i>	0.520

At the order level as shown in table 4.3, presents the ten most abundant microbial orders identified across all study samples on the basis of their mean relative abundance. Among these orders, *Lachnospirales* was the most dominant, with a mean relative abundance of 20.602%. It was followed by *Eubacteriales* and *Bacteroidales*, which exhibited mean relative abundances of 16.639% and 13.059%, respectively. Collectively, these three dominant orders accounted for approximately 50.300% of the overall microbial composition at the order level. This finding indicates that a substantial proportion of the gut microbial community was concentrated within a relatively small number of highly abundant bacterial orders. The predominance of *Lachnospirales* and *Eubacteriales* reflects the strong representation of bacterial groups commonly associated with the anaerobic environment of the human gastrointestinal tract. The next most abundant orders were *Lactobacillales* and *Bifidobacteriales*, with mean relative abundances of 2.689% and 2.687%, respectively. Although their abundances were considerably lower than those of the three dominant orders, both remained consistently represented within the overall microbial profile. *Erysipelotrichales* and *Coriobacteriales* showed mean relative abundances of 1.983% and 1.919%, respectively.

TABLE 4.4: Top ten microbial families ranked according to their mean relative abundance across all study samples.

Rank	Family	Mean relative abundance (%)
1	<i>Lachnospiraceae</i>	20.555
2	<i>Oscillospiraceae</i>	13.926
3	<i>Bacteroidaceae</i>	5.813
4	<i>Prevotellaceae</i>	4.467
5	<i>Bifidobacteriaceae</i>	2.687
6	<i>Streptococcaceae</i>	2.033
7	<i>Coprobacillaceae</i>	1.816
8	<i>Coriobacteriaceae</i>	1.688
9	<i>Rikenellaceae</i>	1.365
10	<i>Veillonellaceae</i>	1.134

At the family level as shown in table 4.4, the dominant families included *Lachnospiraceae* and *Oscillospiraceae*, followed by *Prevotellaceae*, *Bacteroidaceae*, *Bifidobacteriaceae*, *Streptococcaceae*, *Coprobacillaceae*, *Coriobacteriaceae*, *Rikenellaceae*, *Veillonellaceae*, *Akkermansiaceae*, and *Clostridiaceae*.

TABLE 4.5: Top ten microbial genera ranked according to their mean relative abundance across all study samples.

Rank	Genus	Mean relative abundance (%)
1	<i>Faecalibacterium</i>	6.235
2	<i>Blautia</i>	4.765
3	<i>Bacteroides</i>	4.167
4	<i>Bifidobacterium</i>	2.672
5	<i>Anaerostipes</i>	2.080
6	<i>Streptococcus</i>	2.013
7	<i>Mediterraneibacter</i>	1.930
8	<i>Segatella</i>	1.928
9	<i>Ruminococcus</i>	1.919
10	<i>Anaerobutyricum</i>	1.758

The family-level plot showed greater diversity compared with higher ranks, as multiple families contributed measurable proportions within each sample. *Lachnospiraceae* was consistently abundant across the dataset, while *Bacteroidaceae* and *Prevotellaceae* showed stronger variation between samples. This level highlights a broader distribution of gut microbial families and provides a clearer view of differences among individual metagenomic profiles. Together, these three families accounted for approximately 40.294% of the overall microbial composition at the family level, indicating their major contribution to the structure of the gut microbiome. *Prevotellaceae* was the fourth most abundant family, with a mean relative abundance of 4.467%, while *Bifidobacteriaceae* and *Streptococcaceae* contributed 2.687% and 2.033%, respectively. The remaining families, including *Coprobacillaceae*, *Coriobacteriaceae*, *Rikenellaceae*, and *Veillonellaceae*, exhibited

comparatively lower mean relative abundances ranging from 1.134% to 1.816%. Collectively, the ten leading families represented approximately 55.484% of the total microbial abundance, demonstrating that more than half of the family-level microbial community was concentrated within these dominant taxa.

At the genus level as shown in table 4.5, presents the ten most abundant microbial genera identified across all study samples according to their mean relative abundance. Among these genera, *Faecalibacterium* was the most dominant, with a mean relative abundance of 6.235%, followed by *Blautia* at 4.765% and *Bacteroides* at 4.167%. Together, these three genera accounted for approximately 15.167% of the overall microbial composition at the genus level, indicating their substantial contribution to the gut microbial community. *Bifidobacterium* and *Anaerostipes* showed mean relative abundances of 2.672% and 2.080%, respectively, while *Streptococcus* contributed 2.013%. The remaining genera, including *Mediterraneibacter*, *Segatella*, *Ruminococcus*, and *Anaerobutyricum*, exhibited relatively similar mean abundances ranging from 1.758% to 1.930%. Collectively, the ten dominant genera represented approximately 29.467% of the total microbial abundance, suggesting that the genus-level gut microbiome consisted of several prominent taxa together with a large proportion of less abundant genera.

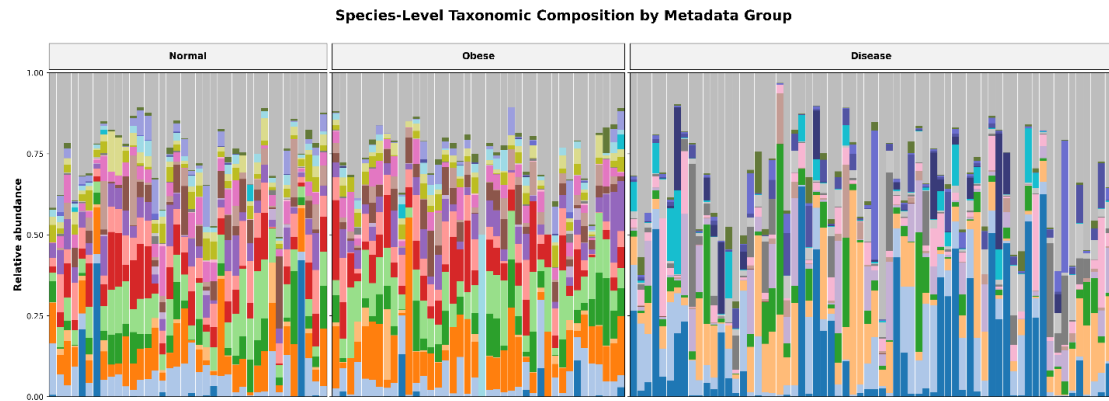
TABLE 4.6: Top ten microbial species ranked according to their mean relative abundance across all study samples.

Rank	Species	Mean relative abundance (%)
1	<i>Anaerostipes hadrus</i>	2.071
2	<i>Anaerobutyricum hallii</i>	1.758
3	<i>[Ruminococcus] torques</i>	1.422
4	<i>Faecalibacterium prausnitzii</i>	1.383
5	<i>Faecalibacillus intestinalis</i>	1.348
6	<i>Segatella copri</i>	1.196
7	<i>Dorea longicatena</i>	1.043
8	<i>Blautia obeum</i>	0.955
9	<i>Phocaeicola vulgatus</i>	0.905
10	<i>Bifidobacterium longum</i>	0.738

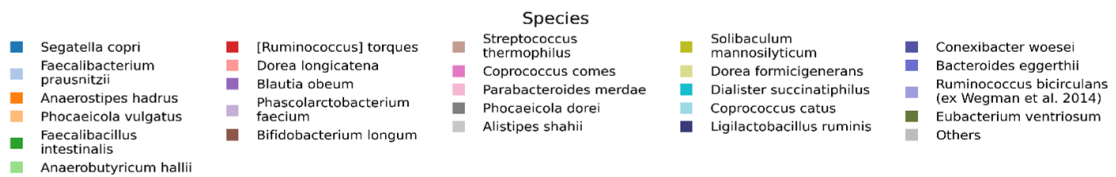
At the species level as shown in table 4.6, presents the ten most abundant microbial species identified across all study samples according to their mean relative abundance. Among these species, *Anaerostipes hadrus* was the most abundant, with a mean relative abundance of 2.071%, followed by *Anaerobutyricum hallii* at 1.758% and *[Ruminococcus] torques* at 1.422%. *Faecalibacterium prausnitzii* and *Faecalibacillus intestinalis* also contributed substantially to the overall species-level profile, with mean relative abundances of 1.383% and 1.348%, respectively. The remaining dominant species included *Segatella copri* at 1.196%, *Dorea longicatena* at 1.043%, *Blautia obeum* at 0.955%, *Phocaeicola vulgatus* at 0.905%, and *Bifidobacterium longum* at 0.738%. Collectively, these ten species accounted for approximately 12.819% of the overall microbial abundance, indicating that the species-level gut microbiome was highly diverse and that a considerable proportion of the community consisted of numerous lower-abundance species. The predominance of *Anaerostipes hadrus*, *Anaerobutyricum hallii*, and *Faecalibacterium prausnitzii* indicates the strong representation of anaerobic gut microorganisms within the study population. However, because these values were calculated as averages across normal, obese, and disease samples, they describe the overall taxonomic composition rather than condition-specific changes.

The species-level taxonomic composition grouped by metadata category showed visible differences in microbial community structure among the Normal, Obese, and Disease groups as shown in figure 4.6. The metadata summary indicated that the dataset contained 72 Disease samples, 44 Normal samples, and 40 Obese samples. In the stacked bar plot, Normal and Obese samples showed a more mixed and relatively balanced species distribution, with several species contributing moderate proportions across samples, including *Segatella copri*, *Faecalibacterium prausnitzii*, *Anaerostipes hadrus*, *Phocaeicola vulgatus*, *Faecalibacillus intestinalis*, *Anaerobutyricum hallii*, *[Ruminococcus] torques*, *Dorea longicatena*, *Blautia obeum*, and *Bifidobacterium longum*. In contrast, the Disease group showed stronger sample-to-sample variation, with some samples dominated by fewer species and a larger proportion of reads assigned to Others or selected taxa. The statistical screening further showed that significant taxonomic differences were detected across multiple ranks, with 2 significant phyla out of 22 tested taxa, 6 significant families out

of 99 tested taxa, 12 significant genera out of 191 tested taxa, and 11 significant species out of 242 tested taxa at $p < 0.05$. Overall, the figure shows that species-level composition varies across the metadata groups, with the clearest taxonomic differentiation appearing at the genus and species levels.



(A) Species-level taxonomic composition across metadata groups.



(B) Summary of statistically significant taxa across taxonomic levels.

FIGURE 4.6: Species-level taxonomic composition and summary of statistically significant taxa across metadata groups. Stacked bar plots show the relative abundance of dominant species across Normal, Obese, and Disease groups, while the summary plot shows the number of significant taxa detected at different taxonomic levels using a threshold of $p < 0.05$.

4.7 Wilcoxon Rank-Sum Test

4.7.1 Normal vs Obese Comparison

The Mann–Whitney U test-based differential taxa analysis as shown in figure 4.7 between the Normal and Obese groups identified significant differences across multiple taxonomic ranks. At the phylum level, *Actinomycetota* and *Verrucomicrobiota* showed significant differences between the two groups, with *Actinomycetota* showing comparatively higher abundance in the Obese group, while *Verrucomicrobiota* remained low in both groups but differed significantly.

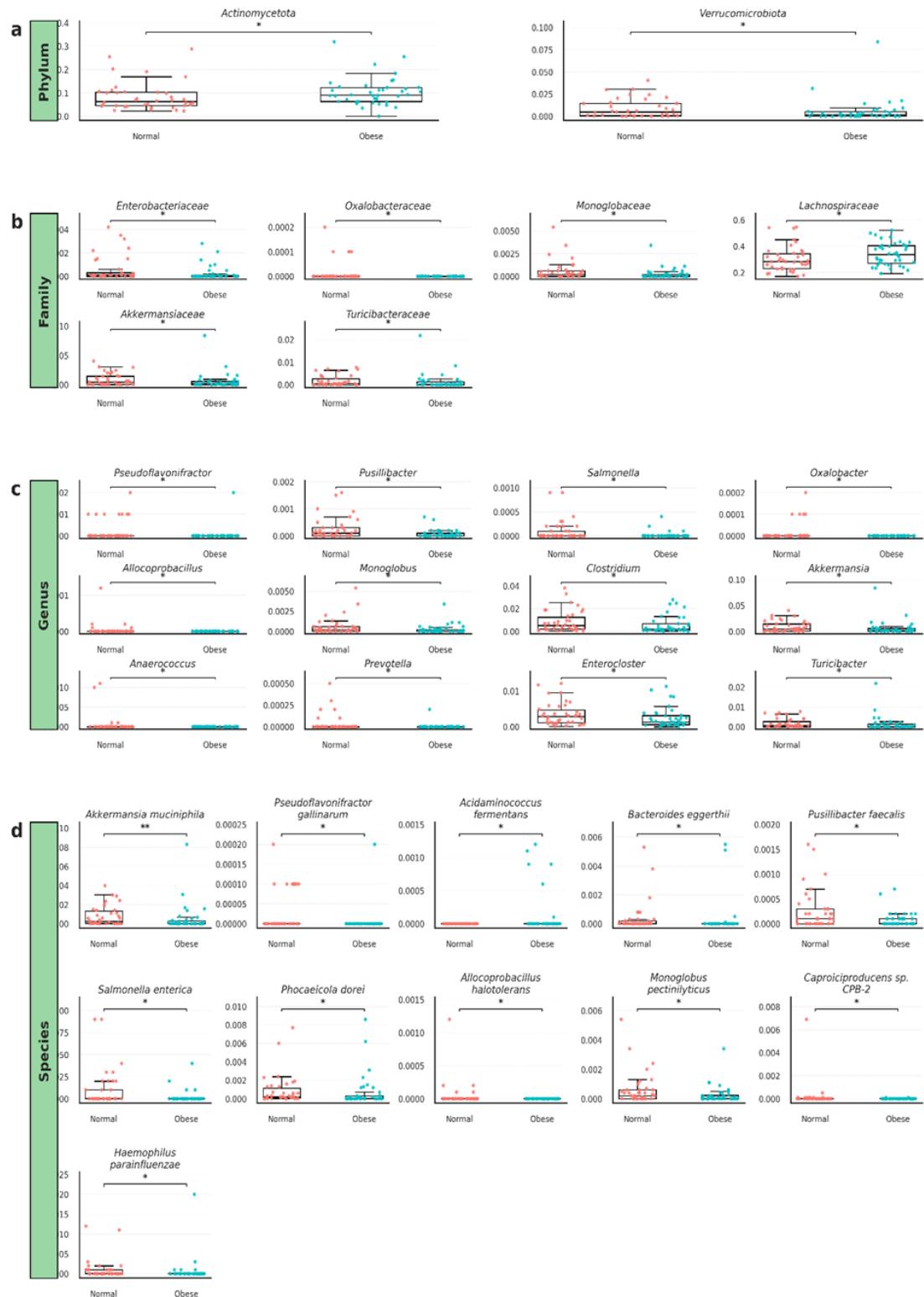


FIGURE 4.7: Mann–Whitney U test-based differential abundance analysis between the Normal and Obese groups across multiple taxonomic levels. Panel present the significantly different taxa identified at the phylum, family, genus, and species levels, respectively. Red boxplots represent the Normal group, whereas blue boxplots represent the Obese group. Each point corresponds to the relative abundance of a taxon in an individual sample. The central line within each box represents the median, the box indicates the interquartile range, and the whiskers illustrate the variability of the abundance values.

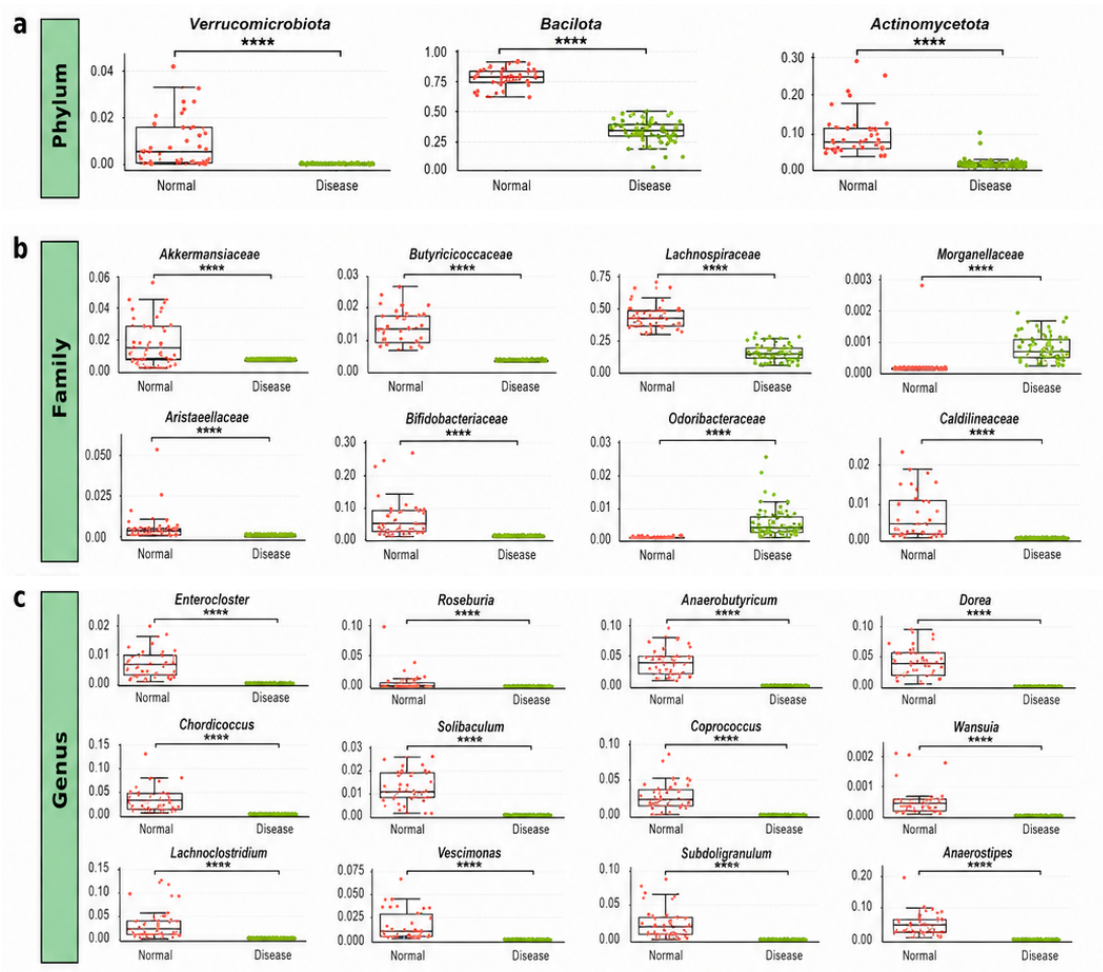
At the family level, six families were significantly different, including *Enterobacteriaceae*, *Oxalobacteraceae*, *Monoglobaceae*, *Lachnospiraceae*, *Akkermansiaceae*, and *Turicibacteraceae*. Among these, *Lachnospiraceae* showed relatively higher abundance and wider distribution in both groups compared with the other significant families. At the genus level, twelve genera were significantly different, including *Pseudoflavonifractor*, *Pusillibacter*, *Salmonella*, *Oxalobacter*, *Allocoprobaillus*, *Monoglobus*, *Clostridium*, *Akkermansia*, *Anaerococcus*, *Prevotella*, *Enterocloster*, and *Turicibacter*. Most of these genera showed low median abundance, with individual samples showing higher outlier values. At the species level, *CPB-2*, and *Haemophilus parainfluenzae*. The species-level plots showed generally low abundance for most taxa, with several taxa appearing as sparse or sample-specific signals rather than uniformly abundant across all samples.

4.7.2 Normal vs Disease Comparison

The Mann–Whitney U test revealed extensive differences in microbial abundance between the Normal and Disease groups across all taxonomic levels, as shown in Figure 4.8. At $p < 0.05$, significant differences were identified in 19 of 26 tested phyla, 86 of 130 families, 158 of 240 genera, and 210 of 306 species, with the greatest number occurring at the genus and species levels. At the phylum level, *Verrucomicrobiota*, *Bacillota*, and *Actinomycesetota* differed significantly, with *Bacillota* and *Actinomycesetota* showing greater abundance in Normal samples. Several families, including *Akkermansiaceae*, *Butyricicoccaceae*, *Lachnospiraceae*, *Bifidobacteriaceae*, and *Clostridiaceae*, were also more abundant in the Normal group, whereas *Morganellaceae* and *Odoribacteraceae* were comparatively enriched in Disease samples. At the genus level, beneficial or commensal taxa such as *Roseburia*, *Anaerobutyricum*, *Dorea*, *Coproccoccus*, *Subdoligranulum*, and *Anaerostipes* were significantly reduced in Disease samples. Similar differences were observed at the species level, indicating substantial microbial dysbiosis and a general reduction in several potentially beneficial gut microorganisms in the Disease group. The taxa showing statistically significant differences in relative abundance between the Normal and Disease groups across phylum, family, genus, and species lev-

els. The phylum-level plots demonstrate broad shifts in the microbial community, while the family- and genus-level panels provide greater taxonomic resolution and reveal that several commensal microbial groups were more abundant in Normal samples. At the species level, the larger number of significant taxa indicates that disease-associated alterations were most clearly detected at finer taxonomic ranks. Differences in medians, interquartile ranges, and sample-wise distributions further show that some taxa were consistently altered across the groups, whereas others displayed substantial inter-individual variability.

Overall, the figure demonstrates widespread restructuring of the gut microbiome in the Disease group, characterized by the reduction of several normally abundant taxa and the enrichment of selected disease-associated microorganisms. These consistent taxonomic shifts suggest that the identified microbial groups may contribute to the distinction between healthy and disease-associated gut profiles and act as non-invasive biomarkers.



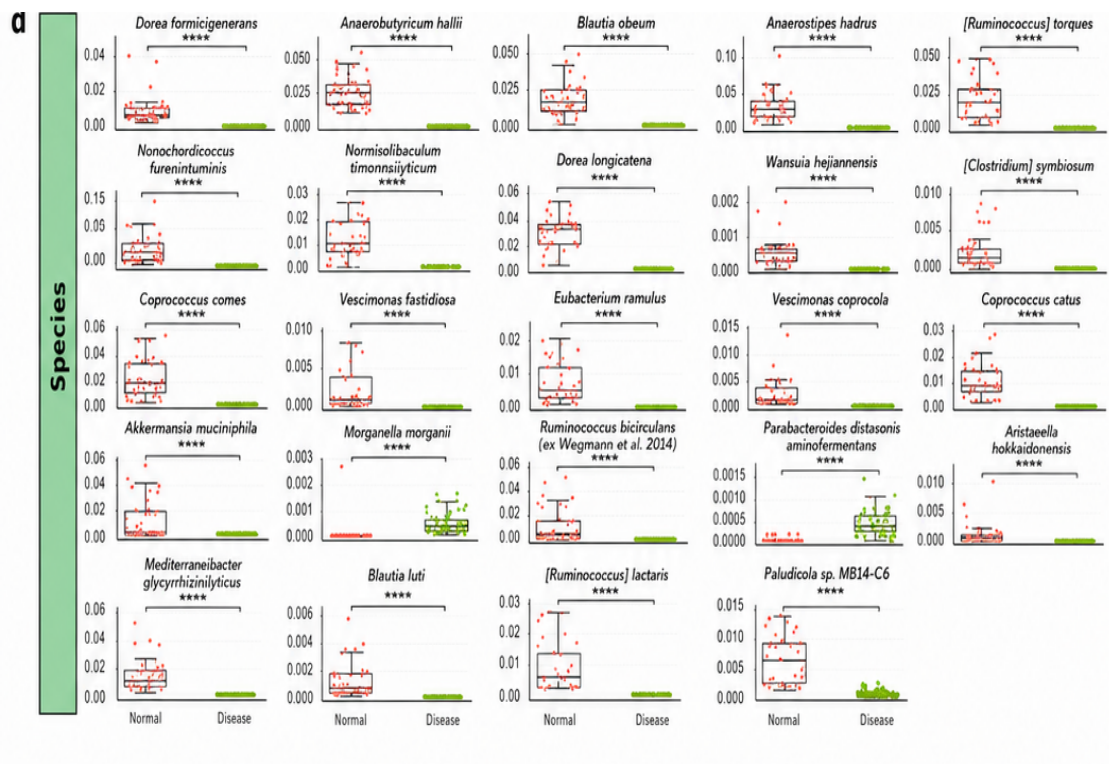


FIGURE 4.8: Mann–Whitney U test-based differential taxa analysis between Normal and Disease groups. Boxplots show significantly different taxa between Normal and Disease samples across the phylum, family, genus, and species levels. Each point represents an individual sample, boxes show group-wise abundance distributions, and asterisks indicate statistically significant differences at $p < 0.05$. The figures demonstrate the progression of disease-associated microbial differences from broad to specific taxonomic levels.

4.7.3 Obese vs Disease Comparison

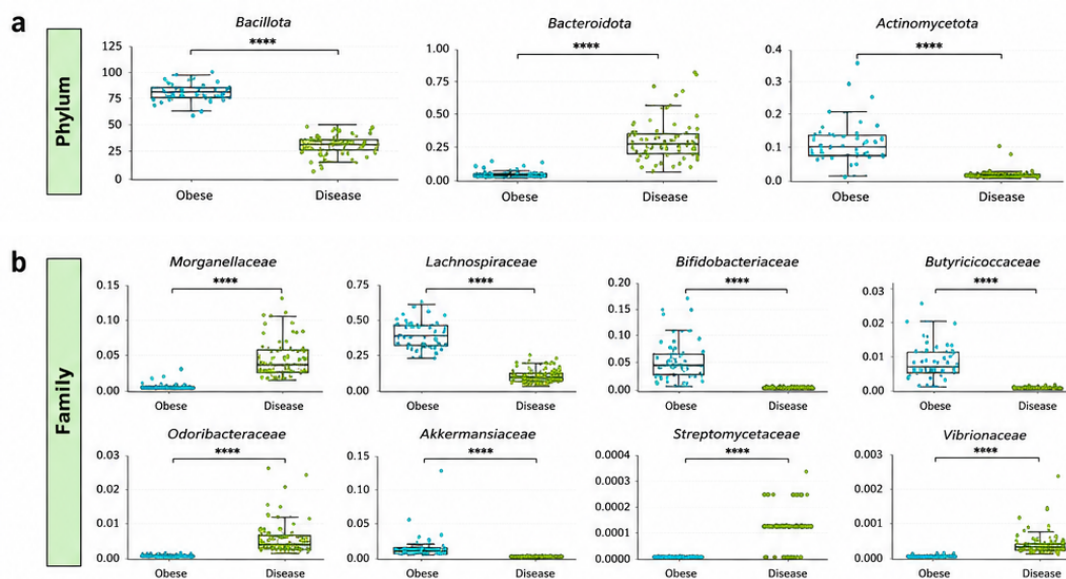
The Mann–Whitney U test-based differential abundance analysis showed extensive taxonomic differences between the Obese and Disease groups across all evaluated taxonomic ranks as shown in figure 4.9. The console summary indicated that, at $p < 0.05$, 18 of 26 phyla, 81 of 123 families, 154 of 234 genera, and 211 of 295 species were significantly different between the two groups, with the largest number of significant taxa detected at the species level. At the phylum level, *Bacillota* and *Actinomycetota* showed higher relative abundance in the Obese group, whereas *Bacteroidota* was higher in the Disease group.

At the family level, several families including *Lachnospiraceae*, *Bifidobacteriaceae*, *Butyrivibrionaceae*, and *Akkermansiaceae* were more abundant in Obese samples, while *Morganellaceae*, *Odoribacteraceae*, *Streptomyetaceae*, and *Vib-*

rimonaceae showed higher abundance in Disease samples. At the genus level, Obese samples showed higher abundance of multiple gut-associated genera, including *Anaerobutyricum*, *Enterocloster*, *Dorea*, *Roseburia*, *Solibaculum*, *Coprococcus*, *Anaerostipes*, *Vescimonas*, and *Subdoligranulum*, whereas *Morganella* and *Wan-*
suia were increased in Disease samples.

At the species level, several species such as *[Ruminococcus] torques*, *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *Blautia obeum*, *Dorea formicigenerans*, *Solibaculum mannosilyticum*, *Ethordicoccus fuertiuminis*, *Dorea longicatena*, *Bifidobacterium longum*, *Coprococcus comes*, *Eubacterium ramulus*, *Vescimonas coprophila*, *Vescimonas fastidiosa*, *Coprococcus catus*, *Bifidobacterium animalis*, *Enorma phocaeensis*, and *[Ruminococcus] lactaris* were higher in the Obese group, while *Morganella morganii*, *Salmonella enterica*, *Paenoreptidonas aminofrementans*, and *Odoribacter splanchnicus* showed higher abundance in Disease samples. Overall, the figure demonstrates strong group-wise separation between Obese and Disease samples, with Disease-associated samples showing enrichment of selected *Bacteroidota*- and *Proteobacteria*- related taxa, while Obese samples retained higher abundance of several *Bacillota*-associated gut taxa.

These findings suggest that disease-related dysbiosis involves coordinated changes across multiple taxonomic levels rather than variation in a single microbial group. The significantly altered taxa represent potential candidates for biomarker prioritization.



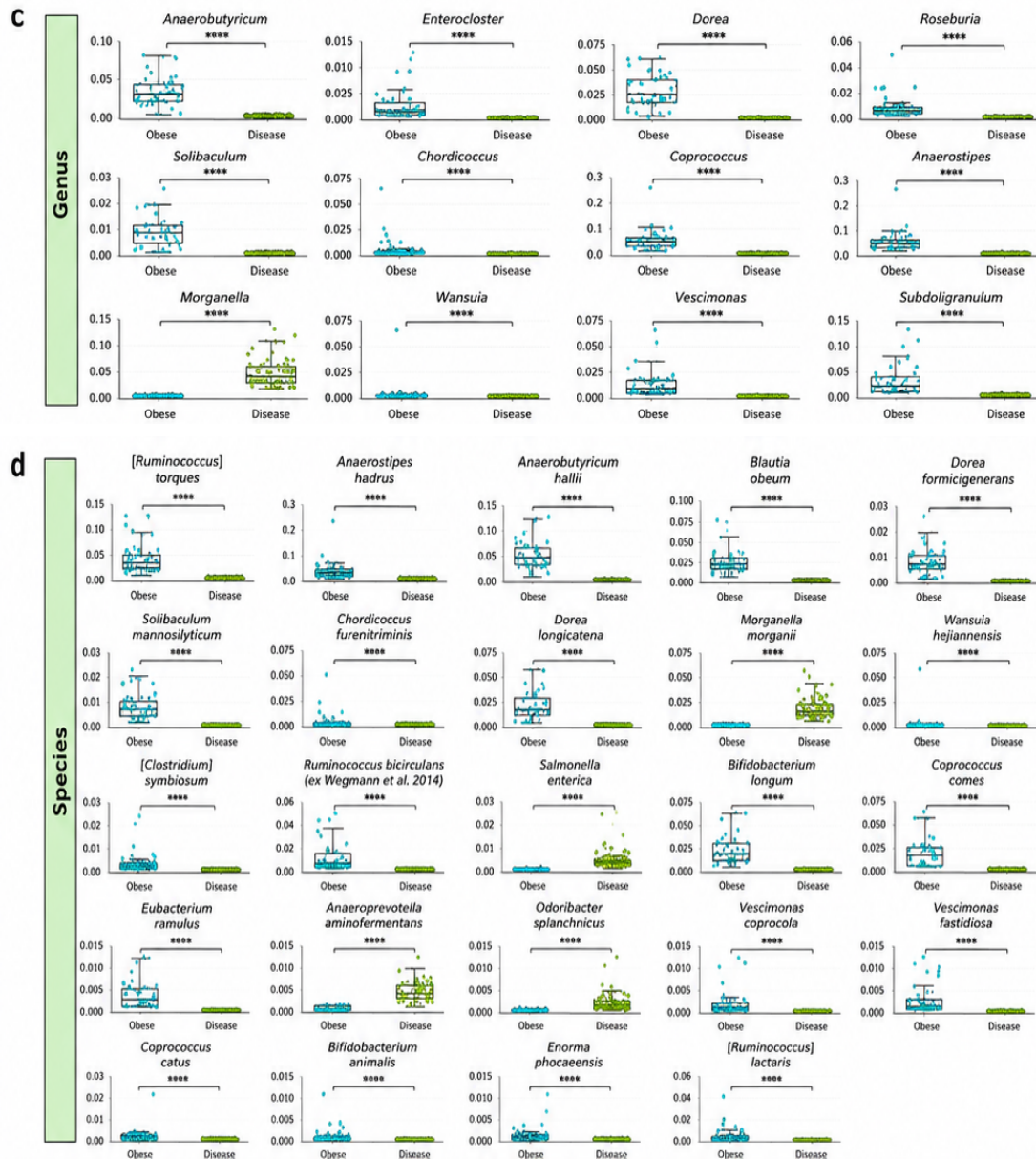


FIGURE 4.9: Mann–Whitney U test-based differential taxa analysis between Obese and Disease groups. Boxplots show significantly different taxa between Obese and Disease samples across the phylum, family, genus, and species levels. Each point represents an individual sample, boxes show group-wise abundance distributions, and asterisks indicate statistically significant differences at $p < 0.05$. The figures demonstrate the progression of disease-associated microbial differences from broad to specific taxonomic levels.

4.8 Alpha Diversity Analysis

The species-level alpha diversity comparison showed differences in microbial richness and diversity among the Normal, Obese, and Disease groups. In the alpha diversity plots, the Normal group generally showed higher observed taxa, Shannon

diversity, Simpson diversity, and Chao1 richness, while the Disease group showed comparatively lower diversity values across most indices.

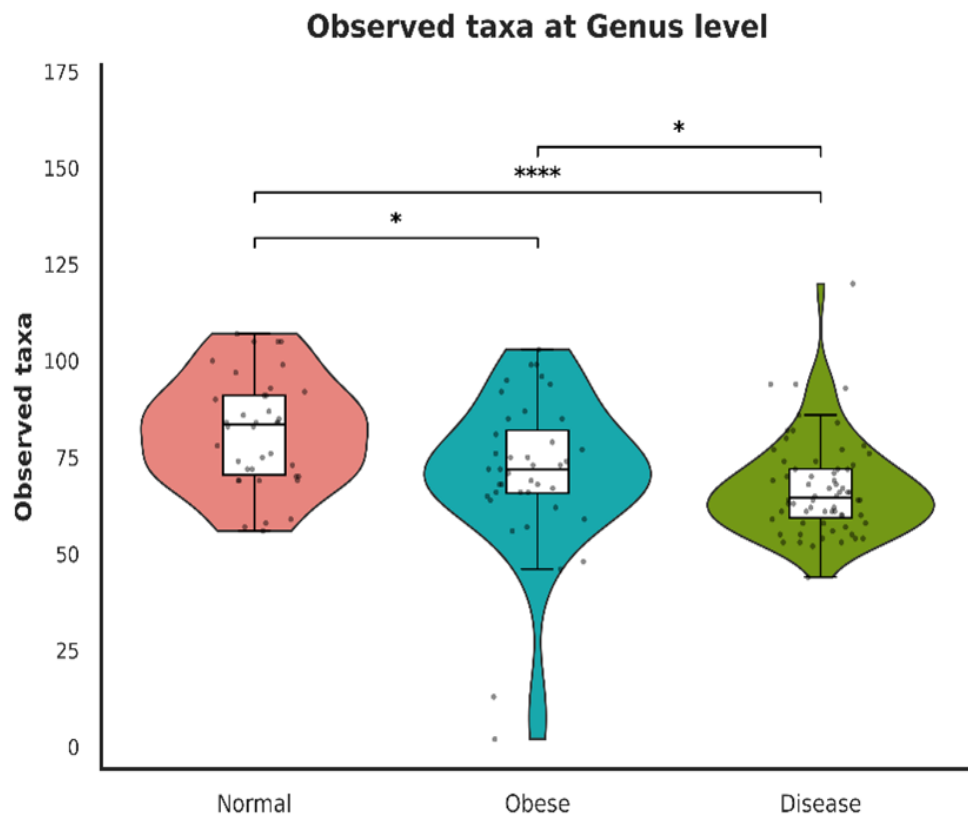
For observed taxa, Normal samples displayed a higher median richness than Disease samples, while the Obese group showed an intermediate distribution.

The Shannon diversity index showed a similar pattern, with lower species diversity in the Disease group compared with Normal and Obese groups.

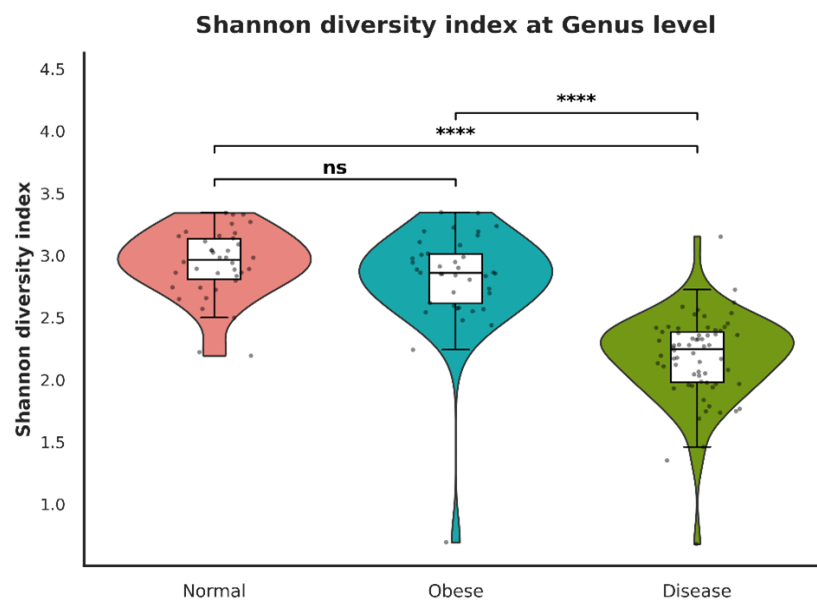
The Simpson diversity index also decreased in the Disease group, indicating reduced evenness and dominance structure at the species level.

For Chao1 richness, Normal samples showed higher estimated richness than Disease samples, whereas the Normal–Obese comparison was not significant.

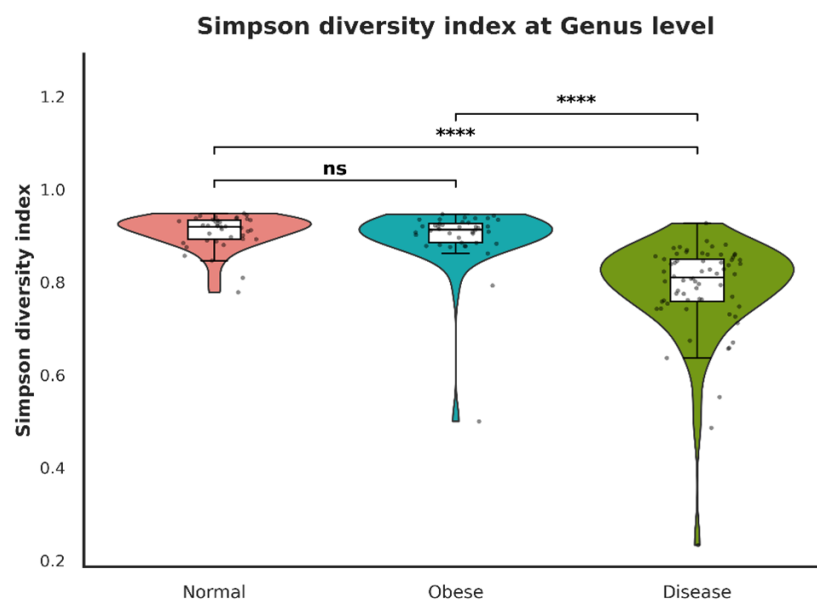
Overall, the alpha diversity results indicate that species-level microbial richness and diversity were reduced in Disease samples compared with the Normal and Obese groups as shown in figure 4.8.



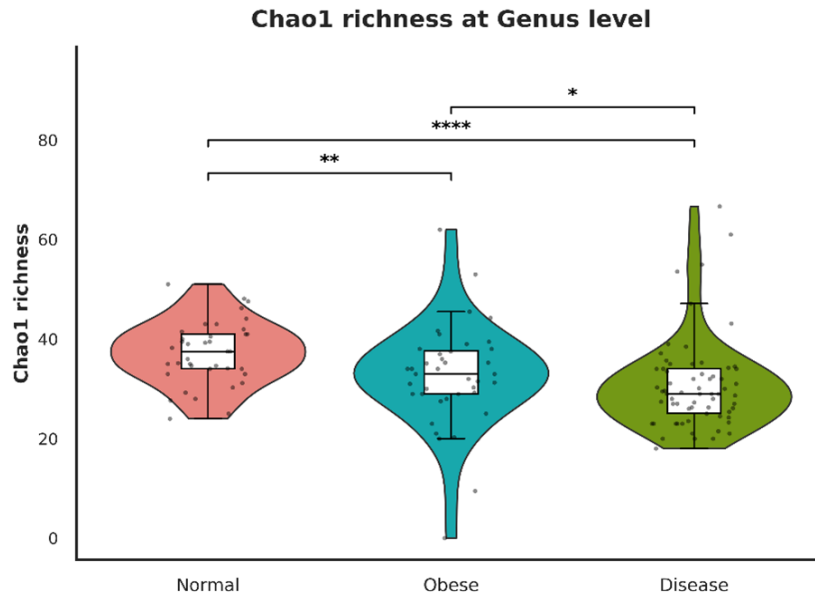
(A) **Observed taxa:** Species richness based on the number of observed taxa, showing lower richness in the Disease group compared with the Normal and Obese groups.



(B) Shannon diversity index: Species-level diversity incorporating richness and evenness, showing reduced diversity in the Disease group.



(C) Simpson diversity index: Diversity index reflecting species dominance and evenness, with lower values observed in Disease samples.



(D) **Chao1 richness:** Estimated species richness across the groups, showing a significant reduction in Disease samples compared with Normal samples, while the Normal–Obese comparison was non-significant.

4.9 Beta Diversity Analysis

The species-level Bray–Curtis beta diversity analysis showed clear differences in community composition among the metadata groups as shown in figure 4.10. The Bray–Curtis distance heatmap displayed pairwise dissimilarity among samples, with values ranging from 0 to 1, where darker colors indicate lower dissimilarity and yellow colors indicate higher dissimilarity.

The diagonal line represents zero distance between identical samples, while the block-like color patterns indicate that some sample groups shared more similar species-level profiles internally than with other groups. The contrast suggests structured microbial variation among the clinical categories.

The PCoA plot further showed separation of the groups based on Bray–Curtis distances. Disease samples formed a distinct cluster on the negative side of PCoA1, clearly separated from the Normal and Obese samples, while Normal and Obese samples clustered closer together on the positive side of PCoA1 with partial overlap. PCoA1 explained 32.56% of the total variation, and PCoA2 explained 9.18%, together capturing 41.74% of the species-level community variation. The shape

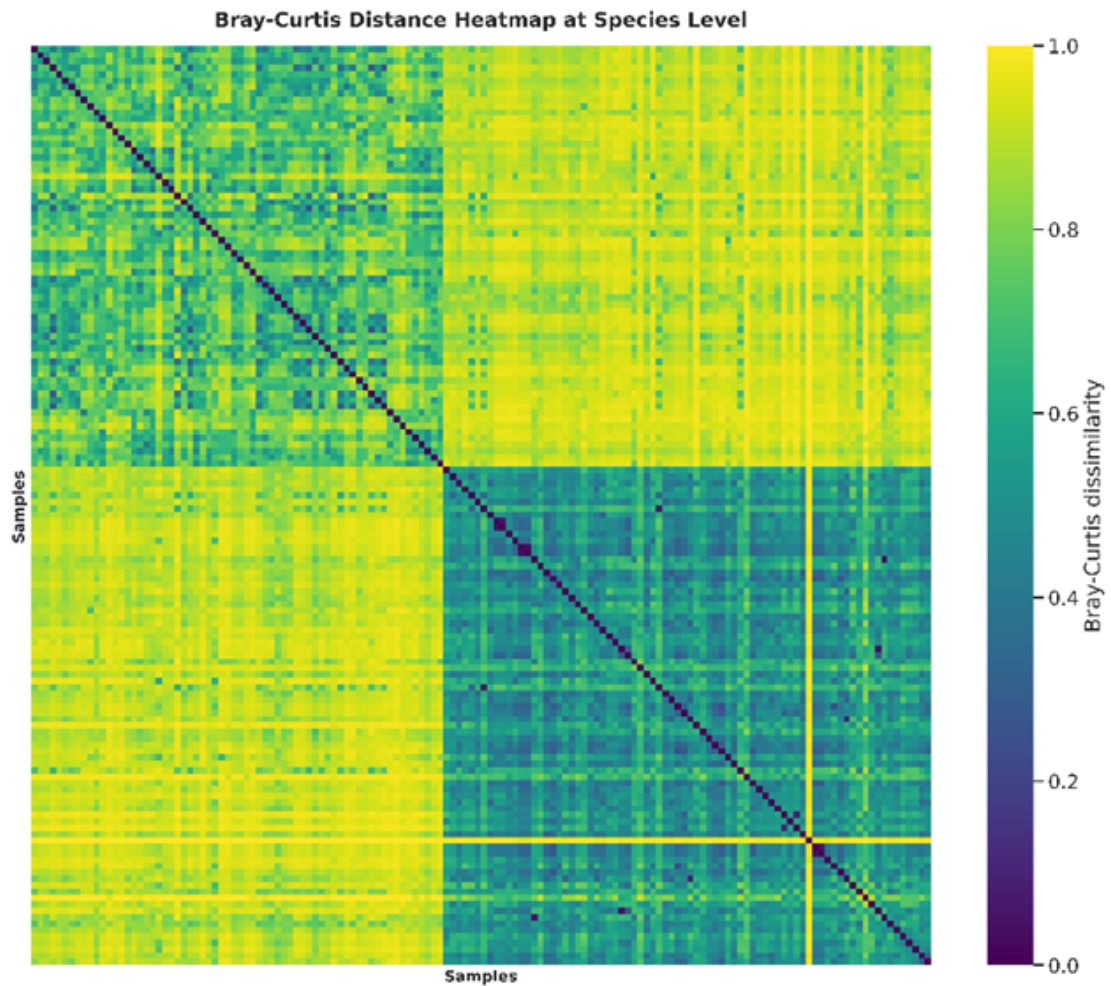


FIGURE 4.10: Bray–Curtis distance heatmap: Pairwise species-level dissimilarity matrix across samples, where darker colors represent lower dissimilarity and yellow colors represent higher dissimilarity.

coding for sex showed both male and female samples within the groups, but the dominant separation pattern was mainly associated with disease status rather than sex. Overall, the Bray–Curtis analysis indicates strong species-level compositional differences between Disease samples and the other two groups, while Normal and Obese communities showed comparatively closer clustering as shown in figure 4.11. This distinct clustering indicates that disease status was the major factor driving variation in species-level gut microbial composition.

4.10 PERMANOVA Analysis

PERMANOVA analysis based on Bray–Curtis distances showed significant beta diversity differences among the Disease, Obese, and Normal groups across all tax-

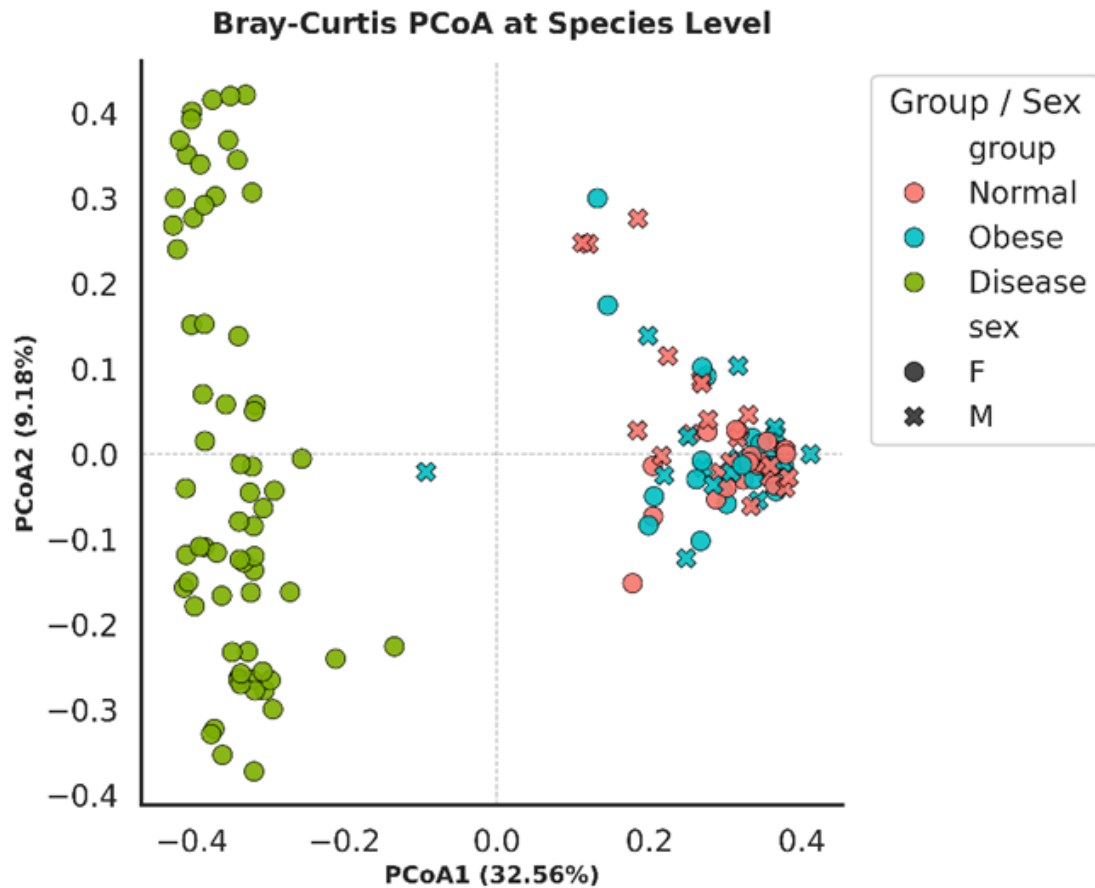


FIGURE 4.11: Bray–Curtis PCoA plot: Principal coordinate analysis based on species-level Bray–Curtis distances, showing clear separation of Disease samples from Normal and Obese samples along PCoA1.

onomic ranks as shown in figure 4.12. A total of 144 metadata-matched samples were included, comprising 66 Disease, 40 Obese, and 38 Normal samples. Significant group-wise differences were observed at the phylum, class, order, family, genus, and species levels, with all ranks showing $p = 0.001$ based on 999 permutations. The highest PERMANOVA test statistic was observed at the phylum level (196.42), followed by class (143.27), order (81.26), family (48.98), genus (45.78), and species (35.89). The decreasing test statistic from higher to lower ranks indicates that group-level compositional separation was strongest at broader taxonomic levels, while significant differences were still retained at genus and species levels. These results confirm that disease status was associated with significant shifts in overall gut microbial community structure across all taxonomic ranks. The decreasing PERMANOVA test statistic from phylum to species suggests stronger group-level separation at broader taxonomic levels.

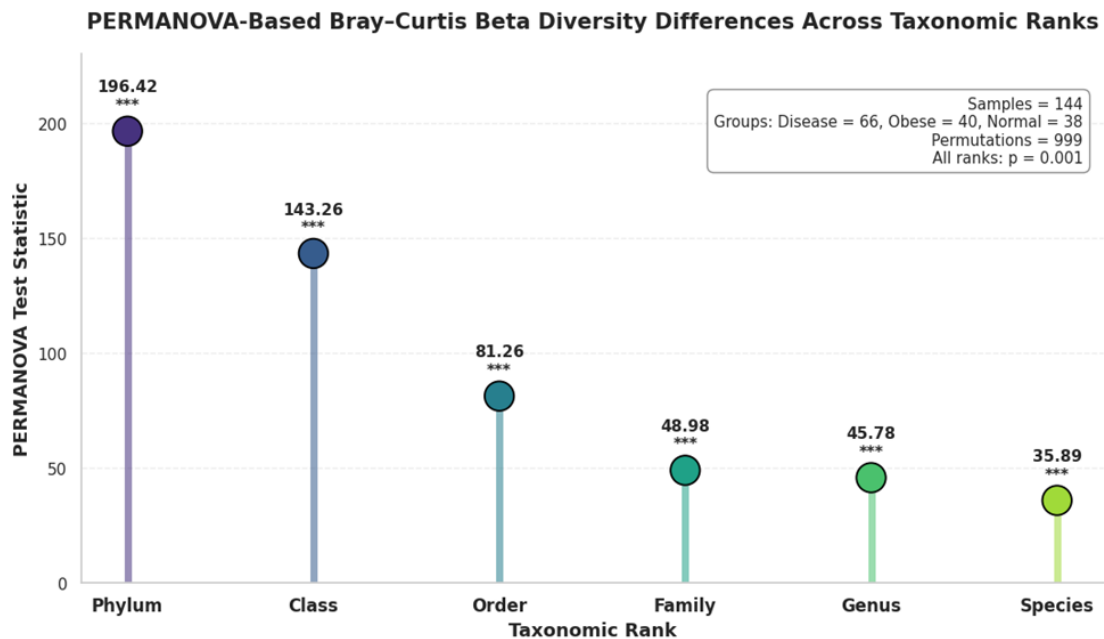


FIGURE 4.12: PERMANOVA-based comparison of Bray–Curtis beta diversity across taxonomic ranks. The plot shows PERMANOVA test statistics for phylum, class, order, family, genus, and species levels across Disease, Obese, and Normal groups. All taxonomic ranks showed significant group-wise compositional differences ($p = 0.001$) based on 999 permutations, with the strongest separation observed at the phylum level followed by class, order, family, genus, and species levels.

4.11 Machine Learning Based Classification using XGBoost and Explainable AI for Candidate Biomarker Prioritization

4.11.1 Random Forest Species-level Biomarker Classification Results

Species-level Random Forest biomarker classification showed strong discriminatory performance for disease-associated comparisons but weak separation between Normal and Obese groups. For Normal vs Disease, the model achieved an AUC and cross-validation AUC of 1.000, with the confusion matrix showing correct classification of all Disease and Normal test samples. Similarly, the Obese vs Disease model achieved an AUC and cross-validation AUC of 1.000, with all Disease and Obese samples correctly classified. In contrast, the Normal vs Obese model showed poor

classification performance, with a test AUC of 0.340 and cross-validation AUC of 0.554, indicating substantial overlap in species-level microbial profiles between these groups.

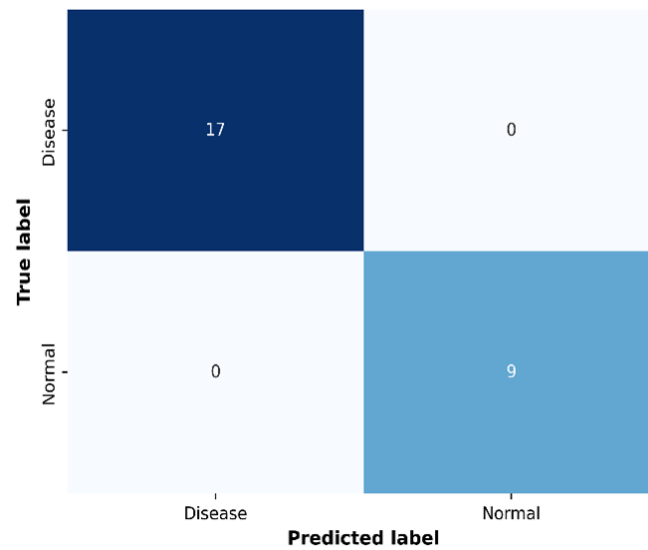
4.11.1.1 Normal vs Disease Comparison

For the Normal vs Disease comparison, the species-level Random Forest biomarker model showed complete classification performance, with both the test AUC and cross-validation AUC reaching 1.000. The confusion matrix confirmed perfect separation between the two groups, correctly classifying all 17 Disease samples and all 9 Normal samples, with no false-positive or false-negative predictions. The ROC curve also showed an ideal profile, with the curve reaching the upper-left corner and an AUC of 1.000, indicating complete discrimination between Normal and Disease samples as shown in figure 4.13.

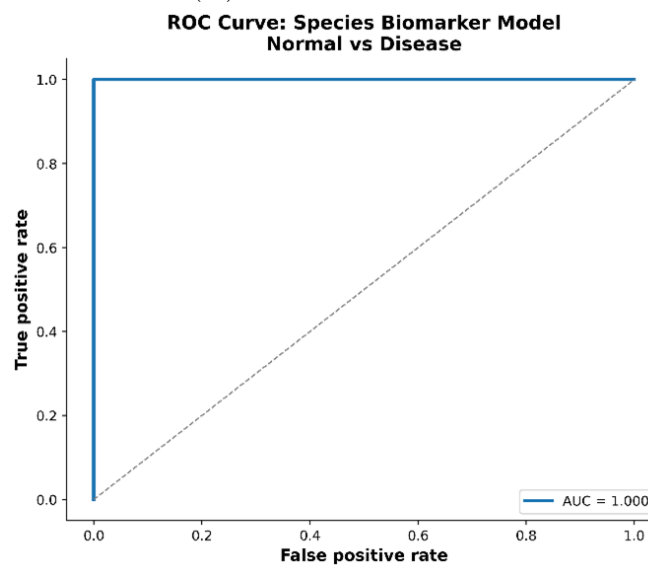
The corresponding biomarker abundance heatmap showed a clear group-wise separation pattern, where several top-ranked species displayed higher relative abundance in one group and reduced abundance in the other. The Random Forest importance plot identified the top species-level biomarkers contributing to this classification, with *Dorea formicigenerans*, *Anaerobutyricum hallii*, *Blautia obeum*, *Chordicoccus furentiruminis*, *Anaerostipes hadrus*, *Dorea longicatena*, *[Ruminococcus] torques*, *Morganella morganii*, *Solibaculum mannosilyticum*, and *Wansuia hejianensis* among the highest-ranking taxa. Overall, the model results show that species-level microbial features strongly differentiated Normal and Disease groups as shown in figure 4.14.

4.11.1.2 Normal vs Obese Comparison

For the Normal vs Obese comparison, the species-level Random Forest model showed weak classification performance, with a test AUC of 0.340 and a cross-validation AUC of 0.554. The confusion matrix showed substantial misclassification between the two groups, with only 2 Normal samples correctly classified and 8 Normal samples incorrectly predicted as Obese, while 7 Obese samples were correctly classified and 3 Obese samples were incorrectly predicted as Normal.



(A) Confusion matrix

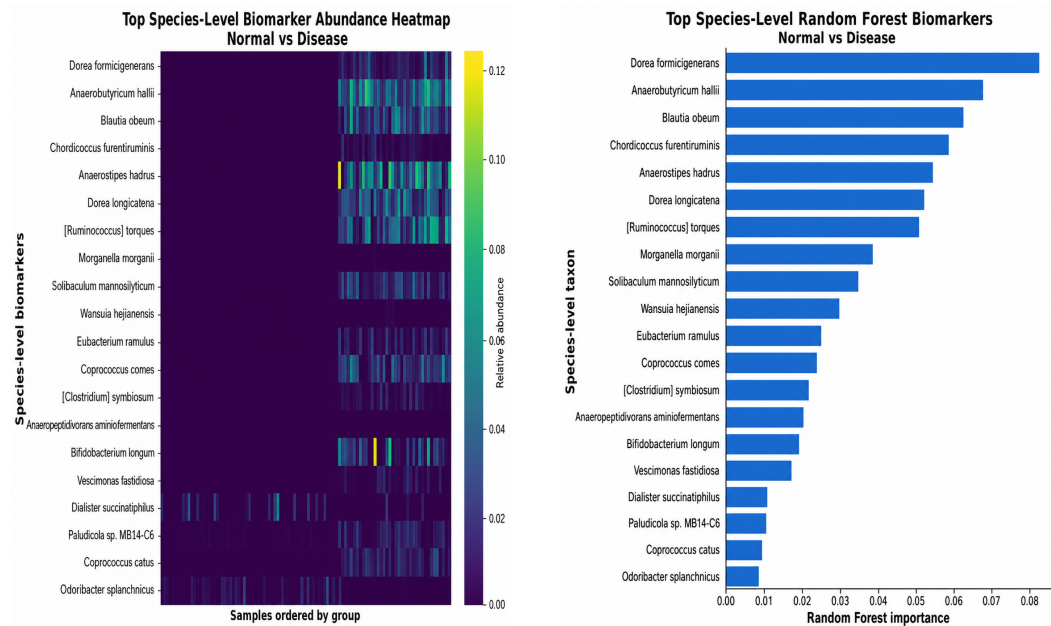


(B) ROC curve

FIGURE 4.13: Normal vs Disease comparison using species-level Random Forest biomarker classification. The confusion matrix shows classification performance, while the ROC curve represents model discrimination ability.

The ROC curve remained close to the diagonal reference line and produced an AUC below 0.5, indicating poor discriminatory ability between Normal and Obese samples at the species level as shown in figure 4.15.

The biomarker heatmap showed some variation in abundance patterns across selected species, but the separation between groups was not as clear as observed in the disease comparisons. The Random Forest importance plot ranked *Akkermansia muciniphila*, *Bifidobacterium longum*, *[Clostridium] symbiosum*, *Anaerostipes*



(A) Biomarker abundance heatmap (B) Random Forest feature importance

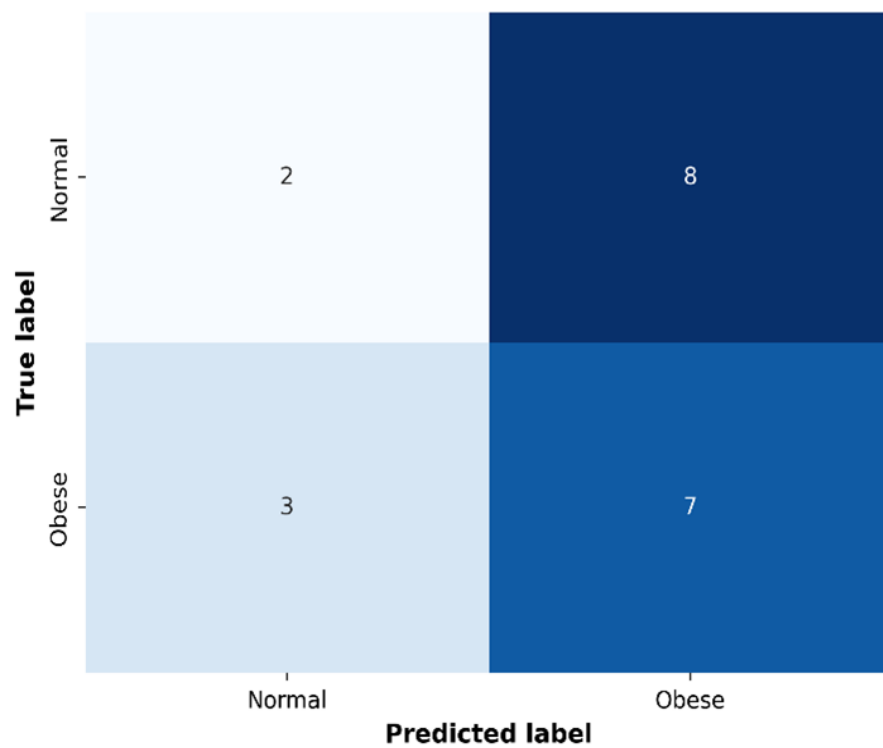
FIGURE 4.14: Normal vs Disease comparison using species-level Random Forest biomarker classification. The heatmap shows the abundance pattern of top species-level biomarkers across samples, while the feature importance plot highlights the most influential microbial taxa contributing to classification.

hadrus, *Blautia luti*, *Phocaeicola dorei*, *Chordicoccus furentiruminis*, *Dorea longicatena*, *Monoglobus pectinilyticus*, and *Anaerobutyricum hallii* among the top contributing species. Overall, the Normal and Obese groups showed overlapping species-level profiles, resulting in limited classification accuracy as shown in figure 4.16.

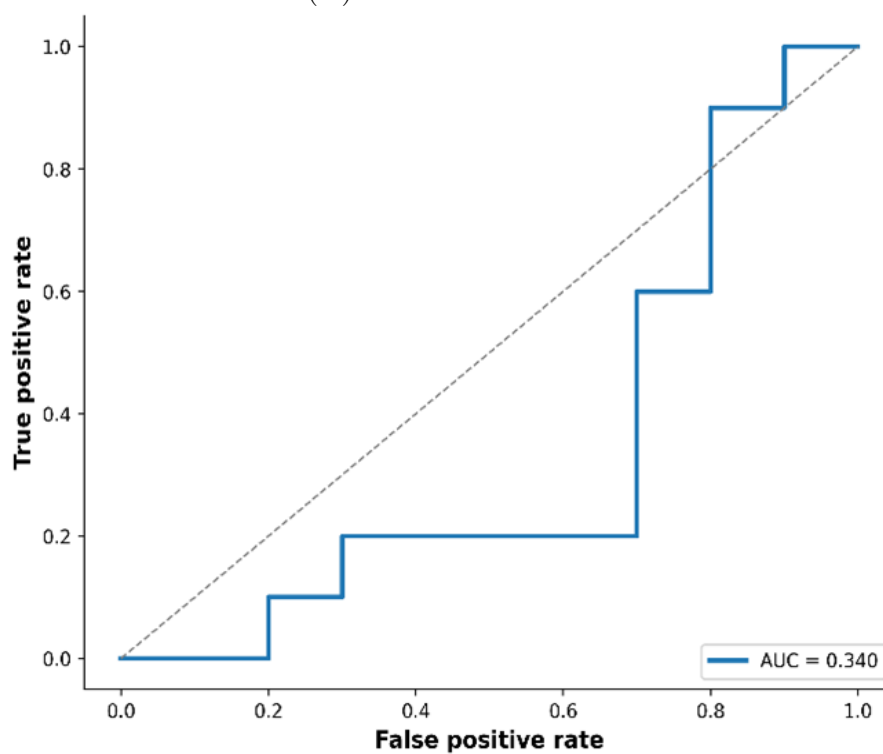
4.11.1.3 Obese vs Disease Comparison

For the Obese vs Disease comparison, the species-level Random Forest biomarker model again showed complete classification performance, with both the test AUC and cross-validation AUC reaching 1.000. The confusion matrix demonstrated perfect prediction of both classes, correctly identifying all 17 Disease samples and all 10 Obese samples, with no misclassified samples. The ROC curve confirmed this complete discriminatory performance, with an AUC of 1.000 as shown in figure 4.17.

The biomarker abundance heatmap showed strong group-wise separation between Obese and Disease samples, with several top-ranked species displaying distinct

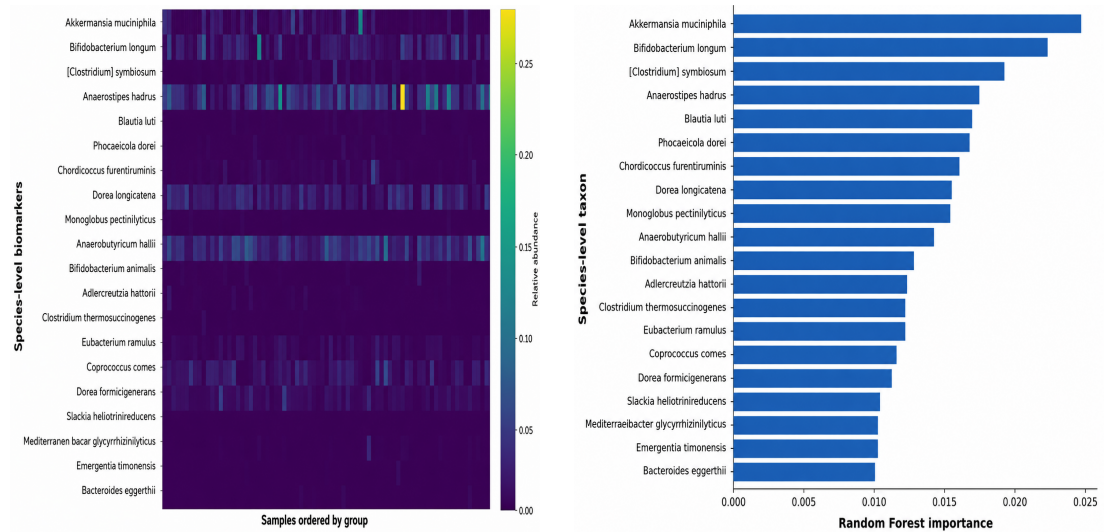


(A) Confusion matrix



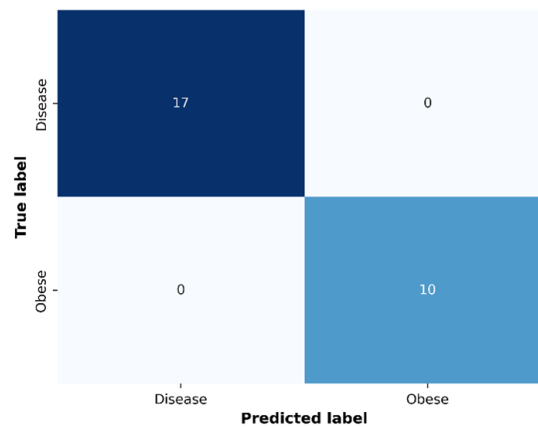
(B) ROC curve

FIGURE 4.15: Normal versus Obese comparison at the species level using Random Forest biomarker classification. The confusion matrix shows misclassification between the two groups, while the ROC curve indicates poor discriminatory performance with an AUC below 0.5.

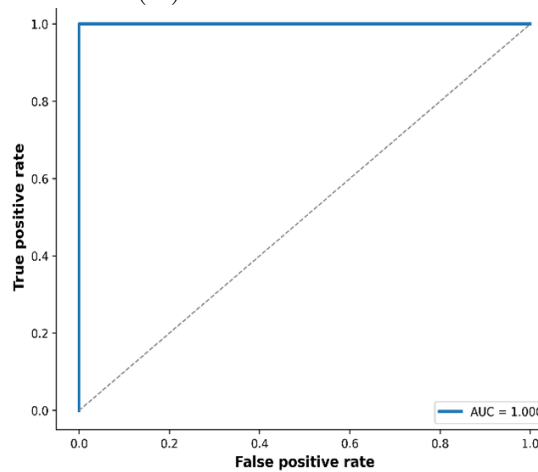


(A) Biomarker abundance heatmap (B) Random Forest feature importance

FIGURE 4.16: Normal vs Obese Comparison Specie Level Random Forest Biomarker Classification Heatmap and Feature Importance.



(A) Confusion matrix



(B) ROC curve

FIGURE 4.17: Obese vs Disease Comparison Specie Level Random Forest Biomarker Classification Confusion Matrix and ROC Curve.

abundance patterns across the two groups. The Random Forest importance plot identified *Morganella morganii* as the highest-ranking biomarker, followed by *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *[Ruminococcus] torques*, *Solibaculum mannosilyticum*, *Blautia obeum*, *Bifidobacterium longum*, *Salmonella enterica*, *Bacteroides eggerthii*, and *Dorea formicigenerans*. Overall, these results indicate that species-level microbial markers provided strong separation between Obese and Disease groups as shown in figure 4.18.

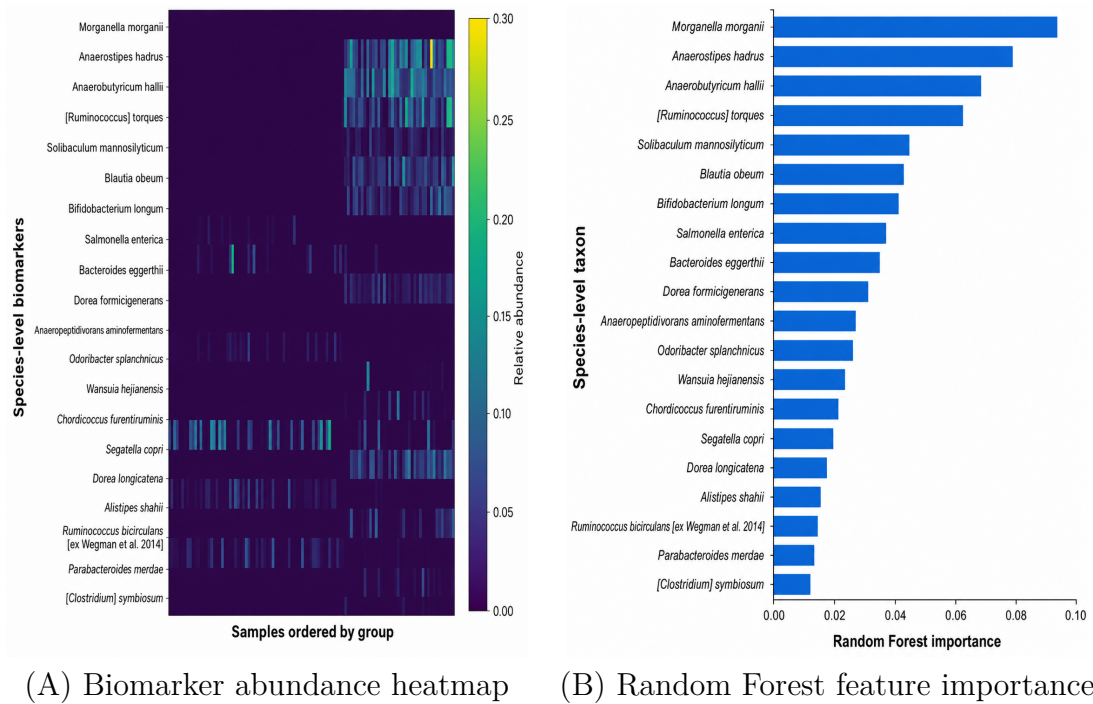


FIGURE 4.18: Obese vs Disease Comparison Specie Level Random Forest Biomarker Classification Heatmap and Feature Importance.

4.11.2 XAI-based Disease Biomarker Analysis

The final integrated biomarker output identified 825 total integrated biomarkers, of which 401 were disease-enriched biomarkers. Among these, 303 biomarkers met the high-confidence disease biomarker criterion of $p < 0.05$. Across both Normal vs Disease and Obese vs Disease comparisons, XAI-based classification highlighted a consistent set of disease-associated species, including *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *[Ruminococcus] torques*, *Dorea longicatena*, *Faecalibacterium prausnitzii*, and *Phocaeicola vulgatus*. Model performance was strongest for Normal vs Disease, with perfect classification and $AUC = 1.000$, while Obese vs

Disease also showed high discrimination with $AUC = 0.976$ and only one misclassified sample. Overall, the XAI-based biomarker analysis identified a robust set of species-level disease-associated taxa supported by both model performance and SHAP-based feature contribution.

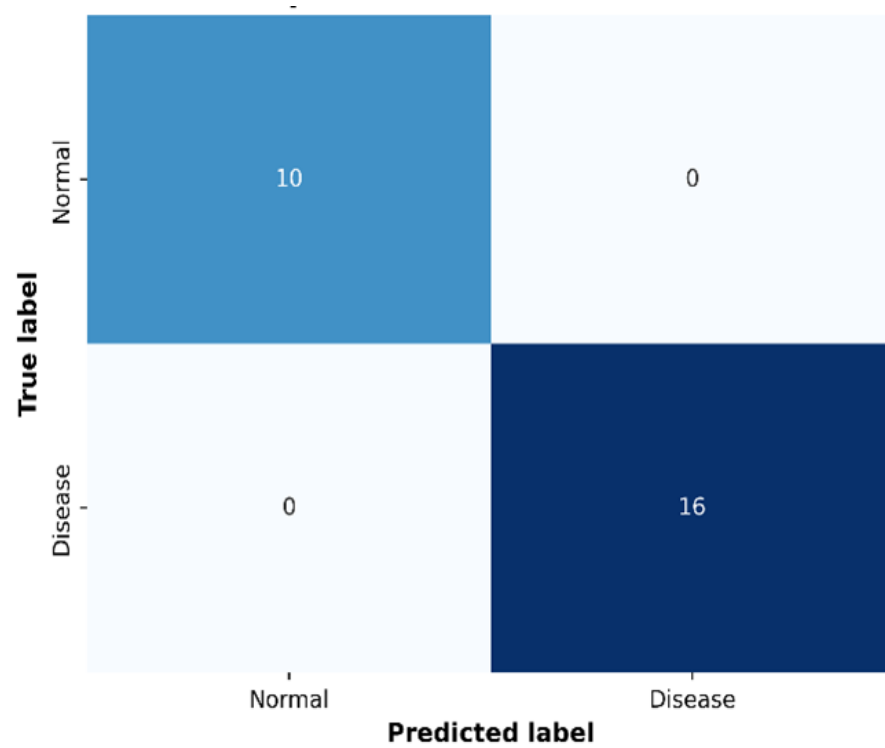
4.11.2.1 Normal vs Disease Comparison

For the Normal vs Disease comparison, the XAI-based species-level biomarker model showed strong disease-classification performance. The ROC curve produced an AUC of 1.000, indicating complete discrimination between Normal and Disease samples. This was supported by the confusion matrix, where all 10 Normal samples and all 16 Disease samples were correctly classified, with no false-positive or false-negative predictions as shown in figure 4.19.

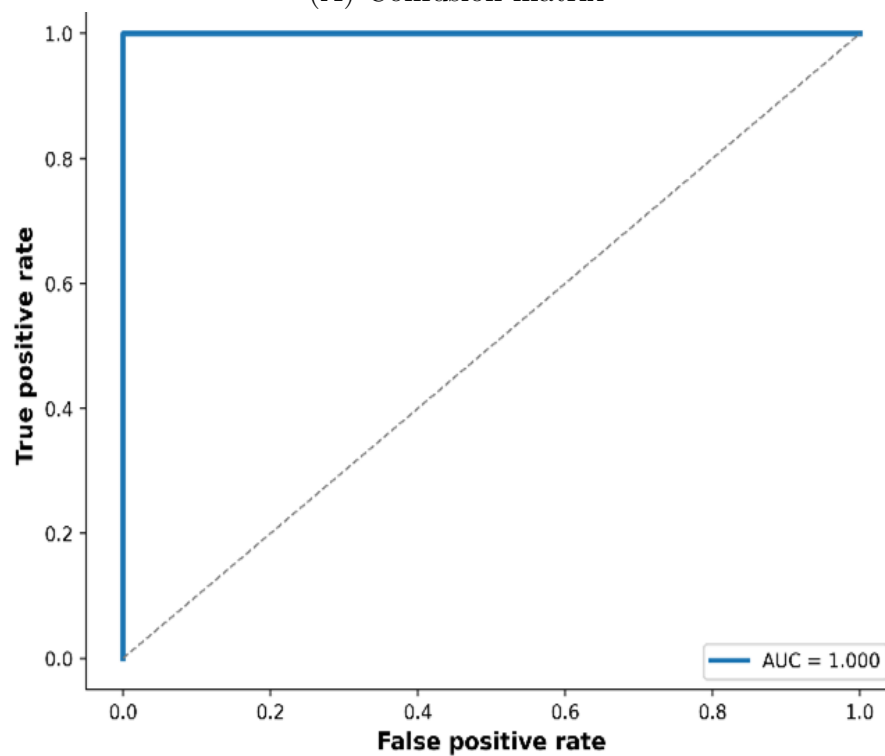
The table 4.7 presents the biomarkers ranked using integrated machine-learning and statistical criteria, including SHAP importance, p-value, FDR-adjusted significance, and log2 fold change.

Most biomarkers, including *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *Dorea longicatena*, and *Faecalibacterium prausnitzii*, were enriched in the normal group, whereas *Clostridium perfringens*, *Phocaeicola vulgatus*, and *Haemophilus parainfluenzae* showed higher abundance in the disease group, indicating potential microbial signatures associated with disease status.

The SHAP biomarker ranking identified *Anaerostipes hadrus* as the most influential disease-positive biomarker, followed by *Anaerobutyricum hallii*, *[Ruminococcus] torques*, *Dorea longicatena*, *Dorea formicigenerans*, *Pauldosa sp. MB14-C6*, *Faecalibacterium prausnitzii*, *Clostridium perfringens*, *Phocaeicola vulgatus*, and *Haemophilus parainfluenzae*. The disease-associated SHAP heatmap showed clear group-wise abundance differences for the selected biomarkers, with several taxa displaying stronger signal in the Disease group compared with Normal samples. The SHAP summary plot further showed that the top-ranked biomarkers contributed positively toward the Disease class prediction, confirming their importance in model-based disease discrimination as shown in figure 4.20.



(A) Confusion matrix



(B) ROC curve

FIGURE 4.19: XAI-based disease-associated species biomarker analysis for Normal vs Disease comparisons Confusion Matrix and ROC Curve.

TABLE 4.7: Top species-level biomarkers identified using SHAP importance, statistical significance, FDR correction, and $\log_2$ fold-change analysis between Normal and Disease groups.

Taxon	Higher in	SHAP	p-value	FDR	$\log_2$ FC
<i>Anaerostipes hadrus</i>	Normal	2.8362	1.05e-22	4.91e-21	-24.93
<i>Anaerobutyricum hallii</i>	Normal	0.3792	1.05e-22	4.91e-21	-24.81
<i>[Ruminococcus] torques</i>	Normal	0.1544	1.05e-22	4.91e-21	-24.51
<i>Dorea longicatena</i>	Normal	0.0344	6.24e-22	1.83e-20	-24.30
<i>Dorea formicigenerans</i>	Normal	0.0151	1.05e-22	4.91e-21	-23.09
<i>Paludicola</i> sp. MB14-C6	Normal	0.0108	7.32e-15	7.14e-14	-3.73
<i>Faecalibacterium prausnitzii</i>	Normal	0.0075	2.08e-06	6.15e-06	-1.26
<i>Clostridium perfringens</i>	Disease	0.0065	3.05e-06	8.82e-06	1.79
<i>Phocaeicola vulgatus</i>	Disease	0.0053	1.99e-06	5.97e-06	1.95
<i>Haemophilus parainfluenzae</i>	Disease	0.0046	7.14e-03	1.12e-02	0.99
<i>Christensenella</i> sp. MSJ-20	Normal	0.0025	1.28e-02	1.87e-02	-2.48
<i>Blautia obeum</i>	Normal	0.0000	1.05e-22	4.91e-21	-24.08
<i>Chordicoccus furentiruminis</i>	Normal	0.0000	6.19e-22	1.83e-20	-21.26
<i>Solibaculum mannosilyticum</i>	Normal	0.0000	6.23e-22	1.83e-20	-23.10
<i>Wansuia hejianensis</i>	Normal	0.0000	1.85e-20	4.25e-19	-18.43

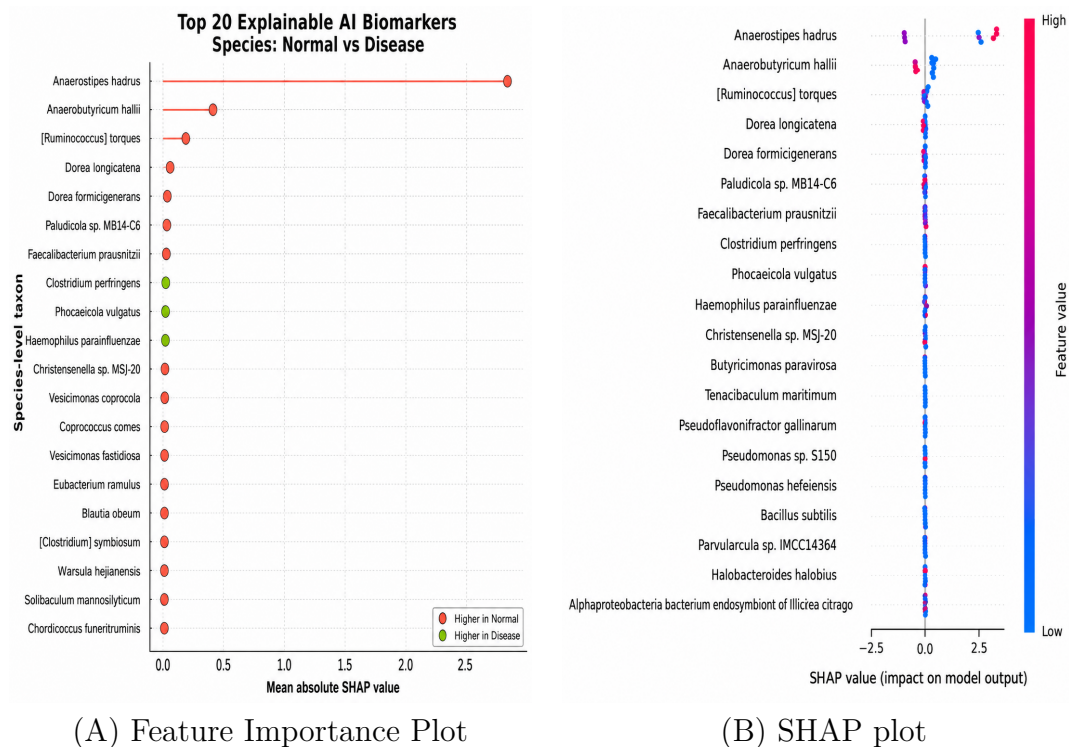


FIGURE 4.20: XAI-based disease-associated species biomarker analysis for Normal vs Disease comparisons Feature Importance and SHAP plot

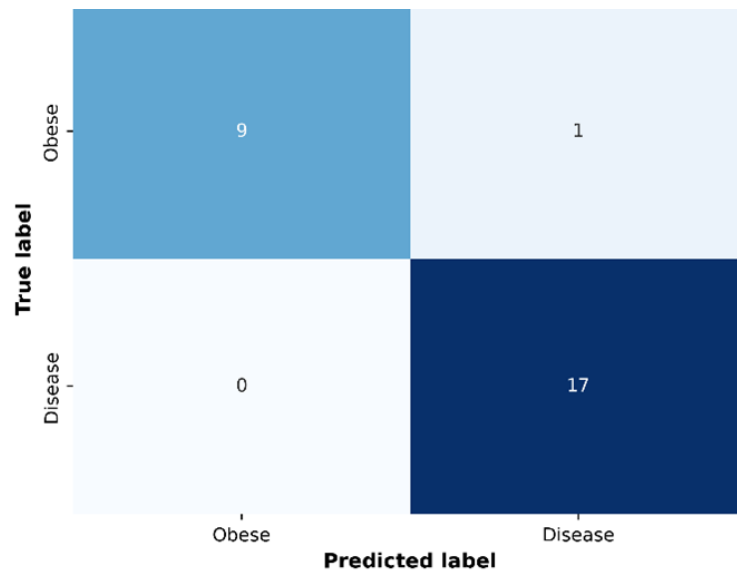
4.11.2.2 Obese vs Disease Comparison

For the Obese vs Disease comparison, the XAI-based species-level biomarker model also showed high predictive performance. The ROC curve achieved an AUC of 0.976, indicating strong separation between Obese and Disease samples. The confusion matrix showed that 17 Disease samples were correctly classified as Disease, while 9 Obese samples were correctly classified as Obese and 1 Obese sample was misclassified as Disease. No Disease sample was misclassified as Obese as shown in figure 4.21.

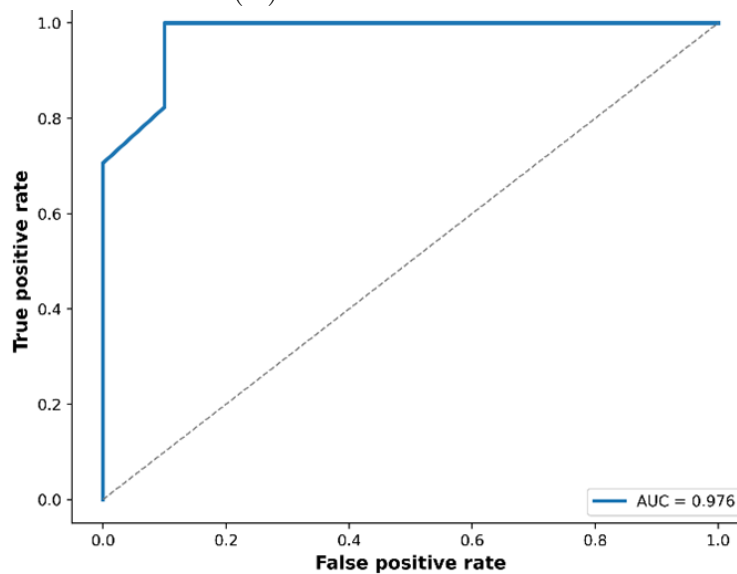
The table 4.8 shows the top species-level biomarkers identified using combined machine-learning and statistical evidence, including SHAP importance, p-value, FDR-adjusted significance, and log₂ fold change. Most discriminatory taxa, including *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *[Ruminococcus] torques*, *Dorea longicatena*, and *Paludicola sp. MB14-C6*, were enriched in the obese group, whereas *Parabacteroides merdae*, *Morganella morganii*, and *Salmonella enterica* were higher in the disease group. These results suggest a clear species-level microbial shift between obesity-associated and disease-associated gut microbiome profiles.

TABLE 4.8: Top species-level biomarkers identified using SHAP importance, statistical significance, FDR correction, and log₂ fold-change analysis between Obese and Disease groups.

Taxon	Higher in	SHAP	p-value	FDR	log ₂ FC
<i>Anaerostipes hadrus</i>	Obese	2.9961	3.06e-22	2.32e-20	-25.42
<i>Anaerobutyricum hallii</i>	Obese	0.3994	3.06e-22	2.32e-20	-25.10
<i>[Ruminococcus] torques</i>	Obese	0.0747	3.06e-22	2.32e-20	-24.78
<i>Dorea longicatena</i>	Obese	0.0139	2.32e-19	6.59e-18	-24.11
<i>Paludicola sp. MB14-C6</i>	Obese	0.0092	2.12e-11	1.26e-10	-3.89
<i>Parabacteroides merdae</i>	Disease	0.0089	1.29e-11	8.40e-11	3.60
<i>Blautia obeum</i>	Obese	0.0000	1.68e-21	9.55e-20	-24.08
<i>Dorea formicigenerans</i>	Obese	0.0000	8.95e-21	3.39e-19	-22.73
<i>Solibaculum mannosilyticum</i>	Obese	0.0000	8.95e-21	3.39e-19	-22.92
<i>Chordicoccus furentiruminis</i>	Obese	0.0000	4.56e-20	1.48e-18	-21.79
<i>Morganella morganii</i>	Disease	0.0000	7.36e-19	1.86e-17	18.47
<i>Wansuia hejianensis</i>	Obese	0.0000	1.02e-18	2.32e-17	-20.63
<i>[Clostridium] symbiosum</i>	Obese	0.0000	5.33e-18	1.10e-16	-21.34
<i>Ruminococcus bicirculans</i>	Obese	0.0000	2.45e-17	4.63e-16	-22.52
<i>Salmonella enterica</i>	Disease	0.0000	3.84e-17	6.70e-16	5.61



(A) Confusion Matrix



(B) ROC Curve

FIGURE 4.21: XAI-based disease-associated species biomarker analysis for Normal vs Disease comparisons Confusion Matrix and ROC Curve.

The top disease-positive SHAP biomarkers included *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *[Ruminococcus] torques*, *Dorea longicatena*, *Pauldosa sp. MB14-C6*, *Faecalibacterium prausnitzii*, *Phocaecicola vulgatus*, and additional taxa with lower SHAP contribution. The biomarker heatmap showed distinct abundance patterns across Obese and Disease samples, with several top-ranked taxa showing stronger signal in the Disease group. The SHAP summary plot confirmed that selected species contributed positively toward Disease prediction, supporting their role as disease-associated microbial markers in the Obese vs Disease comparison.

as shown in figure 4.22.

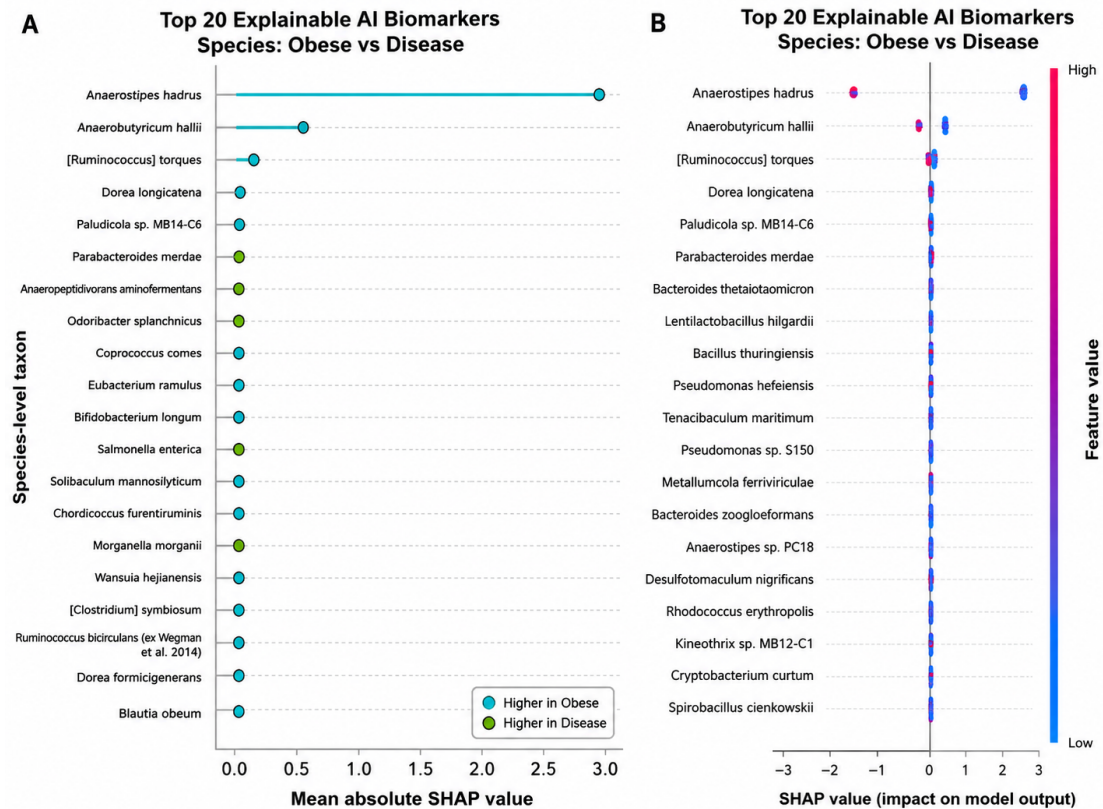
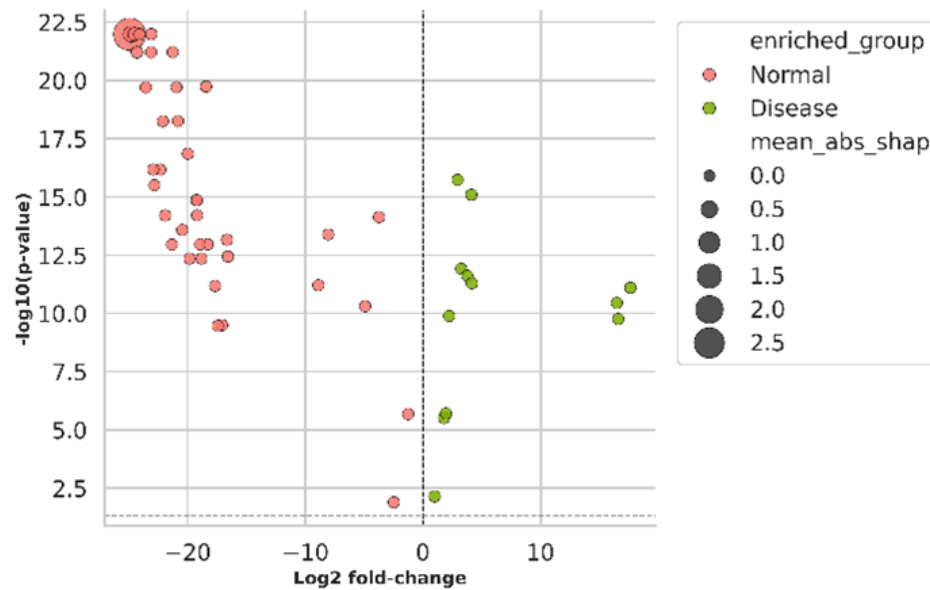


FIGURE 4.22: XAI-based disease-associated species biomarker analysis for Obese versus Disease comparison, including biomarker heatmap, feature importance, and SHAP summary plot.

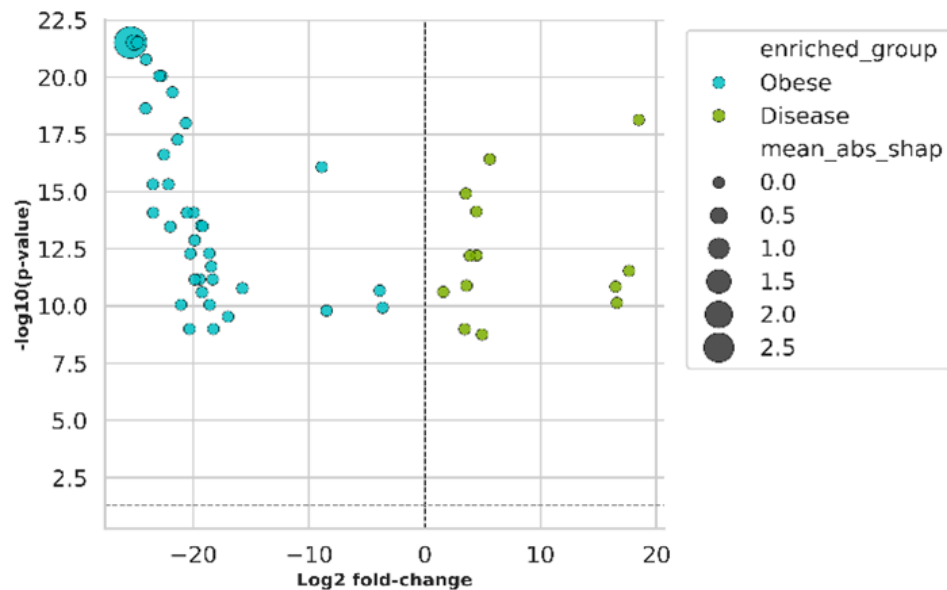
4.12 Differential Abundance Analysis

Integrated XAI and Mann–Whitney U test-based biomarker analysis identified species-level taxa that were both statistically significant and relevant to machine-learning classification. In the Normal vs Disease comparison, most biomarker species were enriched in the Normal group, as shown by the dense cluster of red bubbles on the negative log₂ fold-change side, with several taxa reaching high $-\log_{10}(\text{p-value})$ values. Disease-enriched taxa were fewer and appeared on the positive log₂ fold-change side. In the Obese vs Disease comparison, a similar pattern was observed, where many species were enriched in the Obese group and showed strong statistical significance, while a smaller number of Disease-enriched taxa appeared on the positive side. Across both comparisons, larger bubbles indi-

cated taxa with higher mean absolute SHAP values, representing stronger model contribution. These plots show that the integrated statistical and XAI approach successfully highlighted candidate species-level biomarkers that distinguish Disease samples from Normal and Obese groups as shown in figure 4.23.



A: Normal vs Disease comparison



B: Obese vs Disease comparison

FIGURE 4.23: Integrated XAI and Mann–Whitney U test-based species-level biomarker bubble plots. The plots show candidate biomarker species identified by combining statistical differential abundance testing with SHAP-based model importance. The x-axis represents $\log_2$ fold-change, the y-axis represents $-\log_{10}(p\text{-value})$, point color indicates the enriched group, and bubble size represents mean absolute SHAP value.

The combined interpretation of differential abundance and SHAP importance provides stronger evidence for biomarker prioritization than either approach alone. Taxa showing both statistically significant abundance differences and high mean absolute SHAP values are more likely to contribute consistently to disease classification.

4.13 Performance Matrices Evaluation

The AI model evaluation metrics showed strong classification performance for disease-associated comparisons, while Normal vs Obese comparisons showed weak discrimination. In the heatmap, Genus | Obese vs Disease and Species | Normal vs Disease achieved perfect values across all metrics, including accuracy, precision, recall/sensitivity, specificity, F1-score, test ROC-AUC, and cross-validation AUC. Genus | Normal vs Disease also showed very high performance, with accuracy 0.96, precision 0.94, recall 1.00, specificity 0.90, F1-score 0.97, test ROC-AUC 1.00, and CV-AUC 1.00. Similarly, Species | Obese vs Disease performed strongly, with accuracy 0.96, precision 0.94, recall 1.00, specificity 0.90, F1-score 0.97, test ROC-AUC 1.00, and CV-AUC 1.00 as shown in figure 4.24.

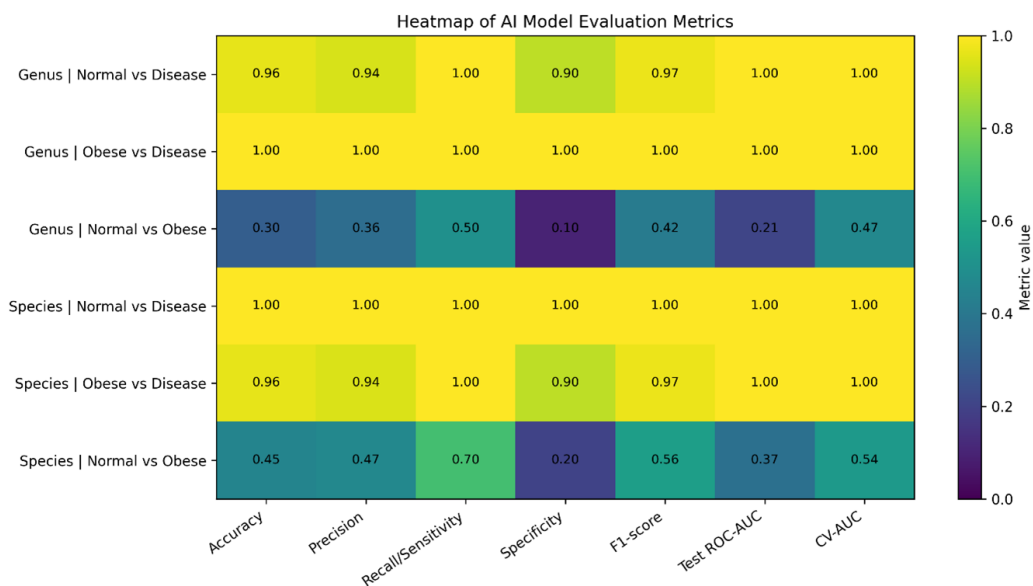


FIGURE 4.24: Heatmap of AI model evaluation metrics. Comparative performance summary across genus- and species-level classification models, showing high metric values for disease-related comparisons and lower values for Normal versus Obese models.

In contrast, the Normal vs Obese models showed poor performance at both genus and species levels. The genus-level Normal vs Obese model had low accuracy 0.30, precision 0.36, specificity 0.10, F1-score 0.42, test ROC-AUC 0.21, and CV-AUC 0.47, while the species-level Normal vs Obese model showed only modest improvement, with accuracy 0.45, precision 0.47, recall 0.70, specificity 0.20, F1-score 0.56, test ROC-AUC 0.37, and CV-AUC 0.54. The radar plots supported these findings, showing nearly complete metric coverage for disease-related comparisons and compressed, irregular profiles for Normal vs Obese models.

Overall, the AI model evaluation confirmed that microbial features strongly separated Disease samples from both Normal and Obese groups, whereas Normal and Obese samples showed overlapping microbial profiles and weaker classification performance.

The group-level presence heatmap of the top 100 species showed clear differences in microbial detection patterns across the normal, obese, and disease groups. Species were ranked according to their maximum presence across the three clinical groups, and the heatmap represents the percentage of samples in which each species was detected.

Overall, several species showed high presence in the normal and obese groups, whereas the disease group displayed a comparatively reduced detection pattern for many taxa. This reduction was visible as lighter regions across several species columns in the disease row, suggesting a loss or depletion of commonly detected gut microbial members during disease progression.

A number of species, including *Eubacterium ventriosum*, *Paludicola* sp. MB14-C6, *Phocaeicola vulgatus*, *Dorea formicigenerans*, *Ruminococcus torques*, *Anaerostipes hadrus*, and *Anaerobutyricum hallii*, showed strong presence across at least one clinical group. Some taxa were broadly detected in normal and obese samples but were reduced or absent in the disease group, indicating a possible decline in commensal microbial signatures.

In contrast, a subset of species showed relatively stronger presence in the disease group, suggesting their potential association with disease-related microbial imbalance. The obese group generally showed an intermediate pattern between normal

and disease, supporting the possibility of a gradual microbial shift along the normal $\rightarrow$ obese $\rightarrow$ disease axis. These findings indicate that species-level presence patterns can distinguish clinical groups and may help identify microbial biomarkers associated with metabolic syndrome-related dysbiosis as shown in figure 4.25.

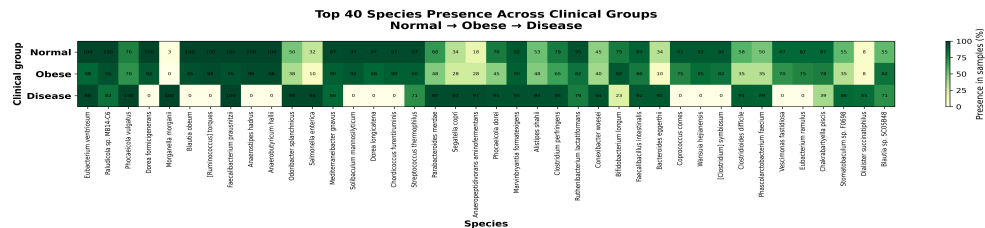


FIGURE 4.25: Presence–Absence heatmap of the top 40 species across normal, obese, and disease groups.

4.14 Cooccurrence Analysis

The genus-level microbial co-occurrence analysis was performed separately for the Normal, Obese, and Disease groups after matching taxonomic abundance data with metadata. The genus matrix contained 388 taxa across 146 samples, while the final group-wise analysis retained 38 Normal, 40 Obese, and 66 Disease samples. For each group, 40 taxa were retained before network filtering. The Normal group retained 16 edges, the Obese group retained 10 edges, and the Disease group retained only 1 edge, indicating a marked reduction in microbial co-occurrence complexity in the Disease group compared with the Normal and Obese groups. The network summary showed clear differences in genus-level microbial interaction structure across the three groups. The Normal network contained 16 nodes and 16 edges, including 11 positive and 5 negative associations, indicating a more connected and interaction-rich microbial community. The Obese network contained 7 nodes and 10 edges, all of which were positive associations, suggesting fewer taxa but a relatively dense positive co-occurrence structure. In contrast, the Disease network contained only 2 nodes and 1 positive edge, with no negative associations detected. Overall, the Disease group showed the lowest network complexity, while the Normal group showed the highest number of nodes and edges.

The Disease group showed the simplest microbial co-occurrence network, consisting of only 2 nodes and 1 positive edge. The detected association connected *Bacteroides* and *Phocaeicola*, indicating a single retained positive genus-level co-occurrence relationship within the Disease samples. No negative associations were observed in this network. The sparse structure indicates that only a limited number of strong genus-level associations passed the applied filtering threshold in the Disease group as shown in figure 4.26.

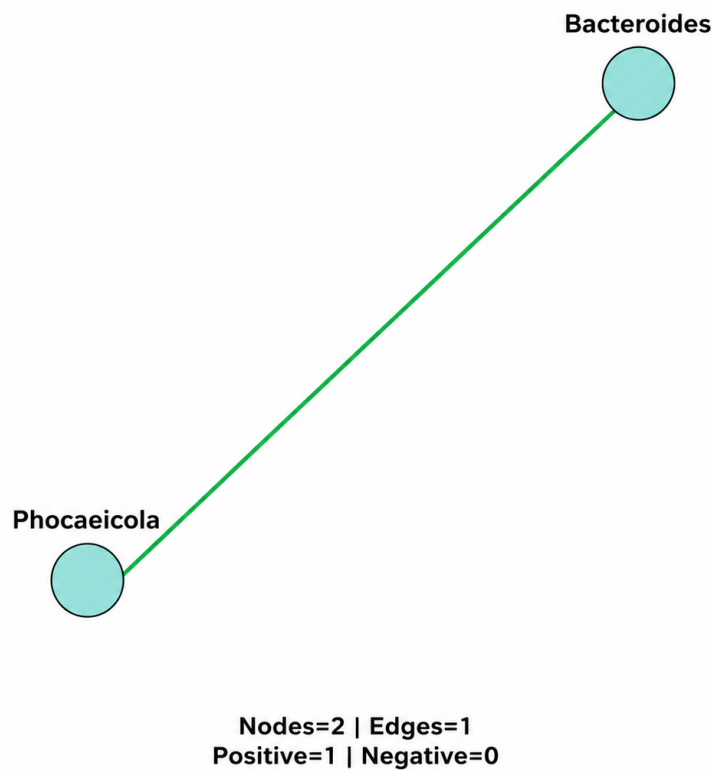


FIGURE 4.26: Disease-group microbial co-occurrence network. The network shows a single positive association between *Bacteroides* and *Phocaeicola*, with 2 nodes, 1 edge, 1 positive connection, and no negative edges.

The Normal group displayed the most complex microbial co-occurrence network among the three groups, containing 16 nodes and 16 edges. The network included both positive associations and negative associations, with 11 positive edges and 5 negative edges retained after filtering. Several small connected components were observed, indicating that different groups of genera formed separate co-occurrence modules rather than one fully connected network. This pattern reflects a more diverse and structured genus-level association profile in Normal samples compared

with the Disease group as shown in figure 4.27.

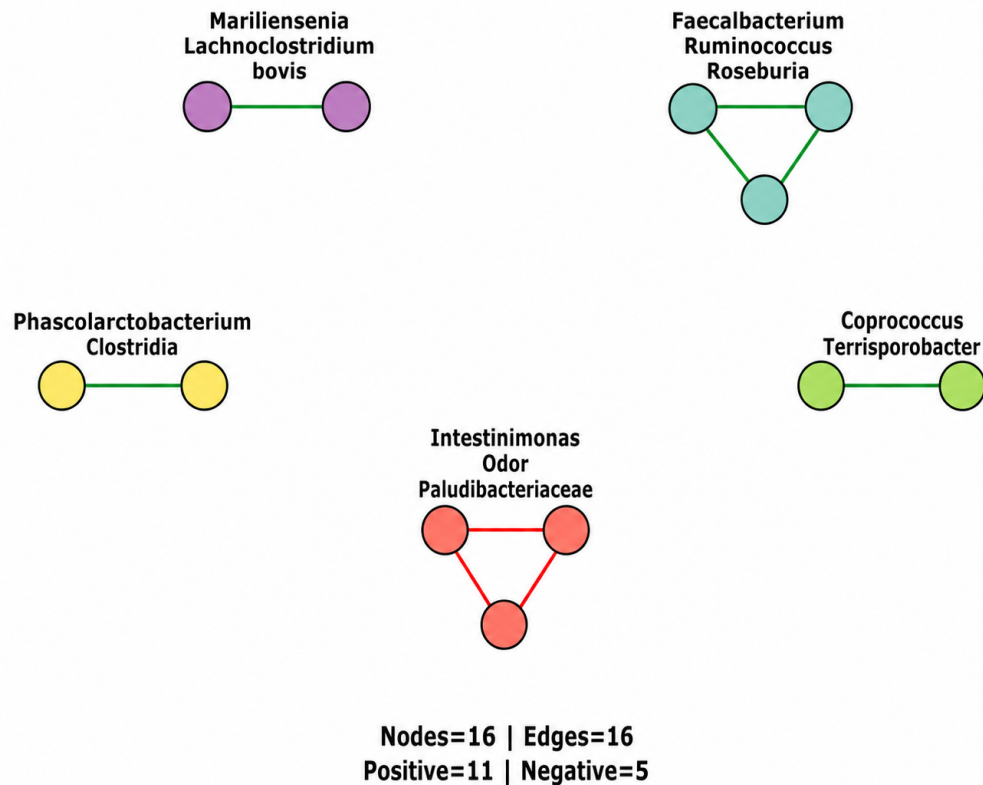


FIGURE 4.27: Normal-group microbial co-occurrence network. The network displays genus-level microbial associations in Normal samples, showing 16 nodes and 16 retained edges, including 11 positive and 5 negative relationships.

The Obese group network contained 7 nodes and 10 edges, with all retained associations classified as positive and no negative edges detected. The network showed a compact connected component among several genera, along with a separate positive association between two additional taxa. Compared with the Normal group, the Obese network had fewer nodes but retained multiple positive co-occurrence relationships, indicating a smaller but positively connected genus-level network structure. Compared with the Disease group, the Obese network showed greater microbial interaction complexity as shown in figure 4.28. The absence of negative associations in the Obese network suggests that the retained microbial relationships were primarily based on coordinated increases or similar abundance patterns among the connected genera. Although the network contained fewer taxa than the Normal network, the presence of multiple positive edges indicates that some microbial groups remained closely associated under obese conditions. This reduced

network size may reflect the loss of certain microbial interactions or a reorganisation of the gut ecosystem during the transition from a normal metabolic state to obesity. However, co-occurrence does not necessarily indicate direct biological interaction, because the observed associations may also be influenced by shared environmental conditions, dietary factors, host characteristics, or indirect relationships with other microorganisms. Therefore, these network patterns should be interpreted as indicators of microbial community organisation rather than evidence of direct functional dependence between the connected taxa.

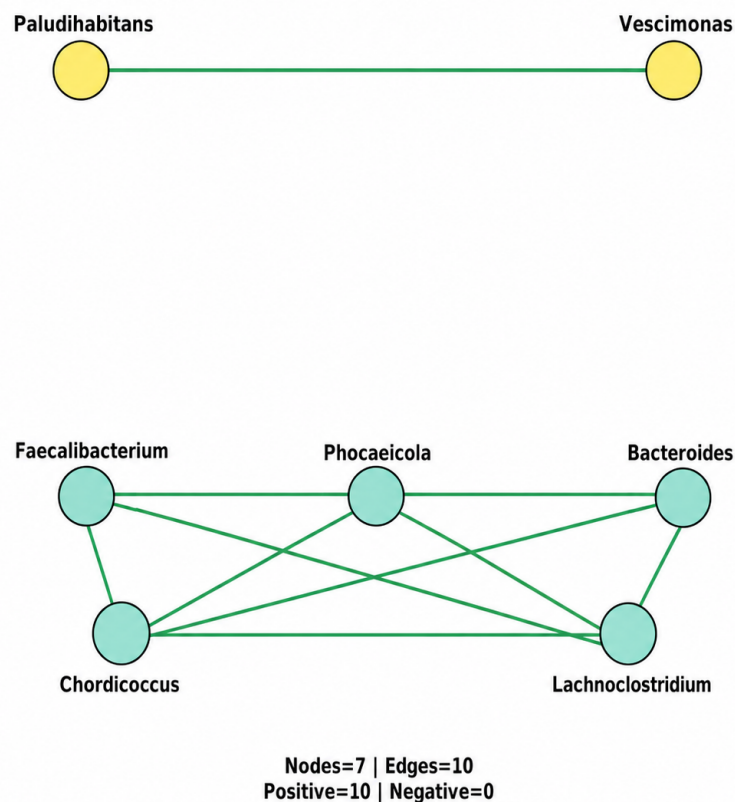


FIGURE 4.28: Obese-group microbial co-occurrence network. The network shows genus-level microbial co-occurrence associations in Obese samples, containing 7 nodes and 10 positive edges, with no negative associations detected.

Chapter 5

Discussion

The gut microbiome can be considered a functional alteration that contributes to the complexity of metabolic diseases, alongside host genetics and lifestyle/diet. The intestinal microbiota plays a role in processes related to the metabolic homeostasis of the host via mechanisms of nutrient harvesting, production of short-chain fatty acids, transformation of bile acids, immune system regulation, maintenance of intestinal barrier and modulation of systemic inflammation. This microbial ecosystem is known as gut dysbiosis and has been associated with obesity, insulin resistance, type 2 diabetes, metabolic syndrome and related inflammatory sequelae [56, 57]. Recent reviews stress the importance of metabolic disease-associated dysbiosis, which is largely characterized by changes in microbial diversity, loss of beneficial commensals and butyrate-producing bacteria, enrichment of opportunistic/pathobiont taxa, and disruption of microbial metabolic pathways [58]. The gut microbiome has emerged as an important biological interface between diet, host metabolism, immune regulation, and disease progression. The association of obesity, metabolic syndrome, type 2 diabetes, inflammatory diseases, and other chronic disease states with alterations of gut microbial composition has been repeatedly reported [57, 59]. Taxonomic, diversity, biomarker, functional, and network-level changes were observed and confirmed in this study with the comparison between the Normal, Obese and Disease groups, where disease status indicated significant taxonomic, diversity, biomarker, functional, and network-level changes. In general, the results indicate that there is a significant gut dysbiosis

in disease, with changes in species composition, lower alpha diversity, increased beta diversity separation, enrichment of several taxa and lower complexity of the microbial co-occurrence network.

The first quality check of the reads was found to be suitable for further metagenomic processing. The FastQC results indicated high quality scores per base, good GC balance, low rate of ambiguous bases, and low adapter contamination. This is relevant as the taxonomic classification and functional predictions of metagenomics sequencing reads is highly susceptible to poor quality reads, sequencing artifacts, and adapter contamination.

The paired reads retained by the next step of Trimmomatic in the various samples were high, suggesting good sequence depth in the data following quality filtering. FastQC and Trimmomatic usage conforms to typical pre-processing steps in shotgun metagenomic sequencing pipelines, such as quality control and adapter/low quality base removal prior to assembly and taxonomic classification [49, 60].

Reconstruction of the assembly and binning revealed successful metagenomic reconstruction and limitations in genome-resolved reconstruction. A large number of contigs were assembled using metaSPAdes, which served as a foundation to create bins and undertake genome level analysis. It is consistent with the nature of metaSPAdes as a metagenomic assembler for complex microbial communities [61].

The metaBAT2 binning process resulted in several initial genome bins, but the contamination values for the bins shown by CheckM were very high. By accepted standards of the MAG, the extent of genome quality should be assessed based on the completeness and contamination thresholds of the MAGs, with a medium quality MAG having 50% genome completeness and contamination 10%, and high quality MAGs having 90% genome completeness and contamination 5% [35, 62].

Hence, while binning resulted in some initial genome bins, the binned genome estimates of contamination suggest that these genomes are not high-quality MAGs and should not be taken as such. The result indicates that re-binning, more stringent contig filtering, bin refinement or other binning methods might be needed before definitive genome-resolved conclusions can be reached.

The gut microbial community was highly variable throughout the dataset using taxonomic classification by Kaiju. Kaiju can classify metagenomic reads at the protein level and is helpful for identifying microbial taxa in the shotgun metagenomic sequencing of data, especially if nucleotide-level classification fails to capture more divergent microbes [39]. In all samples there were 31 phyla, 52 classes, 114 orders, 203 families, 388 genera and 535 species identified through the taxonomic analysis. *Bacillota*-related lineages, particularly *Clostridia*, *Lachnospirales*, *Lachnospiraceae* and related genera were prevalent in the community. In line with the human gut microbiome, *Bacillota* and *Bacteroidota* are common dominant bacterial groups [56, 57]. The microbial community at genus and species level was more variable, with *Faecalibacterium*, *Blautia*, *Bacteroides*, *Bifidobacterium*, *Anaerostipes*, *Anaerobutyricum*, *Segatella*, *Dorea* and *Phocaeicola* among the genera that accounted for inter-sample variation. Several members of these genera are common organisms associated with the intestinal tract, including many species that are responsible for the fermentation of carbohydrates and the production of short-chain fatty acids. The taxonomic composition results indicated that *Bacillota*- and *Bacteroidota*-associated lineages predominated in the gut microbial community, and there was more taxonomic diversity within these classes at both genus and species levels. This was similar to the established pattern in gut microbiome research, where differences in the gut microbiome at lower taxonomic levels are more likely to be associated with biological processes such as disease than differences at higher levels [14]. At the genus level, the genera *Faecalibacterium*, *Blautia*, *Bacteroides*, *Bifidobacterium*, *Anaerostipes*, *Anaerobutyricum*, *Dorea*, *Phocaeicola*, and *Segatella* were some of the most prominent genera in the present study. For several of these genera, there have been previously known associations with metabolic health, carbohydrate fermentation, production of short-chain fatty acids, and regulation of the immune system in the intestine. Enrichment of pro-inflammatory (*Faecalibacterium*, *Roseburia*) and dysbiotic (*Firmicutes*, *Clostridium*) taxa has also been previously reported to be associated with metabolic disease states [63]. Thus, the taxonomic changes that were observed in the microbial communities were best understood at the genus and species level, rather than just the phylum level.

The visible separation between Normal, Obese and Disease samples was observed at the group-wise species level composition. Species profiles were relatively consistent for Normal and Obese groups, and were more varied for Disease groups, with more taxa or “Others” present. This is similar to the pattern observed in dysbiosis: disease-associated communities are typically characterized by shifts in dominant taxa, decreased stability, and increased heterogeneity. Plotting the significant taxa also revealed that significant taxa were found on several ranks with more significant taxa at genus and species level than at other levels. This is inevitable, as lower taxonomic levels offer greater biological detail and are more likely to reflect changes in the composition of microbes that are relevant to disease than the summary of the phylum level.

Differential abundance analysis using the Mann–Whitney U test was also used to confirm the microbial differences between the two groups. The number of significant taxa was lower for the Normal vs Obese comparison when compared to the disease-related comparisons. At phylum level, there were differences between *Actinomycetota* and *Verrucomicrobiota*, and significant differences were observed between the families: *Lachnospiraceae*, *Akkermansiaceae*, *Enterobacteriaceae*, *Oxalobacteraceae*, *Monoglobaceae* and *Turicibacteraceae*. The separation between the Normal and Obese groups was not as pronounced as the separation between the Disease groups. This is consistent with other obesity microbiome studies that have identified the differences between obesity and normal weight but have also found high inter-cohort variability and effects of diet, location, drug treatment and host factors on microbiome differences [64]. However, the Normal vs Disease comparison resulted in considerable differential abundance at all taxonomic levels with 210 species, 158 genera, 86 families and 19 phyla being significant. The positive or gut commensal genera such as *Lachnospiraceae*, *Roseburia*, *Anaerobutyricum*, *Dorea*, *Coprococcus*, *Subdoligranulum* and *Anaerostipes* were depleted in Disease samples compared to Normal samples. Biologically significant because it has been observed that depletion of butyrate-producing and anti-inflammatory gut bacteria is associated with metabolic and inflammatory disease states [15, 56, 63]. Some comparisons also revealed higher abundances of selected taxa, including *Morganelaceae* and *Odoribacteraceae*, in the Disease group. The expansion of *enterobacte-*

riaceae is often cited as a sign of inflammatory dysbiosis because the members of this family are able to carry pro-inflammatory molecular patterns and may expand under disturbed gut conditions [59].

The Obese vs Disease comparison also resulted in good separation among the species, with 18 significant phyla, 81 families, 154 genera and 211 species. The abundance of the *Bacillota* and *Actinomycetota* was greater in Obese samples, while the abundance of the *Bacteroidota* was greater in Disease samples. Several gut-associated genera and species (*Anaerobutyricum*, *Dorea*, *Roseburia*, *Coproccoccus*, *Anaerostipes*, *Bifidobacterium longum* and *Blautia obeum*) were present in higher abundance in Obese samples at lower ranks than in Disease samples, whereas other taxa (*Morganella morganii*, *Salmonella enterica* and *Odoribacter splanchnicus*) were more abundant in Disease samples. Disease associated dysbiosis is supported by the presence of opportunistic or inflammation-associated taxa in Disease samples.

The results should be viewed in context, however, as associations not causality, as the data from the cross sectional metagenomes cannot prove that a microbial shift is the cause or effect of disease [65].

The alpha diversity analysis revealed decreased species level richness and diversity for the Disease group as compared to the Normal and Obese groups. All three diversity measures (observed taxa, Shannon diversity and Simpson diversity) were lower in Disease samples, as were the values for Chao1 richness. This decrease indicates a decrease in microbial diversity and evenness, a hallmark of dysbiosis in metabolic and inflammatory diseases.

Previous studies have reported that metabolic disease such as type 2 diabetes and obesity-related metabolic dysfunction are linked with decreased microbial diversity and loss of beneficial commensals [64, 66]. The diversity reduction of disease group was the most uniform in this data set, and the groups normal and obese were the two groups that were least different.

The evidence of compositional separation at the group level was stronger for beta diversity analysis. Block-like dissimilarity patterns were displayed in the Bray-Curtis heatmap of samples and clear separation of Disease samples from Normal

and Obese samples along PCoA1 were observed in the species-level PCoA plot. PERMANOVA showed that there were significant differences among groups at all of the taxonomic levels, including phylum ($p = 0.001$), class ($p = 0.001$), order ($p = 0.001$), family ($p = 0.001$), genus ($p = 0.001$), and species ($p = 0.001$). All ranks were significant with the test statistic being highest at the phylum level and gradually decreasing towards the species level. This suggests that the changes in composition were not limited to only a few taxa, but throughout the overall microbial community structure. In microbiome studies, Bray–Curtis-based PCoA and PERMANOVA are often applied to determine if there are differences in microbial community composition among biological groups. Results of alpha and beta diversity also support the existence of disease-associated dysbiosis. This study found reduced species-level richness and diversity of observed taxa, Shannon index, Simpson index and Chao1 richness in the Disease samples, and Bray–Curtis PCoA identified the Disease samples as distinct from the Normal and Obese groups. PERMANOVA also found significant differences in communities at all taxonomic levels. The results are consistent with previous studies, reporting an association between metabolic disorders and type 2 diabetes with altered diversity indices and community restructuring, but the direction and extent of changes to diversity can differ among populations, diets or medication exposure, and analytical methods [58].

Combining statistical significance, effect size, enrichment direction, and the relevance of machine learning with the integrated XAI and Mann–Whitney biomarker bubble plots reinforced the biomarker interpretation. As for the Normal vs Disease comparison, there were many more taxa enriched in the Normal samples than in Disease samples. There was wider clustering of Obese enriched taxa than Disease enriched taxa in the Obese vs Disease comparison. To make the outputs from machine learning interpretable, the use of SHAP values in these models gives feature-level contributions to predict the outcome of the model, and hence, microbial taxa can be ranked by statistical difference as well as by their predictive importance [34]. This is crucial for biomarkers discovery, as two taxa that are statistically significant and can be important contributors to a model are not necessarily the same.

Comparisons between different diseases were highly separable at the genus and species level when using random forest biomarker classification. In the main output of Random Forest, the AUC values were 1.000 for Normal vs Disease and 1.000 for Obese vs Disease model, but the AUC value was relatively small for Normal vs Obese, with high misclassifications. This is reflected in the taxonomic and diversity results, with the Disease samples clearly differing from the other groups, and Normal and Obese samples having more similar microbial profiles. The most important Random Forest biomarkers at species level were *Dorea formicigenerans*, *Anaerobutyricum hallii*, *Blautia obeum*, *Anaerostipes hadrus*, *Dorea longicatena*, *[Ruminococcus] torques*, *Bifidobacterium longum*, *Morganella morganii* and *Salmonella enterica*. Some of these taxa have the potential to be disease discriminatory features due to either being beneficial commensals/butyrate-associated taxa or opportunistic/inflammation-associated taxa.

A species-level signature was further supported by the XAI disease biomarker analysis. For the Normal vs Disease model, the ROC curve had AUC = 1.000 and the confusion matrix had 100% accuracy of the Normal and Disease samples. In the Obese vs Disease model, the AUC = 0.976 and one sample of Obese was misclassified as Disease. *Anaerostipes hadrus*, *Anaerobutyricum hallii*, *[Ruminococcus] torques*, *Dorea longicatena*, *Faecalibacterium prausnitzii*, *Phocaeicola vulgatus* and other taxa were the top disease-positive SHAP biomarkers. Some of these organisms are included and used as useful commensals in other contexts, but in this dataset even these have been assigned a SHAP direction based on their contribution to the trained classifier, and therefore should be interpreted within the specifics of this cohort, preprocessing and model. Hence, the best overall conclusion is that these species are model-relevant discriminators of Disease status, and mechanistic roles need to be validated using functional data.

This trend was clearly summarized by the AI model evaluation heatmap and radar plots. The models of the disease had good performance at both genus and species level, and the results showed high accuracy, precision, recall, F1 score, ROC-AUC, and cross validation AUC. Significantly lower specificity and ROC-AUC, with low overall discrimination, were found in Normal vs Obese models. This reinforces the explanation that the disease status resulted in a higher microbial

signal than obesity status alone. The perfection or near perfection of some models, however, should be taken with a pinch of salt. A very high AUC in microbiome machine learning studies can be a sign of strong biological separation, but sample size, feature selection, data leakage, overfitting and/or cohort effects can also affect AUC. So, it would be important to have external validation on independent datasets before claiming this as a clinical biomarker.

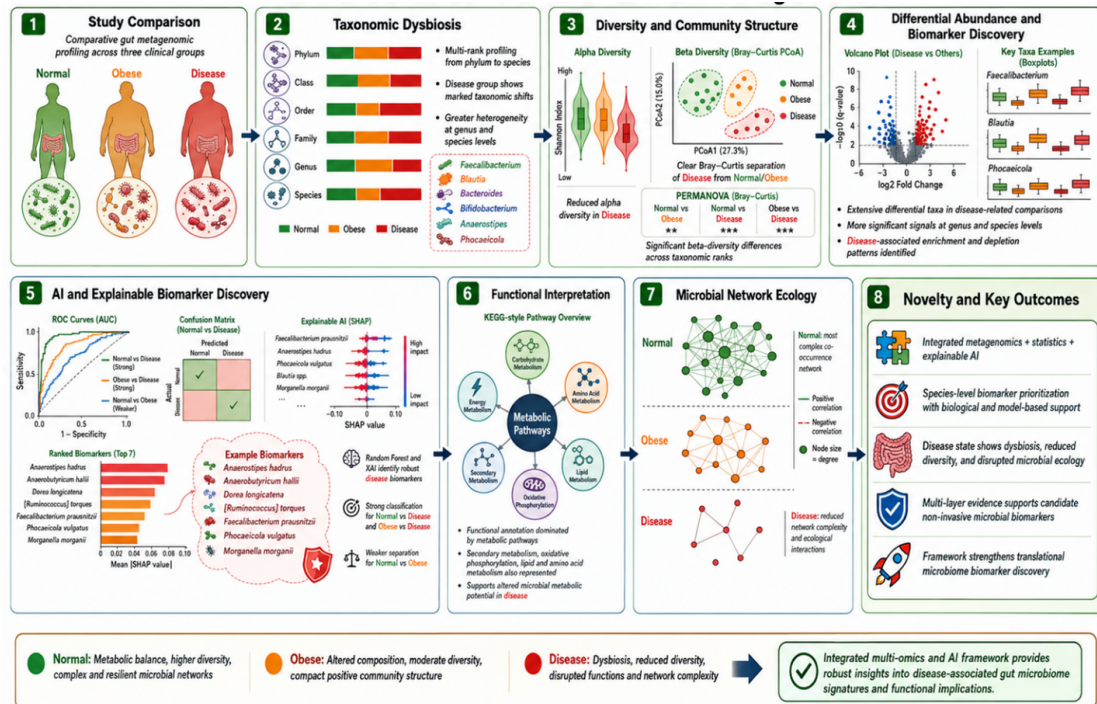


FIGURE 5.1: Graphical abstract of the integrated metagenomic and explainable AI framework for disease-associated gut microbiome biomarker discovery.

The microbial co-occurrence network analysis revealed that the Network Complexity of the Normal and Obese samples was greater than that of the Disease samples. The Normal network contained the most nodes and edges (positive and negative associations), indicating a more complex and balanced microbial interaction structure. The Obese network had less nodes and also a lot of positive associations, which means a less, yet relatively connected network. The Disease network was significantly smaller with only two nodes and one positive edge, indicating a less complex microbial network association. Less microbial network complexity could be due to disturbance of interactions between gut microbes, elimination of cooperation/competition balance, or the decreased stability of the

gut microbial community. This discovery is consistent with the alpha diversity findings, in which Disease samples had lower richness and diversity.

The graphical abstract as shown in figure 5.1 summarizes the overall discussion of this study, showing that the integrated metagenomic and explainable AI framework identified clear disease-associated gut microbiome signatures. Across the workflow, taxonomic profiling revealed multi-rank dysbiosis, diversity analysis showed reduced alpha diversity and distinct Bray–Curtis separation in the Disease group, and differential abundance testing highlighted stronger disease-related shifts at genus and species levels. Machine-learning and SHAP-based explainable AI further prioritized species-level biomarkers with strong discriminatory performance for disease-related comparisons, while functional annotation and co-occurrence network analysis indicated altered microbial metabolic potential and reduced ecological network complexity in Disease samples. Overall, this integrated framework strengthens the identification of candidate non-invasive microbial biomarkers by combining taxonomic, ecological, functional, statistical, and AI-based evidence.

Chapter 6

Conclusion and Future Work

In this study, we perform an integrated metagenomic and explainable AI analysis of the alteration of gut microbiome in three groups: Normal, Obese, and Disease. Taxonomic profiling identified the microbial diversity at various levels, and the differences in microbial profiles related to disease symptoms were more apparent at the genus level and at the species level. The Alpha diversity analysis revealed a decrease in microbial richness and diversity in the Disease group, and Bray–Curtis beta diversity and PERMANOVA indicated significant separation in microbial communities at taxonomic ranks. Differential abundance analysis also revealed that there were significant taxonomic differences in disease-related comparisons at the genus and species level. Random Forest classification combined with the Explainable AI (SHAP) was able to find strong species-level biomarker signatures that distinguished the Disease samples from both the Normal and Obese groups. High classification accuracy occurred within the disease-related models, while weak classification occurred within the Normal vs Obese models suggesting a stronger microbiome signal in the disease-related models than in the obesity-related models. Based on functional annotation, microbial genes were primarily linked to metabolic pathways, and co-occurrence network analysis revealed a decrease in microbial network complexity for Disease group. The study shows that a multi-layered approach to identifying candidate disease-associated gut microbiome biomarkers is possible with a combination of taxonomic profiling, diversity analysis, differential abundance testing, machine learning, explainable AI, functional annotation and

microbial network analysis. The results of this study indicate the possible use of gut metagenomic signatures as non-invasive biomarkers, but further validation with independent cohorts and experimental confirmation are needed for clinical use.

Future work should focus on validating the identified disease-associated microbial biomarkers in larger, independent, and geographically diverse cohorts with well-recorded clinical and lifestyle metadata, including diet, medication history, antibiotic exposure, age, sex, disease duration, and metabolic parameters. Longitudinal studies are required to determine whether the observed microbial shifts occur before disease onset, during disease progression, or as a consequence of treatment or altered host physiology. Future research should also integrate metagenomics with metatranscriptomics, metabolomics, proteomics, and host clinical data to confirm whether predicted functional pathways, particularly those related to carbohydrate metabolism, lipid metabolism, amino acid metabolism, oxidative phosphorylation, inflammation, and short-chain fatty acid production, are biologically active and reflected in measurable metabolic changes. Genome-resolved analysis should be improved using multi-binner strategies, bin-refinement tools, stringent contig filtering, and updated quality assessment approaches such as CheckM or CheckM2 to recover high-quality MAGs for strain-level and functional interpretation. In addition, future machine-learning models should include independent validation datasets, nested cross-validation, confounder adjustment, feature stability analysis, and model calibration to reduce the risk of overfitting and improve clinical reliability. Finally, experimental confirmation using targeted metagenomics, qPCR, metabolomics, and controlled biological studies will be necessary to establish the diagnostic, predictive, and therapeutic relevance of the proposed microbial biomarkers and to support their possible use in non-invasive disease screening, risk stratification, and microbiome-based interventions.

Bibliography

- [1] I. J. Neeland, S. Lim, A. Tchernof, A. Gastaldelli, J. Rangaswami, C. E. Ndumele, T. M. Powell-Wiley, and J.-P. Després, “Metabolic syndrome,” *Nature Reviews Disease Primers*, vol. 10, no. 1, p. 77, 2024.
- [2] G. Fahed, L. Aoun, M. Bou Zerdan, S. Allam, M. Bou Zerdan, Y. Bouferraa, and H. I. Assi, “Metabolic syndrome: updates on pathophysiology and management in 2021,” *International journal of molecular sciences*, vol. 23, no. 2, p. 786, 2022.
- [3] I. Pigeot and W. Ahrens, “Epidemiology of metabolic syndrome,” *Pflügers Archiv-European Journal of Physiology*, vol. 477, no. 5, pp. 669–680, 2025.
- [4] M. S. Islam, P. Wei, M. Suzauddula, I. Nime, F. Feroz, M. Acharjee, and F. Pan, “The interplay of factors in metabolic syndrome: understanding its roots and complexity,” *Molecular Medicine*, vol. 30, no. 1, p. 279, 2024.
- [5] P. Dobrowolski, A. Prejbisz, A. Kuryłowicz, A. Baska, P. Burchardt, K. Chlebus, G. Dzida, P. Jankowski, J. Jaroszewicz, and P. Jaworski, “Metabolic syndrome—A new definition and management guidelines,” *Arterial Hypertension*, vol. 26, no. 3, pp. 99–121, 2022.
- [6] T. Chomiuk, N. Niezgoda, A. Mamcarz, and D. Śliż, “Physical activity in metabolic syndrome,” *Frontiers in physiology*, vol. 15, p. 1365761, 2024.
- [7] N. K. Singh, P. Dubey, S. Singh, and R. Singh, “Pathophysiology Mechanism in Therapeutics of Metabolic Syndrome,” in *Nanotheranostics in Metabolic Syndrome: Transforming Diagnosis and Therapeutics*. Springer, 2025, pp. 33–84.

- [8] R. H. Dhondge, S. Agrawal, R. Patil, A. Kadu, and M. Kothari, “A comprehensive review of metabolic syndrome and its role in cardiovascular disease and type 2 diabetes mellitus: Mechanisms, risk factors, and management,” *Cureus*, vol. 16, no. 8, 2024.
- [9] A. A. Obeidat, M. N. Ahmad, M. A. Ghabashi, A. Y. Alazzeh, S. M. Habib, D. Abu Al-Haijaa, and F. S. Azzeh, “Developmental trends of metabolic syndrome in the past two decades: a narrative review,” *Journal of Clinical Medicine*, vol. 14, no. 7, p. 2402, 2025.
- [10] J. J. Noubiap, J. R. Nansseu, U. F. Nyaga, A. L. Ndoadoumgue, A. T. Ngouo, D. N. Tounouga, F.-L. Tianyi, A. J. Foka, E. Lontchi-Yimagou, J. R. Nkeck, and J. J. Bigna, “Worldwide trends in metabolic syndrome from 2000 to 2023: a systematic review and modelling analysis,” *Nature Communications*, vol. 17, no. 1, p. 573, 2025. [Online]. Available: <https://doi.org/10.1038/s41467-025-67268-5>
- [11] G. A. Kuziel and S. Rakoff-Nahoum, “The gut microbiome,” *Current Biology*, vol. 32, no. 6, pp. R257–R264, 2022.
- [12] P. Gebrayel, C. Nicco, S. Khodor, J. Biliński, E. Caselli, E. Comelli, M. Egert, C. Giaroni, T. Karpiński, I. Loniewski, A. Mulak, R. Julie, P. Samczuk, M. Serino, M. Sikora, A. Terranegra, M. Ufnal, R. Villéger, and C. Pichon, “Microbiota medicine: towards clinical revolution,” *Journal of Translational Medicine*, vol. 20, mar 2022.
- [13] H. A. Tegegne and T. C. Savidge, “Gut microbiome metagenomics in clinical practice: bridging the gap between research and precision medicine,” *Gut microbes*, vol. 17, no. 1, p. 2569739, 2025.
- [14] G. A. Pavlopoulos, F. A. Baltoumas, S. Liu, O. Selvitopi, A. P. Camargo, S. Nayfach, A. Azad, S. Roux, L. Call, and N. N. Ivanova, “Unraveling the functional dark matter through global metagenomics,” *Nature*, vol. 622, no. 7983, pp. 594–602, 2023.

- [15] M. Batool and J. Galloway-Peña, “Clinical metagenomics—challenges and future prospects,” *Frontiers in microbiology*, vol. 14, p. 1186424, 2023.
- [16] T. Zhou and F. Zhao, “AI-empowered human microbiome research,” *Gut*, 2025.
- [17] X.-W. Wang, T. Wang, and Y.-Y. Liu, “Artificial intelligence for microbiology and microbiome research,” *Cell Systems*, vol. 17, no. 2, 2026.
- [18] T. C. Dakal, C. Xu, and A. Kumar, “Advanced computational tools, artificial intelligence and machine-learning approaches in gut microbiota and biomarker identification.” *Frontiers in medical technology*, vol. 6, p. 1434799, 2024.
- [19] P. Przymus, K. Rykaczewski, A. Martín-Segura, J. Truu, E. Carrillo De Santa Pau, M. Kolev, I. Naskinova, A. Gruca, A. Sampri, and M. Frohme, “Deep learning in microbiome analysis: a comprehensive review of neural network models,” *Frontiers in microbiology*, vol. 15, p. 1516667, 2025.
- [20] P. Tripathi, A. Srivastava, C. K. Dubey, V. Mishra, S. Dwivedi, and A. K. Madeshiya, “Implementation of artificial intelligence (AI) and machine learning (ML) in microbiology,” in *Methods in Microbiology*. Elsevier, 2024, vol. 55, pp. 29–41.
- [21] P. Novielli, D. Romano, M. Magarelli, P. D. Bitonto, D. Diacono, A. Chiatante, G. Lopalco, D. Sabella, V. Venerito, P. Filannino, R. Bellotti, M. De Angelis, F. Iannone, and S. Tangaro, “Explainable artificial intelligence for microbiome data analysis in colorectal cancer biomarker identification.” *Frontiers in microbiology*, vol. 15, p. 1348974, 2024.
- [22] J. Fu, J. Zhang, R. He, Q. Dong, H. Mao, W. Shen, W. Wu, X. Chen, W. Ma, and Q. Zhai, “A global metagenomic atlas of aging identifies a microbiota phase transition associated with disease risk,” *npj Biofilms and Microbiomes*, 2026.
- [23] T. C. Dakal, C. Xu, and A. Kumar, “Advanced computational tools, artificial intelligence and machine-learning approaches in gut microbiota and biomarker identification,” *Frontiers in Medical Technology*, vol. 6, p. 1434799, 2025.

- [24] P. Li, M. Li, and W.-H. Chen, “Best practices for developing microbiome-based disease diagnostic classifiers through machine learning,” *Gut Microbes*, vol. 17, no. 1, p. 2489074, dec 2025. [Online]. Available: <https://doi.org/10.1080/19490976.2025.2489074>
- [25] A. R. Goncalves, H. Ranganathan, C. Valdes, H. Zhu, B. Zhang, C. R. Kok, J. M. Martí, N. J. Mulakken, J. B. Thissen, C. Jaing *et al.*, “Beyond microbial abundance: metadata integration enhances disease prediction in human microbiome studies,” *Frontiers in Microbiology*, vol. 16, p. 1695501, 2025.
- [26] Y. Cai, S.-J. Wang, Y.-Y. Tan, D.-L. Liu, S.-F. Chu, and H.-L. Li, “Association of the dietary index for gut microbiota with metabolic syndrome and its components combining interpretable machine learning algorithms.” *Journal of health, population, and nutrition*, vol. 44, no. 1, p. 410, nov 2025.
- [27] S. Romano, J. Wirbel, R. Ansorge, C. Schudoma, Q. R. Ducarmon, A. Narbad, and G. Zeller, “Machine learning-based meta-analysis reveals gut microbiome alterations associated with Parkinson’s disease,” *Nature Communications*, vol. 16, no. 1, p. 4227, 2025. [Online]. Available: <https://doi.org/10.1038/s41467-025-56829-3>
- [28] C.-C. Chang, T.-C. Liu, C.-J. Lu, H.-C. Chiu, and W.-N. Lin, “Explainable machine learning model for identifying key gut microbes and metabolites biomarkers associated with myasthenia gravis,” *Computational and Structural Biotechnology Journal*, vol. 23, pp. 1572–1583, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2001037024001041>
- [29] H. Wu, Y. Li, Y. Jiang, X. Li, S. Wang, C. Zhao, X. Yang, B. Chang, J. Yang, and J. Qiao, “Machine learning prediction of obesity-associated gut microbiota: identifying *Bifidobacterium pseudocatenulatum* as a potential therapeutic target,” *Frontiers in Microbiology*, vol. Volume 15, 2025. [Online]. Available: <https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2024.1488656>
- [30] Y. Liu, J. Zhu, H. Wang, W. Lu, Y. K. LEE, J. Zhao, and H. Zhang, “Machine learning framework for gut microbiome biomarkers discovery and modulation

- analysis in large-scale obese population,” *BMC Genomics*, vol. 23, no. 1, p. 850, 2022. [Online]. Available: <https://doi.org/10.1186/s12864-022-09087-2>
- [31] R. Waheed, F. Deeba, I. Arooj, S. Zafar, K. Ahmad, and U. Ijaz, “Role of artificial intelligence in studying metagenomics of microbes: Decoding the microbial sphere,” in *Methods in Microbiology*. Elsevier, 2025, vol. 56, pp. 147–166.
- [32] S. O. Sewe, G. Silva, P. Sicut, S. E. Seal, and P. Visendi, “Trimming and validation of Illumina short reads using trimmomatic, trinity assembly, and assessment of RNA-seq data,” in *Plant bioinformatics: Methods and protocols*. Springer, 2022, pp. 211–232.
- [33] S. Chen, “fastp 1.0: An ultra-fast all-round tool for FASTQ data quality control and preprocessing,” *Imeta*, vol. 4, no. 5, p. e70078, 2025.
- [34] A. L. Lapidus and A. I. Korobeynikov, “Metagenomic data assembly—the way of decoding unknown microorganisms,” *Frontiers in Microbiology*, vol. 12, p. 613791, 2021.
- [35] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, “metaSPAdes: a new versatile metagenomic assembler,” *Genome research*, vol. 27, no. 5, pp. 824–834, may 2017.
- [36] V. Mallawaarachchi and Y. Lin, “Accurate binning of metagenomic contigs using composition, coverage, and assembly graphs,” *Journal of Computational Biology*, vol. 29, no. 12, pp. 1357–1376, 2022.
- [37] D. D. Kang, J. Froula, R. Egan, and Z. Wang, “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities,” *PeerJ*, vol. 3, p. e1165, 2015.
- [38] A. Chklovski, D. H. Parks, B. J. Woodcroft, and G. W. Tyson, “CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning,” *Nature methods*, vol. 20, no. 8, pp. 1203–1212, 2023.

- [39] P. Menzel, K. L. Ng, and A. Krogh, “Fast and sensitive taxonomic classification for metagenomics with Kaiju,” *Nature Communications*, vol. 7, no. 1, p. 11257, 2016. [Online]. Available: <https://doi.org/10.1038/ncomms11257>
- [40] Z. Ling, Y. Cheng, Z. Lan, X. Liu, Z. Zhu, W. Ding, X. Xu, P. Yu, X. Xu, and L. Shao, “Gut mycobiota dysbiosis and systemic immune dysfunction in Chinese schizophrenia patients with metabolic syndrome,” *Frontiers in Immunology*, vol. 16, p. 1652633, 2025.
- [41] Y. Xia and J. Sun, “Alpha diversity,” in *Bioinformatic and statistical analysis of microbiome data: from raw sequences to advanced modeling with QIIME 2 and R*. Springer, 2023, pp. 289–333.
- [42] C. Casals-Pascual, A. González, Y. Vázquez-Baeza, S. J. Song, L. Jiang, and R. Knight, “Microbial diversity in clinical microbiome studies: sample size and statistical power considerations,” *Gastroenterology*, vol. 158, no. 6, pp. 1524–1528, 2020.
- [43] J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, and R. Knight, “Microbial community resemblance methods differ in their ability to detect biologically relevant patterns,” *Nature methods*, vol. 7, no. 10, pp. 813–819, 2010.
- [44] Y.-J. Hu and G. A. Satten, “A rarefaction-without-resampling extension of PERMANOVA for testing presence–absence associations in the microbiome,” *Bioinformatics*, vol. 38, no. 15, pp. 3689–3697, 2022.
- [45] L. Yang and J. Chen, “A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions,” *Microbiome*, vol. 10, no. 1, pp. 1–23, 2022. [Online]. Available: <https://doi.org/10.1186/s40168-022-01320-0>
- [46] J. T. Nearing, 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. [Online]. Available: <https://doi.org/10.1038/s41467-022-28034-z>
- [47] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, may 1945. [Online]. Available: <http://www.jstor.org/stable/3001968>
- [48] J. Hunt, “Machine Learning in Python,” vol. 12, pp. 633–642, 2023.
- [49] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. [Online]. Available: <https://doi.org/10.1145/2939672.2939785>
- [50] J. Kim, S. Mun, S. Lee, K. Jeong, and Y. Baek, “Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea,” *BMC Public Health*, vol. 22, no. 1, p. 664, 2022.
- [51] J.-K. Kim, S. Mun, and S. Lee, “Detection and analysis of circadian biomarkers for metabolic syndrome using wearable data: cross-sectional study,” *JMIR Medical Informatics*, vol. 13, p. e69328, 2025.
- [52] M. Sokolova, N. Japkowicz, and S. Szpakowicz, “Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation,” in *Australasian joint conference on artificial intelligence*. Springer, 2006, pp. 1015–1021.
- [53] S. Sathyanarayanan and B. R. Tantri, “Confusion matrix-based performance evaluation metrics,” *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023–4031, 2024.
- [54] F. Vahid, R. Homayounfar, M. Farjam, and T. Bohn, “A Global Nutritional Tool for Monitoring Westernized Dietary Transition: Validation of the Westernized Diet Index Using a Large Population Sample and Biomarkers of Metabolic Health,” *Nutrients*, vol. 18, no. 2, p. 349, 2026.

- [55] R. J. Williams, A. Howe, and K. S. Hofmockel, “Demonstrating microbial co-occurrence pattern analyses within and between ecosystems,” *Frontiers in microbiology*, vol. 5, p. 358, 2014.
- [56] Y. Fan and O. Pedersen, “Gut microbiota in human metabolic health and disease,” *Nature Reviews Microbiology*, vol. 19, no. 1, pp. 55–71, 2021. [Online]. Available: <https://doi.org/10.1038/s41579-020-0433-9>
- [57] W. M. D. Vos, H. Tilg, M. V. Hul, and P. D. Cani, “Gut microbiome and health : mechanistic insights,” pp. 1020–1032, 2022.
- [58] M. S. Alqahtani, “The Gut Microbiota-Metabolic Axis: Emerging Insights from Human and Experimental Studies on Type 2 Diabetes Mellitus-A Narrative Review.” *Medicina (Kaunas, Lithuania)*, vol. 61, no. 11, nov 2025.
- [59] M. Gurung, Z. Li, H. You, R. Rodrigues, D. B. Jump, A. Morgun, and N. Shulzhenko, “Role of gut microbiota in type 2 diabetes pathophysiology,” *EBioMedicine*, vol. 51, p. 102590, 2020. [Online]. Available: <https://doi.org/10.1016/j.ebiom.2019.11.051>
- [60] S. Chen, Y. Zhou, Y. Chen, and J. Gu, “fastp: an ultra-fast all-in-one FASTQ preprocessor,” *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, sep 2018. [Online]. Available: <https://doi.org/10.1093/bioinformatics/bty560>
- [61] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes,” *Genome research*, vol. 25, no. 7, pp. 1043–1055, 2015.
- [62] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Elie-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. Ettema, S. Tighe, K. T. Konstantinidis, W. T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D.

- McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, and T. Woyke, “Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea,” *Nature Biotechnology*, vol. 35, no. 8, pp. 725–731, 2017.
- [63] S. Chong, M. Lin, D. Chong, S. Jensen, and N. S. Lau, “A systematic review on gut microbiota in type 2 diabetes mellitus,” *Frontiers in Endocrinology*, vol. Volume 15, 2025. [Online]. Available: <https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2024.1486793>
- [64] M. Pinart, A. Dötsch, K. Schlicht, M. Laudes, J. Bouwman, S. K. Forslund, T. Pischon, and K. Nimptsch, “Gut Microbiome Composition in Obese and Non-Obese Persons: A Systematic Review and Meta-Analysis.” *Nutrients*, vol. 14, no. 1, dec 2021.
- [65] J. Siptroth, O. Moskalenko, C. Krumbiegel, J. Ackermann, I. Koch, and H. Pospisil, “Variation of butyrate production in the gut microbiome in type 2 diabetes patients,” *International Microbiology*, vol. 26, no. 3, pp. 601–610, 2023.
- [66] S. Verhoog, P. E. Taneri, Z. M. Roa Díaz, P. Marques-Vidal, J. P. Troup, L. Bally, O. H. Franco, M. Glisic, and T. Muka, “Dietary Factors and Modulation of Bacteria Strains of *Akkermansia muciniphila* and *Faecalibacterium prausnitzii*: A Systematic Review.” *Nutrients*, vol. 11, no. 7, jul 2019.