AI-Guided Gut Microbiome Profiling for the Diagnostics of Neurological Disorders

By

Arisha Wasim

(Registration No: 00000401274)

Department of Sciences

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

August 2024

AI-Guided Gut Microbiome Profiling for the Diagnostics of Neurological Disorders

By

Arisha Wasim

(Registration No: 00000401274)

A thesis submitted to the National University of Sciences & Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in
Bioinformatics

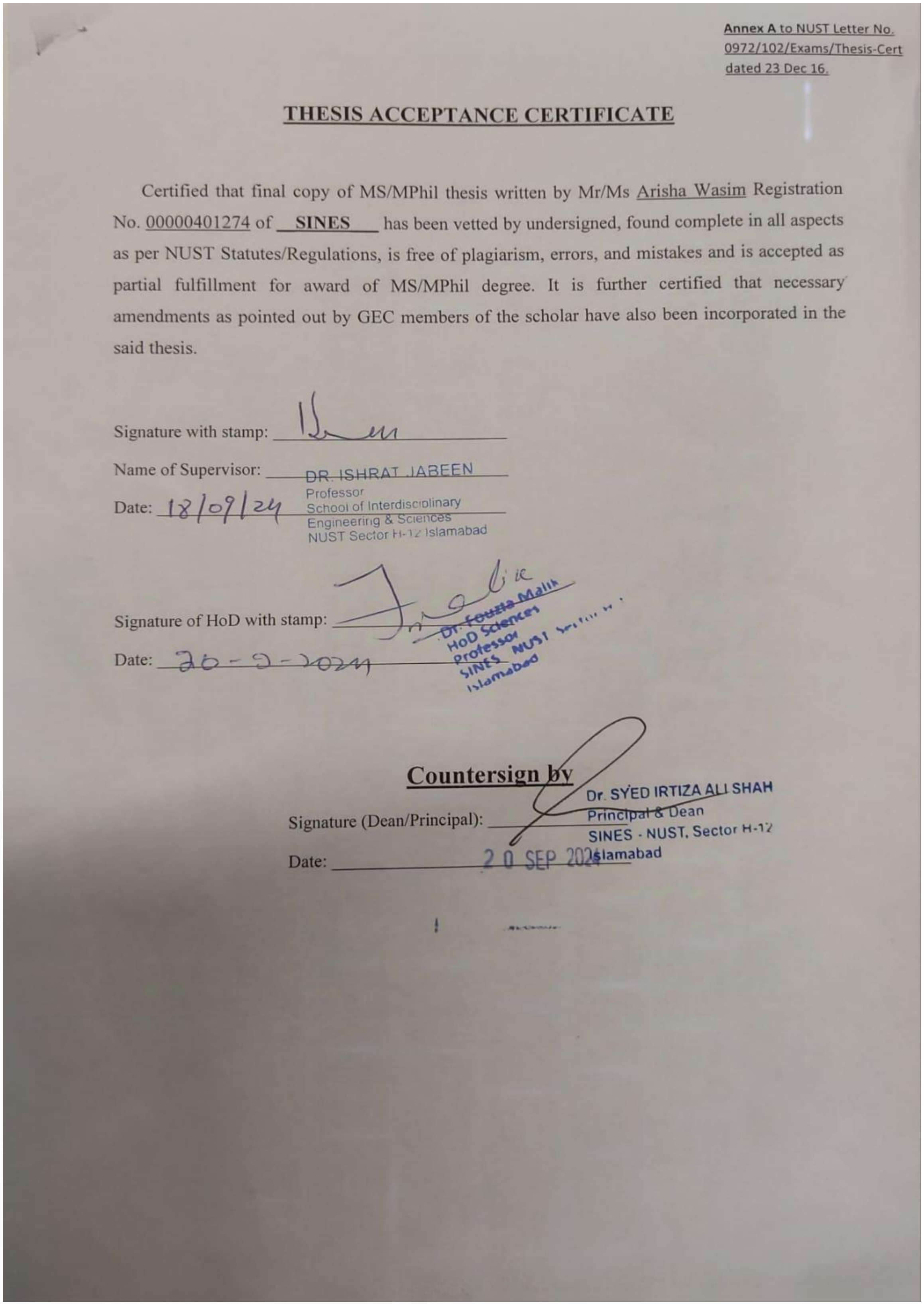
Supervisor: Prof. Dr. Ishrat Jabeen

School of Interdisciplinary Engineering & Sciences (SINES)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

August 2024



Annex A to NUST Letter No.
0972/102/Exams/Thesis-Cert
dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr/Ms Arisha Wasim Registration No. 00000401274 of **SINES** has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: ____________

Name of Supervisor: DR. ISHRAT JABEEN
Professor
School of Interdisciplinary
Engineering & Sciences
NUST Sector H-12 Islamabad

Date: 18/09/24

Signature of HoD with stamp: ____________
Dr. Fouzia Malik
HoD Sciences
Professor
SINES NUST [illegible]
Islamabad

Date: 20-9-2024

Countersign by

Signature (Dean/Principal): ____________
Dr. SYED IRTIZA ALI SHAH
Principal & Dean
SINES - NUST, Sector H-12
Islamabad

Date: 20 SEP 2024

AUTHOR'S DECLARATION

I hereby state that my MS thesis titled "**AI-Guided Gut Microbiome Profiling for the Diagnostics of Neurological Disorders**" is my work and has not been submitted previously by me for taking any degree from the National University of Sciences & Technology, Islamabad, or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Arisha Wasim

Date: August 30, 2024

PLAGIARISM UNDERTAKING

I solemnly declare that the research work presented in the thesis titled "**AI-Guided Gut Microbiome Profiling for the Diagnostics of Neurological Disorders**" 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 the National University of Sciences and Technology (NUST), Islamabad 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 to/cited.

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

Student Signature: _______________

Name: Arisha Wasim

DEDICATION

I dedicate this thesis to my exceptional parents, siblings, friends, and teachers whose unconditional love, support, and guidance led me to this world of accomplishment.

ACKNOWLEDGEMENTS

All praise is for **Almighty Allah**, the ultimate source of all knowledge. By His grace, I have reached this stage of knowledge with the ability to contribute something beneficial to His creation. My deepest respects are to the **Holy Prophet Hazrat Muhammad (PBUH)**, the symbol of guidance and fountain of knowledge.

I earnestly thank my supervisor, **Prof. Dr. Ishrat Jabeen**, for her keen interest, invaluable guidance, encouragement, and continuous support throughout my research journey. I am extremely grateful for her thought-provoking discussions, sound advice, and valuable suggestions. Her mentorship has enabled me to tackle problems more meaningfully and provided me with the resources to pursue my objectives diligently and sincerely.

I am also thankful to my GEC committee members, **Dr. Masood Ur Rehman Kayani** and **Dr. Tariq Saeed**, who guided me throughout my project and offered valuable feedback and suggestions to refine my thesis. Additionally, I acknowledge all the faculty members of SINES for their kind assistance at various phases of this research.

My gratitude extends to my colleagues in the **Computational Drug Design research group**. Special thanks to **Muqadas Chaudhary**, **Amna Tariq**, **Minahil Tariq**, and **Haseeb Manzoor** for their continuous help and feedback at every stage of the research. I am also grateful to my friends **Faheem Raziq, Adnan Tariq,** and **Abdullah Ajmal** for their unwavering support. Lastly, I am deeply thankful to my parents and siblings for their immense love and support throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GBA = Gut-Brain Axis

CNS = Central Nervous System

ENS = Enteric Nervous System

HPA = Hypothalamic-Pituitary-Adrenal

ML = Machine Learning

AD = Alzheimer’s Disease

PD = Parkinson’s Disease

FTD = Frontotemporal Dementia

ADNI = Alzheimer’s Disease Neuroimaging Initiative

PRODIGAL = PROkaryotic Dynamic programming Gene-finding Algorithm

QUAST = Quality Assessment

ANN = Artificial Neural Network

SVM = Support Vector Machine

KNN = K-Nearest Neighbor

XGBoost = eXtreme Gradient Boosting

RF = Random Forest

DAPs = Differentially Abundant Proteins

Abstract

The gut microbiome is a complex ecosystem of microorganisms that reside in the gastrointestinal tract (GIT). The interaction between the gut microbiota and the brain is often called the microbiota/gut-brain axis, which is a bidirectional relationship. Any imbalance in the gut microbiome through dietary changes, medication use, lifestyle choices, environmental factors, and aging has a potential pathophysiological impact on the body in general and CNS in particular.

Early studies linking alterations in the gut microbiome with neurobehavioral phenotypes launched the concept of a microbiota-gut-brain axis whereby intestinal microbiota can influence brain function and behavior. This suggests that targeting the gut microbiome could be a potential strategy for developing new therapies for neurological and neurodegenerative diseases. Previously, various studies implicated the microbiome as one of the key susceptibility factors for neurological disorders. As a result, recent research has focused on identifying potential therapeutic interventions for neurological and neurodegenerative diseases by exploiting the gut microbiome profiles. However, there is a high level of overlap between neurological disease state and normal condition microbiome profiles, which makes it difficult to develop specialized treatments. In addition to this, for now, only a symptomatic cure exists. Therefore, this study aims to leverage machine learning (ML) modeling techniques to identify potential targets within the gut microbiome associated with a neurological disease.

In this study, predictive modeling techniques were employed, including ensemble methods such as XGBoost, alongside other models like Artificial Neural Networks (ANN), Support Vector Machines (SVM) with a non-linear polynomial kernel up to the fifth degree, K-Nearest Neighbors (KNN), and Random Forest. The results indicated that all models converged around 60% accuracy, with specificity and sensitivity metrics following similar trends. These models demonstrated the potential to analyze differentially expressed biomarkers, highlighting areas for future research. Future work should focus on developing models specifically designed for addressing biological problems rather than merely predictive modeling. Additionally, it is essential to examine the species level and the mechanistic aspects of the microbial profile to understand the underlying factors that transition conditions from normal to diseased states.

Keywords: Neurological disease, Microbiome, Machine learning, Feature Extraction, Predictive Modeling.

1 CHAPTER 1: Introduction

1.1 Neurological Disorders

Neurological disorders encompass a wide range of conditions affecting the brain, spinal cord, and nerves. These disorders can have various causes, manifest with diverse symptoms, and require different treatment approaches. Common causes include genetic factors, such as Huntington's disease and certain forms of epilepsy, and infections like meningitis and encephalitis. Traumatic injuries, including head injuries and spinal cord injuries, can also lead to long-term neurological problems like traumatic brain injury (TBI) and chronic traumatic encephalopathy (CTE) [1]. Degenerative diseases, such as Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis (ALS), often arise with age. Autoimmune disorders like multiple sclerosis (MS) and Guillain-Barré syndrome result from the immune system attacking the nervous system, while metabolic and nutritional deficiencies, including Wilson's disease and B12 deficiency, can also cause neurological symptoms [2].

Symptoms of neurological disorders vary widely depending on the specific condition but generally include cognitive decline (memory loss, confusion, difficulty concentrating), motor symptoms (tremors, rigidity, muscle weakness, coordination problems), sensory symptoms (numbness, tingling, pain, sensory loss), seizures (uncontrolled electrical activity in the brain leading to convulsions or loss of consciousness), and behavioral changes (mood swings, depression, anxiety, personality changes) [3].

Treatment approaches for neurological disorders depend on the specific condition and can include medications, surgical interventions, rehabilitation therapies, lifestyle changes, and supportive care. Medications such as antiepileptic drugs for epilepsy, dopaminergic medications for Parkinson's disease, cholinesterase inhibitors for Alzheimer's disease, and immunomodulatory drugs for MS are commonly used. Surgical interventions like deep brain stimulation for Parkinson's disease and neurosurgery for brain tumors and certain epilepsy cases can also be effective [4]. Rehabilitation therapies, including physical, occupational, and speech and language therapy, play a crucial role in managing these disorders. Lifestyle and supportive care, such as nutritional support, exercise, psychological counseling, and support groups, are important for overall well-being. Innovative treatments like gene therapy and stem cell therapy are also emerging as potential options.

Generally, neurological disorders represent a significant global health challenge due to their impact on quality of life, economic burden, and the need for long-term care and management. Worldwide prevalence varies by condition, with approximately 50 million people living with dementia (mostly Alzheimer’s), over 10 million affected by Parkinson's disease, around 50 million with epilepsy, about 2.8 million with multiple sclerosis, and stroke affecting about 13.7 million new cases each year. Migraine is also highly prevalent, affecting approximately 1 billion people globally [5]. As shown in **Figure 1.1** The diverse nature of neurological disorders underscores the importance of continued research and advances in treatment to improve outcomes for those affected.

Neurological disorders are one of the leading causes of disabilities in the world as shown in **Figure 1.1**, while their mortality rate is the second leading cause [6]. In the previous thirty years, the number of deaths due to neurological diseases and the number of people with disabilities due to the same cause has risen significantly, more particularly in low and middle-income countries.

In addition to this, further increases are expected to be seen worldwide due to aging and population growth [7].

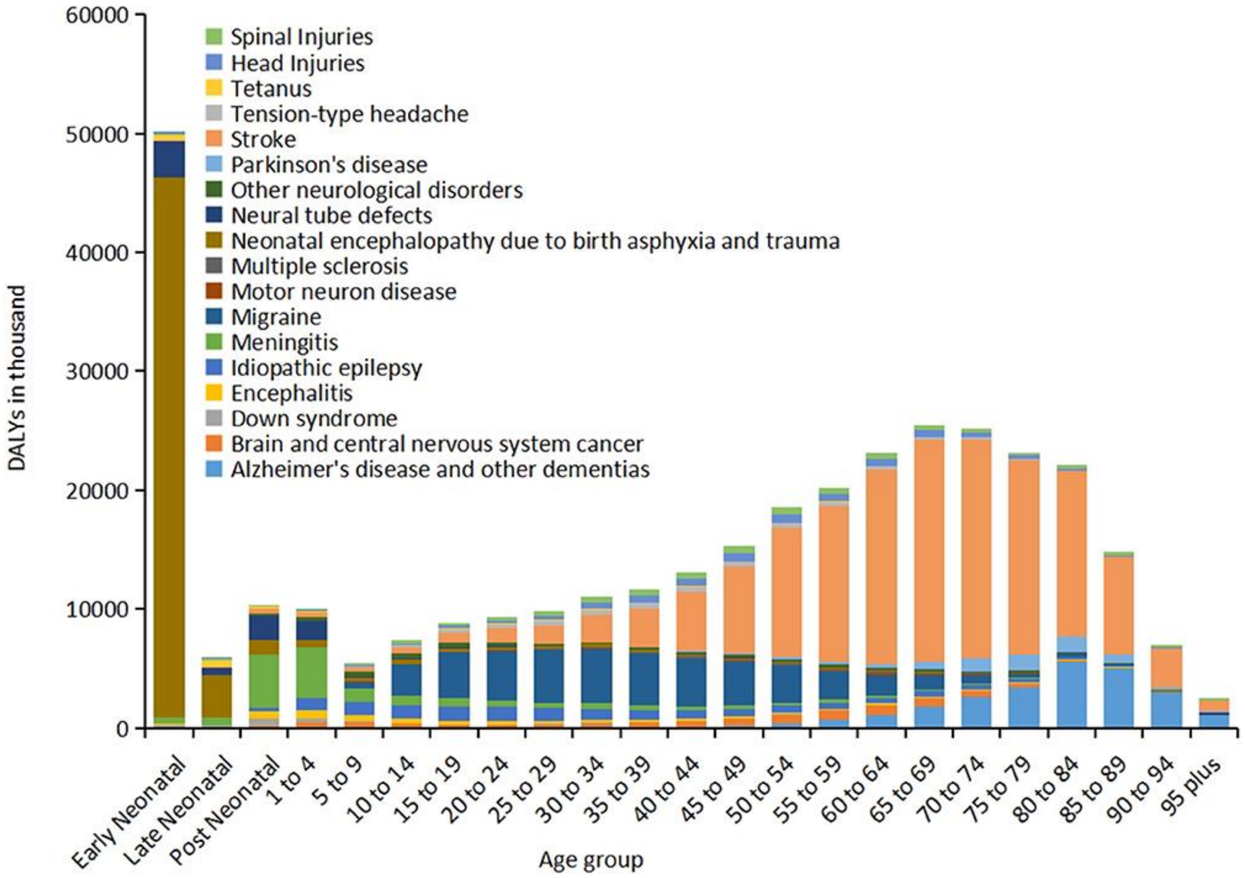


Figure 1.1 Distribution of disability-adjusted life years (DALYs) by neurological condition and age group: This graph illustrates the burden of various neurological conditions across different age group

1.2 Brief Overview of the Mechanistic of Neurological Disorders

Neurobiological pathways reveal the complex interactions between stress, neuronal damage, and genetic predispositions, which contribute to the development and progression of conditions like depression and anxiety.

Stress is a pivotal factor in the pathogenesis of mood disorders, such as depression and anxiety, through a multitude of neurobiological mechanisms. Prolonged exposure to stress exacerbates oxidative stress, which contributes to neuronal damage, particularly in the hippocampus—a brain region integral to mood regulation [8]. Additionally, chronic stress impairs immune function, promoting inflammation, while genetic predispositions modulate individual responses to stress, determining resilience or susceptibility to mood disturbances[9].

The frequent comorbidity of mood disorders with neurological diseases further complicates their clinical management. For instance, increased iron deposition in the thalamus has been correlated with depressive symptoms in geriatric populations, while post-stroke anxiety markedly diminishes quality of life and functional outcomes[10]. Furthermore, the interplay between depression and insomnia is linked to decreased gray matter volume in the orbitofrontal cortex, and alterations in cortical excitability are observed in patients with anxiety disorders. These findings elucidate the complex interrelationship between mood disorders and neurological conditions, underscoring the necessity for integrative treatment approaches and a deeper understanding of these multifaceted interactions[11].

1.3 Symptomatic Diagnosis of Neurological Disorders

For many neurological disorders, the current state of medical knowledge and technology means that only symptomatic diagnosis is available. This approach focuses on identifying and assessing the symptoms presented by the patient, as opposed to pinpointing an underlying cause or definitive biomarker. The diagnostic process begins with a detailed medical history, where healthcare providers gather information about the onset, duration, and progression of symptoms. They inquire about the patient's family history, lifestyle factors, and any previous medical conditions or treatments that might be relevant. This thorough assessment helps in forming a preliminary understanding of the disorder.

A physical examination is conducted to assess general health and identify any physical signs of neurological issues. This is followed by a neurological examination, which evaluates cognitive function, motor skills, sensory perception, reflexes, coordination, and balance. Through these tests, physicians can detect abnormalities that suggest specific neurological conditions. To further investigate the symptoms, various diagnostic tests and imaging studies are employed. These may include MRI (Magnetic Resonance Imaging) and CT (Computed Tomography) scans, which provide detailed pictures of the brain and spinal cord, helping to identify structural abnormalities, tumors, or lesions. An electroencephalogram (EEG) measures electrical activity in the brain and is particularly useful for diagnosing epilepsy and other seizure disorders. Nerve conduction studies and electromyography (EMG) assess the electrical activity of muscles and nerves, aiding in the diagnosis of conditions like peripheral neuropathy and myopathy [12]. Additionally, a lumbar puncture (spinal tap) can be performed to analyze cerebrospinal fluid, which can reveal infections, bleeding, and other neurological conditions [13].

In summary, symptomatic diagnosis involves a comprehensive evaluation of the patient's medical history, physical and neurological examinations, and various diagnostic tests aimed at understanding the nature and severity of the symptoms. While this approach does not provide a definitive cause or biomarker for many neurological disorders, it is essential for developing an effective treatment plan to manage and alleviate symptoms.

1.4 Limitation of Conventional Therapies

Conventional therapies for neurological disorders face significant limitations that impede their effectiveness and accessibility. A primary challenge is the difficulty in achieving adequate drug concentrations within the central nervous system due to the restrictive nature of the blood-brain barrier (BBB). This barrier, while protective, often limits the penetration of therapeutic agents, resulting in suboptimal treatment outcomes for neurodegenerative diseases such as Alzheimer's and Parkinson's[14]. Moreover, these therapies tend to focus on symptomatic relief rather than addressing the underlying pathophysiology of the disorders. For instance, cholinesterase inhibitors used in Alzheimer's disease may alleviate cognitive symptoms but fail to arrest disease progression. Additionally, the side effects associated with many conventional medications, such as the cholinergic and cardiac issues seen with Donepezil, can reduce patient compliance and limit therapeutic success[15].

Further complicating treatment efficacy are the resource limitations and the heterogeneity of neurological disorders, which challenge the development of standardized treatment protocols and lead to potential misdiagnoses or delayed interventions. The inherent complexity of these disorders also presents obstacles in research and development, with drug discovery efforts often hampered by the need for better biomarkers and more precise outcome measures[16]. Consequently, the modest clinical benefits observed with some treatments underscore the necessity for innovative approaches and advanced drug delivery systems to enhance both therapeutic efficacy and patient outcomes in managing neurological conditions.

1.5 The Gut Microbiome as a Potential Therapy

The gut microbiome has emerged as a compelling target for therapeutic interventions in neurological diseases, offering novel avenues for treatment beyond conventional approaches. Modulating the composition of the gut microbiome through probiotics, prebiotics, or fecal microbiota transplantation (FMT) has shown promise in restoring microbial balance and potentially mitigating neurological symptoms. Beneficial bacteria such as Lactobacillus and Bifidobacterium, for instance, have been linked to improvements in cognition and reductions in anxiety and depression[17].

Moreover, improving gut barrier function, which is often compromised in neurological disorders, could prevent the translocation of harmful substances and pathogens, thereby exerting neuroprotective effects. Early intervention strategies aimed at modulating the gut microbiome during the prodromal stages of neurological diseases, or even earlier in life, hold potential for delaying disease onset and progression[18]. Given the highly individualized nature of the gut microbiome, personalized therapeutic approaches tailored to each patient’s unique microbial profile are increasingly being explored, leveraging advances in microbiome profiling and machine learning.

2 CHAPTER 2: Literature Review

2.1 Diagnostics of Neurological Disorders and the Role of Predictive Modeling

Diagnosing neurological disorders is a complex and multifaceted process, necessitating a variety of diagnostic tests and procedures due to the diverse and often overlapping symptoms associated with different conditions. The complexity is further compounded by the limitations of available diagnostic tools. The process typically begins with a thorough neurological examination, which assesses multiple functions such as movement, sensation, coordination, balance, mental status, and mood. This initial evaluation forms the basis for a differential diagnosis and guides subsequent testing. Advanced imaging techniques, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans, are integral to the diagnostic process, providing detailed images of brain structures. MRI is particularly valuable for visualizing soft tissues and detecting abnormalities, while CT scans offer rapid assessment of structural issues, especially in emergency settings. In addition to imaging, electrodiagnostic tests like Electroencephalograms (EEGs) and electromyography (EMG) play a crucial role in assessing brain and nerve function[19].

Genetic testing has also emerged as a vital component in diagnosing hereditary neurological disorders, enabling the detection of specific genetic mutations that inform diagnosis and treatment planning[20].

The diagnostic process is further complicated by several factors, including the significant symptom overlap among neurological disorders. For instance, symptoms such as headaches, memory loss, and cognitive impairments are common across various conditions, making it challenging to distinguish between disorders like Alzheimer's disease and other forms of dementia[21]. Moreover, the lack of definitive biomarkers for certain neurological conditions adds to the difficulty, often leading to diagnostic uncertainty, misdiagnosis, or delays in treatment. The evolving nature of neurological disorders, where symptoms may progress or change over time, requires continuous assessment and sometimes repeated testing to monitor these changes and adjust diagnoses accordingly[22].

Technological limitations also present challenges in the diagnostic process. Despite significant advancements, current imaging and diagnostic technologies still face issues related to sensitivity and specificity, which can hinder the detection of early-stage diseases or subtle changes in brain function. Furthermore, the interdisciplinary nature of diagnosing neurological disorders, which often involves collaboration among neurologists, radiologists, geneticists, and other specialists, adds another layer of complexity. While this approach enhances diagnostic accuracy, it also requires careful coordination of care.

The table below provides a detailed overview of the diagnostics for neurological disorders, including the types of data used, the models employed, the parameters optimized, and the relevance of these methods for accurate diagnosis and treatment. This comprehensive summary highlights the approach and effectiveness of various diagnostic techniques in the context of neurological disease research.

Table 2.1 The table below outlines the diagnostics for neurological disorders, detailing the data types, models employed, parameters used, and their applicability.

Study	**Data Used**	**Models Used**	**Parameters**	**Applicability**	**Reference**
Machine Learning Analysis Reveals Biomarkers for the Detection of Neurological Diseases	The study analyzed clinical and genotyping data from 1,223 participants with neurological diseases (AD, PD, MND, MG) from the UK Biobank, including blood and urine biomarkers, cognitive test scores, and data on 820,967 SNPs.	Multinomial generalized linear model to predict neurological disease (NLD) diagnosis using clinical markers.	The multinomial model was optimized through Monte Carlo randomization to identify the most informative clinical variables, urine tests, and cognitive tests.	The study shows that data-driven methods can identify new diagnostic biomarkers and shared genetic risk factors for neurological diseases (NLDs), with the multinomial model predicting diagnoses with 88.3% accuracy.	[23]
Machine Learning and Novel Biomarkers for the	The study used datasets featuring Alzheimer's disease biomarkers, including	Support Vector Machines (SVM), Random Forest, Logistic Regression,	Biomarker Levels, Imaging Features, Metabolite Profiles, and Model Evaluation Metrics: Area under	This study highlights the importance of combining machine learning with biomarker analysis to	[24]

Diagnosis of Alzheimer’s Disease	cerebrospinal fluid (CSF) markers like amyloid-β, tau protein, and hyperphosphorylated tau, as well as neuroimaging data from PET scans and metabolomics data to identify novel AD-related metabolites.	Convolutionals Neural Networks, Ensemble Learning Techniques	the curve (AUC), accuracy, sensitivity, and specificity for model performance and evaluation.	improve early Alzheimer's disease diagnosis, crucial for timely intervention. The findings suggest that machine learning can effectively detect patterns in complex biomarker data.	
Gut microbiome composition may be an indicator of preclinical Alzheimer’s disease	The study used data from ADNI and AIBL: taxonomic composition and gut microbial function between 164 cognitively normal individuals, including 49 with early preclinical AD biomarkers.	DeepSleepNet deep learning model, originally designed for sleep stage classification, to classify Alzheimer's disease, mild cognitive impairment, and	The DeepSleepNet model was trained on 7 days of actigraphy data, with hyperparameters like learning rate, batch size, and epochs optimized through cross-validation to best classify AD,	The study highlights the potential of using actigraphy-based sleep-wake patterns with deep learning for early Alzheimer's disease detection. The DeepSleepNet model achieved 89.1% accuracy in classifying AD, MCI, and cognitively	[25]

		cognitively normal individuals.	MCI, and cognitively normal individuals.	normal individuals, suggesting sleep-wake disturbances as a promising AD biomarker. Actigraphy offers a non-invasive, cost-effective alternative to CSF and PET scans, potentially enabling earlier interventions and personalized treatments.	
Parkinson’s disease-associated alterations of the gut microbiome predict disease-relevant changes in	Stool samples from the Luxembourg Parkinson’s Study (n = 147 typical PD cases, n = 162 controls).	Random Forest and Support Vector Machine	The study used feature selection to identify key microbial taxa linked to PD, optimized RF and SVM hyperparameters via cross-validation, and evaluated model performance with	The study highlights the gut microbiome's role in Parkinson's disease, suggesting that microbial alterations could serve as biomarkers for early diagnosis and progression monitoring. It also	[26]

metabolic functions			accuracy, sensitivity, and specificity metrics.	indicates that changes in gut microbiome metabolism may lead to targeted microbiome-based therapies.	
Multi-omics analyses of serum metabolome, gut microbiome and brain function reveal dysregulated microbiota-gut-brain axis in bipolar depression	A cohort of 109 unmedicated bipolar disorder (BD) patients and 40 healthy controls (HCs) was studied to characterize the microbial-gut-brain axis in BD.	Random Forest and Support Vector Machine and PCA analysis	Feature Selection, Hyperparameter Tuning, Evaluation Metrics: Used accuracy, sensitivity, and specificity to assess model performance.	The study highlights dysregulation in gut microbiota and neuroactive metabolites in bipolar depression, suggesting potential biomarkers for diagnosis and treatment.	[27]

2.2 The Gut Microbiome

The gut microbiome is a very complex and dynamic ecosystem that contains millions upon billions of microorganisms, including bacteria, archaea, viruses, and eukaryotic microbes, that are present in the gastrointestinal tract (GIT). This ecosystem plays a very important role in maintaining the host's health, all while influencing many physiological processes [28].

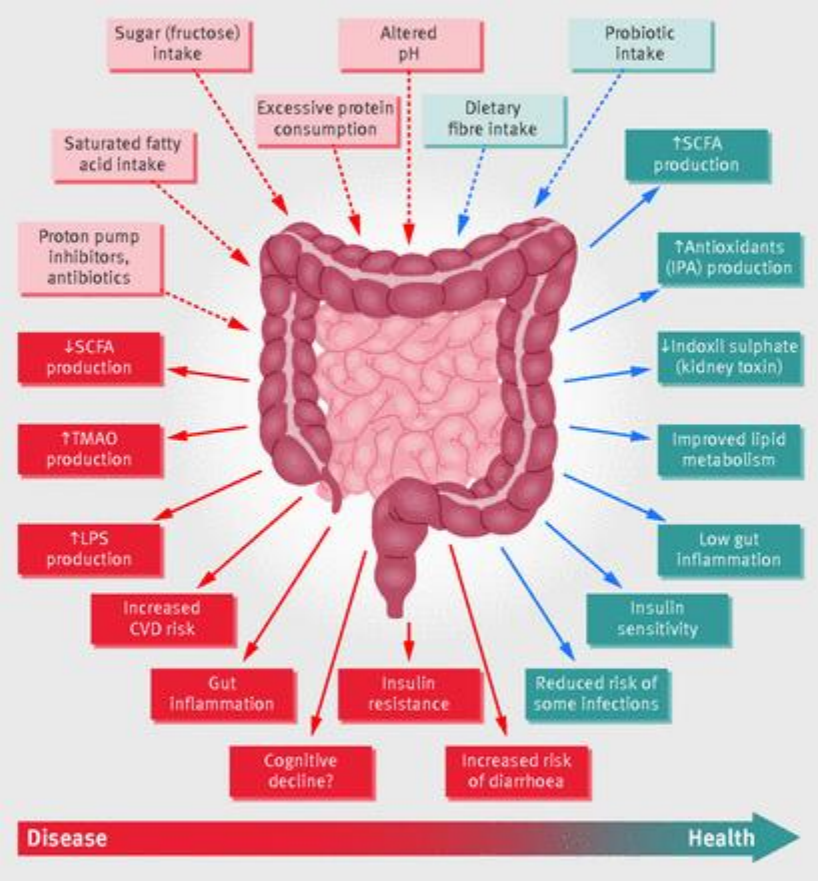


Figure 2.1 Dietary influence on gut health: How dietary factors like sugar and fiber affect gut health, with poor choices leading to disease risks and healthy choices promoting better gut function and overall health.

The gut microbiome is a highly personalized and complex ecosystem, with each individual's microbial community being uniquely shaped by factors such as diet, geography, ethnicity, and genetics. A healthy gut microbiome is typically marked by a high diversity of microbial species, a rich array of microbial genes, and a stable core group of microorganisms [29]. This microbiome plays essential roles in health, including nutrient extraction, metabolism, immune development, and protection against harmful pathogens.

2.2.1 The Gut Microbial Composition

The microbial composition varies throughout the gastrointestinal tract, with distinct differences between the small and large intestines. Factors such as age, diet, antibiotic use, and genetics significantly influence the microbiome's development and changes, particularly in early life [29].

The gut microbiome has many types of microorganisms; however, the most dominant bacteria is from the kingdom Phyla in the human gut. Within this phyla, notable genera include Bacteroides, Clostridium, Faecalibacterium, Lactobacillus, Bifidobacterium, and Eubacterium. Similarly, the most common archaea in the gut are methanogens, particularly Methanobrevibacter smithii, which are involved in the production of methane through the fermentation of substrates like hydrogen and carbon dioxide. Moreover, viruses that infect bacteria, known as bacteriophages, play a role in modulating bacterial populations and genetic exchange. Eukaryotic viruses can also be present and influence host cells and immune responses [28].

The human gut microbiome is a complex and dynamic community of microorganisms, including bacteria, archaea, fungi, and viruses, that reside in the gastrointestinal tract. Moreover, it is highly individualized, with each person harboring a unique microbial composition shaped by factors like diet, geography, ethnicity, and host genetics [30].

2.2.2 Gut Microbiome Dysbiosis

Gut microbiome dysbiosis refers to an imbalance in the complex community of microorganisms residing in the gastrointestinal tract. These microorganisms, including bacteria, viruses, fungi, and other microbes, play crucial roles in maintaining overall health by aiding in digestion, supporting the immune system, and contributing to the synthesis of essential nutrients [31].

Dysbiosis can manifest in several ways. It can involve a loss of beneficial bacteria, such as those from the genera Lactobacillus and Bifidobacterium, which are essential for maintaining gut health by aiding in digestion, producing vitamins, and protecting against pathogenic bacteria [32]. When these beneficial bacteria are significantly reduced, it can lead to impaired digestive functions, weakened immune responses, and an increased risk of infections. Additionally, dysbiosis can result in the overgrowth of harmful bacteria, like Clostridium difficile and certain strains of Escherichia coli, which can cause various gastrointestinal issues, including inflammation, infections, and conditions like irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD) [33].

Another key aspect of dysbiosis is the decreased diversity of gut bacteria. A healthy gut microbiome is characterized by a diverse array of microbial species, crucial for resilience against external stressors and maintaining overall gut health. When microbial diversity is reduced, it compromises the gut's ability to function effectively and can lead to a range of health problems

[34].

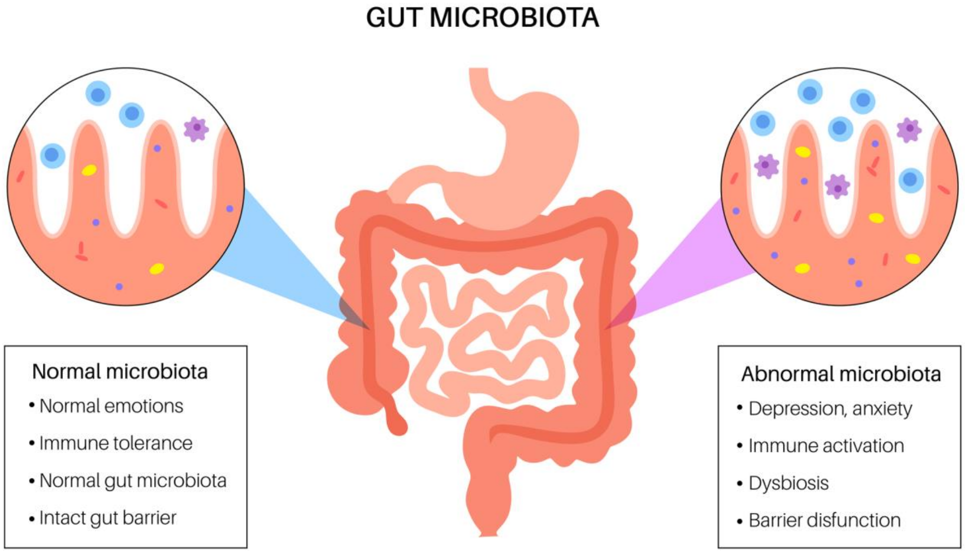


Figure 2.2 Probiotics, microbiome, and gut health; a comparison between normal and dysbiotic gut health

Dysbiosis can be caused by a variety of factors, each contributing to the imbalance in the gut microbiome. Infections, whether bacterial, viral, or fungal, can disrupt the normal microbial community by allowing harmful pathogens to overgrow and suppress beneficial microbes [19]. Inflammation in the gut, often a result of chronic conditions like Crohn's disease or ulcerative colitis, can create an unfavorable environment for beneficial bacteria, further promoting the growth of harmful species [35]. Genetic predispositions can also play a role, as certain genetic variations can affect the immune system's ability to regulate and maintain a balanced microbiome.

Diet and antibiotic use play critical roles in shaping the gut microbiome, which in turn can impact overall health. Diets high in processed foods, sugars, and unhealthy fats can reduce microbial diversity and promote harmful bacteria, whereas a diet rich in fiber, fruits, and vegetables support a healthy microbiome. Antibiotic use can further disrupt this balance by killing both harmful and beneficial bacteria, leading to reduced microbial diversity and the growth of resistant strains[36].

Additionally, disruptions in the gut-brain axis have been linked to various neurological and mental health conditions. For instance, altered gut microbiome profiles in individuals with autism spectrum disorders (ASD) may affect brain function through imbalances in key metabolites and neuroactive molecules produced by gut bacteria[37]. Similarly, disturbances in the gut-brain axis are associated with depression, anxiety, and other mental health issues, underscoring its importance in regulating mood and emotional well-being[38].

In the case of depression, research has shown that individuals with depressive disorders often have an altered gut microbiome composition. This imbalance can affect the production of neurotransmitters such as serotonin, which is primarily produced in the gut and is essential for regulating mood [39]. A disrupted gut microbiome can lead to lower levels of serotonin, contributing to the development of depressive symptoms.

Anxiety is another mental health condition closely linked to the gut-brain axis. The gut microbiome can influence the production of stress-related hormones and the functioning of the hypothalamic-pituitary-adrenal (HPA) axis, which controls the body's response to stress [40]. An imbalanced gut microbiome can result in heightened stress sensitivity and exaggerated stress response, common features of anxiety disorders [41].

Maintaining a healthy gut microbiome is crucial for overall well-being. Factors that can help prevent or manage dysbiosis include a balanced diet rich in fiber, limiting antibiotic use, and considering probiotic supplements when appropriate [42].

2.3 Gut-Brain Axis

The gut-brain axis describes the bidirectional communication system between the gastrointestinal tract and the central nervous system. This intricate interaction operates through several pathways. One of the main routes involves neural connections, particularly the vagus nerve and the enteric nervous system, which enable the gut to transmit signals directly to the brain and receive messages from it [43]. Beyond these neural pathways, the axis also includes hormonal and immune signaling mechanisms. Chemicals produced in the gut, such as neurotransmitters and cytokines, can enter the bloodstream and influence brain function [44].

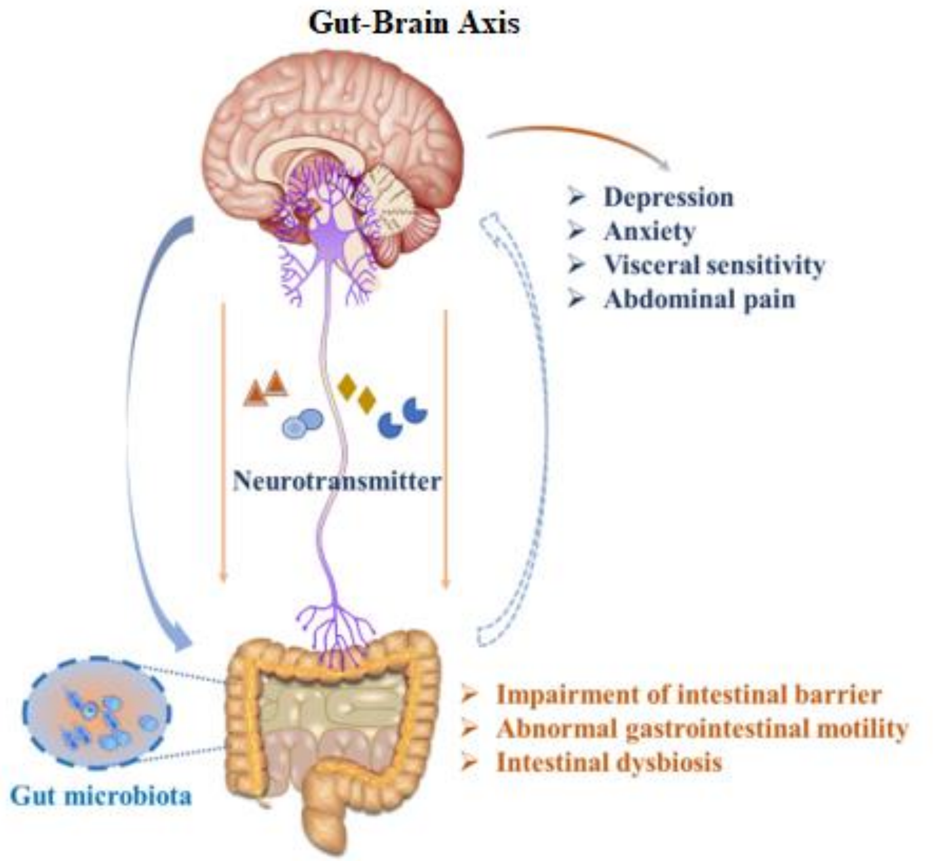


Figure 2.3 The bidirectional relationship between the brain and gut shows how neurotransmitters influence gut health and how gut microbiota impact brain function, contributing to conditions like depression, anxiety, and gastrointestinal issues.

The gut-brain axis (GBA) is a fascinating and intricate communication system that connects the central nervous system (CNS) with the enteric nervous system (ENS) via biochemical signaling pathways. This bi-directional connection plays a crucial role in maintaining overall health and has garnered significant attention in both clinical and scientific research. Clinical and experimental evidence has highlighted the gut microbiota as a pivotal regulator of the GBA[45].

The diverse community of microorganisms residing in the gut interacts with the host in multifaceted ways. Gut microbiota interacts locally with intestinal cells and the enteric nervous system (ENS). These interactions influence gut motility, secretion, and barrier function, which can impact gastrointestinal health. Furthermore, emerging research suggests that gut microbes have the capacity to influence the central nervous system (CNS) through various mechanisms. This includes the production of signaling molecules, such as neurotransmitters and neuroactive metabolites, which can enter the bloodstream and affect brain function [46].

The gut microbiome, composed of trillions of microorganisms residing in the digestive tract, plays a vital role in the gut-brain communication network. These microbes produce neuroactive molecules that can impact brain activity. For instance, some bacteria in the gut produce serotonin, a neurotransmitter that influences mood and behavior [47].

The gut-brain axis is essential for maintaining homeostasis and regulating various functions, including digestion, mood, cognition, and the stress response. Neural connections, such as the vagus nerve, allow the brain to influence gut motility, secretion, and blood flow, which ensures efficient digestion and nutrient absorption [48]. Conversely, signals from the gut can affect brain activity, influencing feelings of hunger and fullness.

Cognition is also influenced by the gut-brain axis. Gut bacteria produce neuroactive compounds, such as short-chain fatty acids and neurotransmitters, which can affect brain function, impacting learning, memory, and decision-making processes [49]. Additionally, the gut-brain axis plays a role in the body’s stress response. The gut microbiome can modulate the hypothalamic-pituitary-adrenal (HPA) axis, which controls the release of stress hormones like cortisol [50]. Dysregulation of this axis can lead to increased stress sensitivity and related health issues.

Overall, the gut-brain axis integrates and regulates multiple physiological processes crucial for maintaining overall health and well-being.

2.4 Diagnostics for the Gut-Brain Axis Problems

The gut microbiome has emerged as a pivotal factor in the diagnosis and treatment of neurological disorders, with growing evidence highlighting its intricate interactions with the central nervous system (CNS) via the gut-brain axis (GBA). This bidirectional communication pathway suggests that the gut microbiota not only plays a significant role in maintaining neurological health but also holds potential as a diagnostic and therapeutic target[51].

Alterations in the microbial composition of the gut can influence neuroinflammation and other pathophysiological processes within the CNS, thereby contributing to the onset and progression of these diseases[52]. The impact of dysbiosis on neuroinflammatory pathways underscores the critical role of the gut microbiota in the pathogenesis of neurological disorders.

Emerging research has explored the potential of gut microbiota profiles as biomarkers for neurological diseases. Changes in the composition of gut microbiota and the presence of specific microbial metabolites have been correlated with the severity and progression of neurological disorders[53]. However, this field is still in its early stages, and the development of standardized diagnostic methods is essential to validate these findings across diverse clinical settings.

Therapeutic strategies targeting the gut microbiome offer promising avenues for the treatment of neurological disorders. Interventions such as probiotics, prebiotics, synbiotics, and fecal microbiota transplantation (FMT) aim to restore a healthy microbial balance within the gut[54]. These approaches have shown potential in not only gastrointestinal diseases but also in the treatment of neurological conditions, including autism spectrum disorder. FMT, in particular, has demonstrated promising results in clinical applications beyond its traditional use, suggesting its potential utility in managing neurological disorders.

In addition to these therapies, Machine learning (ML) techniques have shown promise in leveraging gut microbiome data to aid in the diagnosis and screening of various neurological and mental health disorders associated with gut-brain axis dysfunction. Multiple studies have demonstrated that microbiome-based ML models can accurately classify Parkinson's disease patients, although these models tend to be study-specific and often lack generalizability across different datasets[55]. Meta-analyses of these studies have helped identify Parkinson's-associated microbial pathways that may contribute to disease pathogenesis through the gut-brain axis [56].

Moreover, another recent study developed a machine learning (ML) model using gut microbiome composition that could effectively identify patients with Myasthenia Gravis (MG). This finding suggests the potential for a non-invasive diagnostic screening tool for this neuromuscular disorder, highlighting the promise of gut microbiome-based ML models in medical diagnostics [57].

Furthermore, alterations in the gut microbiome have been linked to the development of autism, and machine learning (ML) models may be able to leverage these microbial signatures for improved diagnosis. By identifying specific patterns and changes in the gut microbiome associated with autism, ML models can potentially enhance the accuracy and early detection of this condition [44].

Machine learning (ML)-based approaches offer several key advantages for addressing gut-brain axis diseases. Firstly, they have the ability to identify complex, multivariate microbial patterns that can distinguish disease states from healthy controls, providing a more nuanced understanding of the microbiome's role in these conditions[58]. Secondly, ML-based methods present the potential for non-invasive and cost-effective screening and diagnosis, making them more accessible compared to current techniques. Lastly, these approaches offer the opportunity to uncover gut microbial biomarkers and provide mechanistic insights into how the gut-brain axis is disrupted in various diseases, potentially leading to more targeted and effective treatments[59].

A study has demonstrated the potential applications in distinguishing between different neurological diseases using machine learning techniques, which addressed the challenging task of multi-class classification, where the goal is to differentiate between multiple neurological diseases simultaneously [60]. This is particularly valuable because it provides a comprehensive approach for distinguishing among diseases like multiple sclerosis, stroke, and others, which can have overlapping clinical presentations. Furthermore, the study wisely considered feature selection techniques to determine which taxonomic levels of microbial abundance data are most informative for classification. Finding that lower taxonomy levels (e.g., genus) yielded better performance suggests that focusing on finer-grained microbial features provides more discriminatory power. Furthermore, evaluating multiple classification algorithms helps identify which ML method is most suitable for this specific task. As a result, the LogitBoost-based prediction model outperformed other classifiers, suggesting that boosting techniques might be particularly effective for distinguishing between these neurological diseases based on gut microbiome data. This study's findings have significant clinical implications. The identified feature subsets can potentially be used to diagnose and differentiate neurological diseases [60]. This could lead to earlier and more accurate diagnoses, which is critical for patient outcomes and the development of targeted treatment strategies.

2.5 Advantages of Data-Guided Decisions

Data-guided predictive modeling, particularly in the context of the gut microbiome, provides a robust framework for advancing our understanding and management of neurological disorders. This approach leverages advanced computational techniques to analyze complex datasets derived from microbiome studies, yielding insights that can significantly enhance both diagnosis and treatment strategies[61].

Predictive modeling excels in integrating data from various sources, including microbial composition, clinical symptoms, and genetic information. This comprehensive approach enables a more holistic understanding of the influence of gut microbiota on neurological disorders, such as Alzheimer's and Parkinson's disease, and their underlying pathophysiologies[62]. By combining diverse datasets, predictive models can uncover intricate relationships between gut microbiota and disease processes, which might not be apparent through traditional analytical methods.

Through the analysis of the gut microbiome and its metabolites, predictive models can identify specific biomarkers associated with neurological disorders. These biomarkers are crucial for early diagnosis, monitoring disease progression, and guiding therapeutic interventions. The ability to pinpoint biomarkers with high specificity and sensitivity can facilitate timely and targeted treatments, potentially improving patient outcomes by addressing disease mechanisms at earlier stages[63].

Moreover, predictive modeling also plays a vital role in the development of personalized treatment strategies based on individual microbiome profiles. This personalized approach allows for the customization of interventions, such as probiotics or dietary modifications, tailored to the specific microbial dysbiosis present in a patient[64]. By aligning treatment strategies with the unique microbiome composition of each individual, predictive modeling can enhance treatment efficacy and minimize adverse effects.

Furthermore, data-driven predictive models are instrumental in elucidating the mechanisms by which gut microbiota influences brain health. By analyzing patterns within complex datasets, these models can identify novel therapeutic targets and contribute to the development of interventions aimed at restoring microbial balance[65]. Understanding these mechanisms is critical for devising strategies that mitigate neurological symptoms and halt disease progression.

Another advantage of data-guided predictive modeling is its capacity for continuous monitoring and adaptation of treatment plans. By regularly assessing a patient's microbiome and its relationship to neurological health, clinicians can make real-time adjustments to therapeutic strategies[66]. This dynamic approach ensures that patient care is continuously optimized based on the latest data, improving long-term outcomes.

2.6 Machine Learning Modeling vs. Neurological Diseases: The Challenges

Diagnosing gut-brain axis diseases using machine learning (ML) modeling presents several key challenges. The multifactorial nature of functional gastrointestinal disorders complicates the diagnostic process. ML algorithms need to integrate diverse types of data—including microbiome signatures, demographics, immunology, and neuroimaging—to effectively identify patterns and provide accurate diagnoses. This complexity requires sophisticated models capable of handling and analyzing varied and intricate datasets to capture the full spectrum of factors involved in these disorders [67].

Diagnosing gut-brain axis diseases using machine learning (ML) modeling presents several key challenges. The multifactorial nature of functional gastrointestinal disorders (FGIDs), such as irritable bowel syndrome, complicates the diagnostic process. ML algorithms need to integrate diverse types of data—including microbiome signatures, demographics, immunology, and neuroimaging—to effectively identify patterns and provide accurate diagnoses [67]. This complexity requires sophisticated models capable of handling and analyzing varied and intricate datasets to capture the full spectrum of factors involved in these disorders.

Choosing appropriate preprocessing steps like normalization and filtering is critical but not always straightforward. In fact, typical preprocessing does not always improve the predictive performance of ML models for diseases like colorectal cancer. Feature selection is crucial to reduce classification error, and multivariate methods like Statistically Equivalent Signatures are more effective than univariate approaches [68]. Validating ML models on separate test datasets is essential to obtain accurate performance estimates. For instance, one study found that random forest modeling combined with the Statistically Equivalent Signatures algorithm provided the best accuracy.

Interpreting ML model results to gain biological insights is another significant challenge. Techniques such as logistic regression with Individual Conditional Expectation plots can help yield interpretable results, making it easier to understand the biological implications of the data and the underlying mechanisms of the diseases [69].

It is essential to translate findings from animal studies on gut-brain axis diseases, such as autism spectrum disorder, to humans to identify potential therapeutic targets. Incorporating established assays for gut dysfunction into the behavioral testing of animal models is crucial for this process [70].

A recent study represents a significant contribution to the growing body of research examining the relationship between gut microbiota and Alzheimer's disease (AD), where the study confirms the previous findings that Alzheimer's disease (AD) patients exhibit alterations in their gut microbiota composition. This aligns with emerging evidence suggesting a link between the gut microbiome and neurological diseases. Moreover, the use of machine learning models in this study is noteworthy. These models are valuable for analyzing complex and multidimensional datasets, such as microbiome data, and can identify patterns and associations that might be challenging to detect through traditional statistical approaches. Specifically, the study investigated associations between gut microbiota composition and AD biomarkers, specifically amyloid and p-tau. The finding that machine learning models could predict amyloid and p-tau status from microbiota composition with AUCs of 0.64 and 0.63, respectively, is promising. It suggests that specific microbial signatures or patterns may be linked to these key AD pathological markers [71].

While the study's findings are promising, it's important to acknowledge that more research is needed to validate these associations and explore the underlying mechanisms. Additionally, the modest AUC values indicate room for improvement in predictive accuracy. Furthermore, these findings support the notion that the gut microbiome may play a role in AD pathogenesis or progression. Further research into the specific microbial taxa or metabolites associated with AD biomarkers could provide deeper insights into disease mechanisms. If specific microbial signatures are confirmed to be linked to AD biomarkers, they could become targets for interventions aimed at modifying the gut microbiome to mitigate AD risk or slow disease progression.

A growing number of cross-sectional studies have investigated the microbiota composition in individuals with a specific neurological disorder versus healthy age-matched individuals. However, these studies provide just a snapshot in time, and longitudinal cohort studies are needed. Experimental models have been essential for moving research in the human microbiota gut-brain axis forward. Experimental models have been essential for moving research in the human microbiota gut-brain axis forward. In tandem, a large experimental effort has been directed towards attempting to dissect the various ways machine learning can be used to study the association of gut microbiome with neurological and neurodegenerative diseases [72].

2.6.1 The Target Overlap Dilemma

Various studies have shown significant overlap in the mechanisms and pathways underlying different neurological diseases, posing challenges for building effective machine learning (ML) models to distinguish between these neurological conditions and normal states. A review article emphasizes that many adult-onset diseases, such as Alzheimer's disease (AD), Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS), and frontotemporal dementia (FTD), share common features like protein aggregation, neuroinflammation, and immune dysregulation [73]. This extensive overlap complicates the development of targeted therapies and suggests that a combination of treatments addressing different disease pathways may be necessary.

Another analysis revealed that 7 out of the top 10 enriched diseases associated with upregulated hub genes were neurological diseases/disorders, highlighting the significant overlap in the underlying biology among different neurological diseases (NDDs). This overlap in disease targets and mechanisms complicates the task of building machine learning (ML) models that can accurately differentiate between NDDs and normal conditions [74].

The discussion notes that a major limitation in developing effective AI/ML models for brain diseases is the limited sample size. This issue is beginning to be addressed by the availability of public datasets like the Alzheimer's disease Neuroimaging Initiative (ADNI) and the Human Connectome Project. Despite this progress, the heterogeneity and overlap between different brain diseases continue to present significant challenges [75].

The extensive overlap in the underlying mechanisms, pathways, and targets between different diseases poses a significant challenge when building accurate and generalizable machine learning (ML) models to distinguish neurological diseases from normal conditions. Addressing this overlap through integrated, multi-target approaches may be necessary to develop effective diagnostic and predictive tools for neurological diseases.

2.7 Research Gap and Problem Statement

An extensive overlap in mechanisms between neurological diseases and normal conditions makes it challenging to build accurate ML models to distinguish them. Many adult-onset NDDs share common pathways of protein aggregation, neuroinflammation, and immune dysregulation, posing difficulties in developing targeted therapies and diagnostic tools.

Limited sample size and availability of high-quality, annotated datasets for many neurological disorders have been a challenge in building robust ML models. While public datasets like ADNI and the Human Connectome Project have started to address this, the inherent complexity and overlap between brain diseases persist [76].

ML models for the gut microbiome are susceptible to biases and variations due to differences in library preparation, sequencing strategy, and choice of reference databases used for taxonomic annotation in bioinformatics pipelines. Reprocessing raw sequences using a single pipeline can help harmonize bacterial profiles [77].

While AI-guided gut microbiome profiling holds promise for understanding and diagnosing neurological diseases, significant challenges remain in building accurate, generalizable, and clinically useful ML models. Addressing the overlap between neurological diseases, expanding high-quality datasets, improving model interpretability, and standardizing microbiome profiling workflows are key research gaps to address.

This study aims to leverage machine learning (ML) modeling techniques to identify potential targets within the gut microbiome associated with these diseases, differentiating between neurological and normal conditions.

2.8 Objectives

- To collect and preprocess the microbiome data for the gut microbiome profiling associated with the neurological disorder.
- To develop predictive ML models based on gut microbiome profiles of the neurological disorder.
- To identify specific biomarkers associated with the disease state, potentially serving as therapeutic targets or diagnostic tool.

3 CHAPTER 3: Materials and Methods

In this study, we performed a machine learning-based analysis on metagenomic samples from patients diagnosed with major depressive disorder. The analysis involved several critical steps to derive meaningful insights from the data.

First, we retrieved and preprocessed the metagenomic data to ensure its quality and integrity. Following this, we carried out metagenomic assembly to reconstruct the genetic material present in the samples. Subsequently, we conducted Prokaryotic Gene Recognition and Translation Initiation Site Identification to obtain genetic and protein data from the metagenomic samples. Feature extraction was then performed to simplify and distill key information from the complex metagenomic data. Finally, we developed machine learning models to classify normal and diseased states using the metagenomic samples.

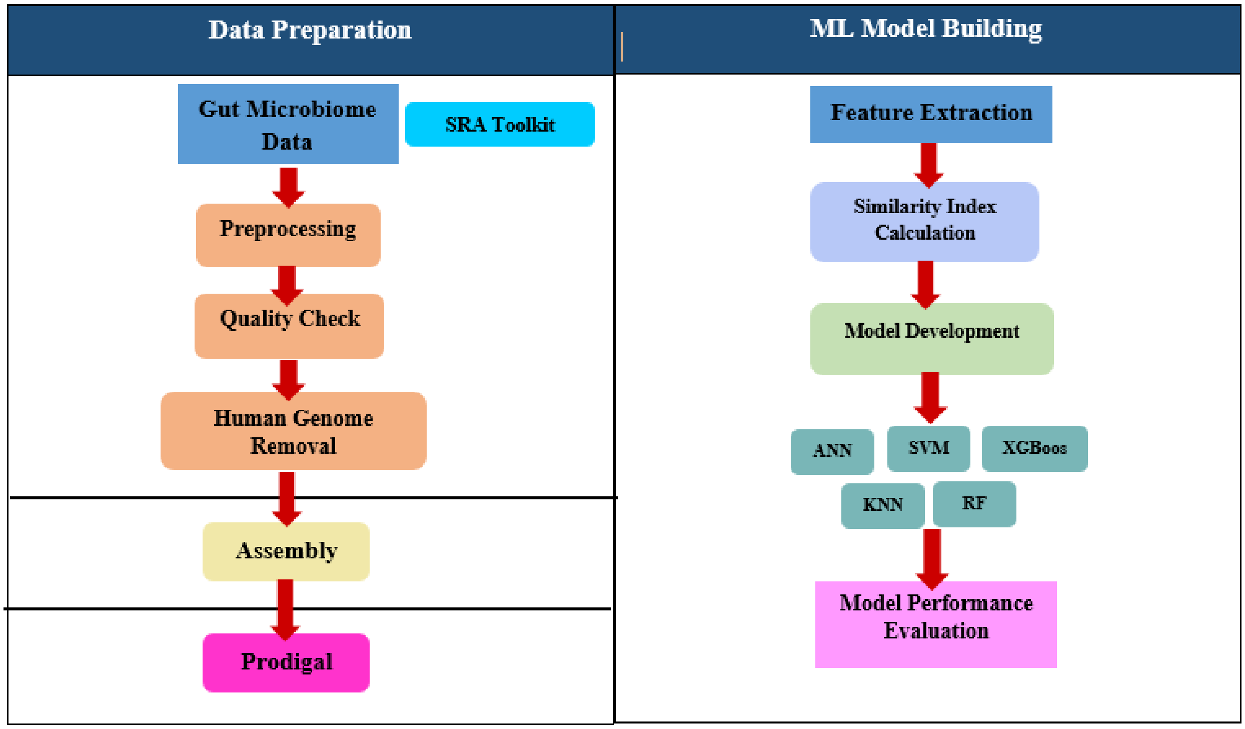


Figure 3.1 Integrated workflow: Gut microbiome data preprocessing and assembly coupled with machine learning model development and performance evaluation

This comprehensive methodology shown in **Figure 3.1** enabled us to achieve a deeper understanding of the microbial community and the functional potential encoded within the metagenomic data from patients with neurological disorders.

3.1 Data Retrieval

The gut microbial data of normal vs. patients with major depressive disorder was retrieved through the NCBI SRA database[78]. A total of 74 paired samples, out of which 36 patients, gut microbiome data was retrieved, and 39 control samples were obtained.

The metagenomic data discussed above is available in the Sequence Read Archive (SRA) under the BioProject accession number PRJNA762199[79]. To retrieve this data, we used the SRAToolkit (v3.0.0) (https://github.com/ncbi/sra-tools), specifically employing the "prefetch" command. After obtaining the data, we utilized the "fastq-dump" command to convert the paired-end SRA samples into the FASTQ format for compatibility. Finally, we compressed these FASTQ paired-end files using the gzip command.

3.2 Data Preprocessing

Data preprocessing is an essential phase in the data analysis pipeline, ensuring the quality and integrity of the data. This step is vital for achieving accurate and meaningful analysis, ultimately resulting in more robust and reliable outcomes.

3.2.1 Quality Check

To begin, the quality of the raw sequencing data was evaluated using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), a tool designed for visual quality control[80]. We executed the “fastqc” command to produce quality reports for the paired-end files. Upon reviewing these reports, it was evident that the data needed preprocessing to eliminate artifacts such as adapters, duplicated reads, and very short reads. To address these issues, we used the fastp tool, applying specific parameters: “-D” to remove duplicated reads, “--detect_adapter_for_pe” to remove adapters, and “--length_required 150” to set a minimum read length of 100 bases. These preprocessing steps were essential to enhance the quality of the sequencing reads and prepare them for further analysis.

3.2.2 Host Reads Removal

Metagenomic samples are often contaminated with host reads, such as human DNA in gut metagenomes, which can lead to false-positive results. To prevent this, it is crucial to remove these host reads before conducting analyses. This was accomplished using the BBDuk script from the BBMap suite[81]. BBDuk used the human genome (GRCh38) as a reference to identify and eliminate host-derived reads from the metagenomic data. After this preprocessing step, the quality of the cleaned data was reassessed with FastQC to confirm its suitability for further analysis

3.3 Metagenome Assembly

De novo metagenome assembly involves reconstructing contiguous sequences (contigs) from short reads generated by next-generation sequencing. For this task, metaSPAdes - v3.15.3 was employed via the KBase online server (https://www.kbase.us). The assembly was performed using the default parameters provided by metaSPAdes[82].

3.4 Quality Assessment (QUAST)

The Quality Assessment Tool for Genome Assemblies (QUAST) was used to evaluate the quality of the metagenome assembly. The "quast.py" script from QUAST (v5.2.0) was utilized to generate detailed quality assessment reports[83]. These reports provided important statistics for each sample, including the distribution of contigs, the size of the largest contig, N50 values, and the total assembly length, among other metrics.

3.5 Prokaryotic Gene Recognition and Translation Initiation Site Identification (PRODIGAL)

Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm) for prokaryotic gene recognition and translation initiation site (TIS) identification in my microbial genome samples was used [84].

This tool aided in microbial genome annotation and metagenomics analysis and was included in accurately identifying genes, proteins, and nucleotides in prokaryotic genomes and microbiome samples.

3.6 Feature Extraction

Metagenomic samples contain DNA from multiple microbial species, complicating their direct use in machine learning (ML) models. To address this, we performed feature extraction using Python with the protpy library, supported by biopython, numpy, and pandas, for data normalization and dataframe creation [85].

We loaded preprocessed data into two input files: control (normal state) and case (diseased state), with the control group serving as the reference. Feature extraction reduced the data's dimensionality, making it suitable for robust ML model training.

Feature selection was then carried out to identify significant features and their associations with metadata classes. This step enhanced the accuracy and interpretability of the ML models by focusing on the most relevant features, thereby providing valuable insights into the biological processes underlying the conditions studied.

3.7 Feature Preprocessing

The initial phase of the data analysis involved rigorous data cleaning to identify and correct errors within the dataset. These steps were essential to ensure the integrity and reliability of the dataset before further analysis.

Firstly, a technique was employed to transform the dataset into a suitable format for analysis, which was normalization. Normalization involved scaling the data to a uniform range, typically between 0 and 1. This process ensured that all features contributed equally to the analysis, preventing any single feature from dominating due to its scale.

Ensuring high-quality data through these preprocessing steps was crucial for achieving better model accuracy and reducing the risks of overfitting and underfitting. Proper data preprocessing also significantly decreased the computational resources required for training models, thereby making the analytical process more efficient [86].

3.8 Decision Tree and Random Forest

Feature selection is a critical step in the modeling process, and utilizing decision trees for this purpose is particularly effective due to their inherent ability to evaluate the importance of features during model construction. To identify the most important features, we employed decision trees that recursively split the dataset based on feature values. At each node, the algorithm selected the feature that provided the best split according to a chosen criterion, which in this study was information gain. Features that appeared higher in the tree were considered more important, as they significantly contributed to reducing uncertainty in the target variable. Information gain equation used for choosing important features is given below:

\[\text{IG}(T, X) = H(T) - \sum_{v \in \text{Values}(X)} \frac{|T_v|}{|T|} H(T_v) \]

Where:

- \(T \) is the set of all instances.

- \(X \) is the feature on which the split is being made.

- \(H(T) \) is the entropy of the original set \(T \).

- \(\text{Values}(X) \) are the possible values of feature \(X \).

- \(T_v \) is the subset of \(T \) for which feature \(X \) has value \(v \).

- \(H(T_v) \) is the entropy of the subset \(T_v \).

- \(\frac{|T_v|}{|T|} \) is the proportion of instances in subset \(T_v \) relative to the original set \(T \).

Also, the formula for entropy calculation is:

The entropy \(H(T) \) for a set \(T \) with binary classification is calculated as:

\[H(T) = - p_+ \log_2(p_+) - p_- \log_2(p_-) \]

Where:

- \(p_+ \) is the proportion of positive instances in \(T \).

- \(p_- \) is the proportion of negative instances in \(T \).

Information Gain is the reduction in entropy or impurity after the dataset is split on a feature. A higher information gain indicates a more significant reduction in uncertainty, making the feature a good choice for splitting.

In addition to an individual decision tree, we incorporated Random Forests to enhance feature selection. Random Forests, which aggregate multiple decision trees, naturally rank features based on their importance during the tree-building process. This approach allowed for the selection of the most significant features without requiring a separate feature selection step.

Random Forest (RF) is primarily known as an ensemble method to enhance predictive performance, but it also shows promise for feature selection, particularly in high-dimensional datasets. Studies have demonstrated that RF can effectively identify key features due to its ability to handle large data sets and resist overfitting. Additionally, RF can be integrated into feature selection frameworks, especially when combined with Wrapper techniques, improving both model accuracy and feature interpretability[87].

3.8.1 Recursive Feature Elimination

Once feature importance scores were obtained from the Random Forest model, Recursive Feature Elimination (RFE) was applied to further refine feature selection. RFE involved recursively removing the least important features and building models on the remaining subset of features. This iterative process continued until the optimal set of features that maximized the model's performance was identified.

By employing decision trees and Random Forests for feature selection, we effectively identified and ranked the most important features in the dataset. The integration of Recursive Feature Elimination further refined this selection, ensuring that the final set of features used in the model provided the best possible performance. This methodology facilitated a streamlined and efficient feature selection process, ultimately contributing to the robustness and accuracy of the model.

3.8.2 Cost Complexity Pruning

After constructing a decision tree, pruning techniques were applied to remove branches with minimal significance. This process not only simplified the model but also helped identify the most relevant features that contributed to the final predictions

Cost Complexity Pruning (CCP) was utilized to prevent overfitting by optimizing the size of the decision tree. The objective of CCP is to find the subtree that minimizes the sum of the impurity of its leaves and a penalty term proportional to the number of leaves. The following function was used:

\[R_\alpha(T) = R(T) + \alpha \cdot |T| \]

Where:

- \(R(T) \) represents the impurity (e.g., Gini index) of the entire tree.

- \(|T| \) denotes the number of leaves in the tree.

- \(\alpha \) is a hyperparameter that controls the tradeoff between accuracy and tree size.

3.9 Grey Area Removal

The Human Genome Project revealed that 99.9% of human genomes are identical across all individuals, while the Human Microbiome Project highlighted significant genetic differences within our microbiomes [88]. To enhance the distinction between case and control groups, we decided to remove the common amino acid regions that are conserved between both groups. By excluding these grey regions which was all the instances with 90% or less similarity, we aimed to create a dataset with a higher range of differences between case and control indexes, thus improving the analytical clarity and potential for discovering meaningful insights.

3.10 Similarity Index Calculation

To effectively remove the grey area between case and control samples, we employed the cosine similarity method to calculate the similarity index between protein abundance profiles. The process involved the following steps:

We calculated the similarity of each case to each control in batches of 25,000 to manage the computational load efficiently. The total number of comparisons made was 14,579,813, covering all possible pairings between case and control samples.

Using the cosine similarity formula, the similarity index for each pair was determined as follows:

\[

\text{similarity}(P, Q) = \cos(\theta) = \frac{P \cdot Q}{\|P\| \|Q\|}

\]

This similarity index provided a measure of how similar each case was to each control based on their protein abundance profiles. The resulting similarity indices were then used to construct a similarity matrix.

To focus on the most distinctive profiles, we converted the similarity matrix into a dissimilarity matrix by subtracting each entry from 1:

\[

\text{dissimilarity}(P, Q) = 1 - \text{similarity}(P, Q)

\]

From the total dissimilarity data, we identified and retained only the lowest similarity indices, specifically the top 10% of these indices, representing the least similar pairs. This approach allowed us to effectively remove the grey area, or the regions of high similarity, between the case and control samples. By excluding these regions, we enhanced the distinction between the two groups, providing a clearer dataset for subsequent analysis.

By calculating these indices in large batches, we efficiently handled the extensive dataset while maintaining the accuracy and reliability of the similarity and dissimilarity measures. This method ensured that the remaining data highlighted the most significant differences between case and control samples, facilitating more robust and insightful analysis.

3.11 Model Building

Machine learning techniques were employed for predictive modeling to accurately classify cases and controls. By leveraging different algorithms, the model was trained on the preprocessed dataset, which included features such as "moreaubroto_autocorrelation," "sequence_order_coupling_number," "conjoint_triad," "ctd_composition," "ctd_transition," and "ctd_distribution." These significant features enabled the identification of crucial patterns and relationships within the data, improving the accuracy and robustness of the predictions. This approach facilitated a data-driven understanding of the underlying patterns and relationships within the dataset, enhancing the overall effectiveness of the predictive model.

3.11.1 Data Splitting and Evaluation

Data manipulation was conducted to clearly define the classes of case and control, with case labeled as 1 and control as 0. The dataset was then randomly split into training and test sets, allocating 80% of the samples for training and 20% for testing. The training set was utilized to build the classification model, while the held-out test set was reserved for evaluating the model's performance.

To assess the performance of the model, several metrics were considered, including accuracy, precision, recall, F1 score, sensitivity, and specificity. These metrics provided a comprehensive evaluation of the model's predictive capabilities. The equations are given as below:

- **Precision:** Measures the accuracy of positive predictions.

Precision = TPTP+FP\text{Precision} = \frac{TP}{TP + FP}Precision=TP+FPTP

Where:

TPTPTP = True Positives

FPFPFP = False Positives

- **Sensitivity:** Measures the ability to find all positive samples.

  Recall = TPTP+FN\text{Recall} = \frac{TP}{TP + FN}Recall=TP+FNTP

  Where:

  TPTPTP = True Positives

  FNFNFN = False Negatives

- **F1 Score:** The harmonic mean of precision and recall, providing a single metric for model performance.

  F1 Score=2×Precision×RecallPrecision+Recall\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}F1 Score=2×Precision+RecallPrecision×Recall

- **Sensitivity:** Measures the ability to correctly identify negative samples.

  Specificity=TNTN+FP\text{Specificity} = \frac{TN}{TN + FP}Specificity=TN+FPTN

  Where:

  TNTNTN = True Negatives

  FPFPFP = False Positives

Furthermore, the model was interpreted to identify the most important predictive exposures, ensuring that the most relevant features were highlighted and understood. This approach facilitated a thorough understanding of the model's strengths and potential areas for improvement, contributing to the overall robustness and reliability of the classification model.

3.11.2 Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) was employed, consisting of four layers with the following structure:

1. Layer 1: 614 * 1024 neurons, with a ReLU activation function.
2. Layer 2: 1024 * 512 neurons, with a ReLU activation function.
3. Layer 3: 512 * 256 neurons, with a ReLU activation function.
4. Layer 4: 256 * 1 neuron, with a Sigmoid activation function.

The ANN's parameters are represented by the weights associated with each layer. These weights were optimized using the Adam optimizer, which leverages gradient descent to minimize the loss function. In this study, Binary Cross-Entropy (BCE) loss was utilized as the loss function, which is appropriate for binary classification tasks.

To enhance generalization and mitigate the risk of overfitting, a Dropout layer with a 20% probability was applied between the third and fourth layers. This technique helps prevent the model from becoming overly reliant on specific neurons during training, thereby improving its ability to generalize to unseen data.

3.11.3 Support Vector Machines (SVM)

Support Vector Machine (SVM) model was employed, with hyperparameters optimized using GridSearchCV. The optimization focused on parameters such as the regularization parameter

(`C`) where a value of 10 was selected for the regularization parameter, which controls the trade-off between maximizing the margin and minimizing classification errors. This value was found to provide the best balance for the given dataset.

In addition to that, a kernel coefficient was used where the `gamma` parameter was set to 'auto,' allowing the model to automatically determine the influence of individual data points based on the inverse of the number of features. This setting was chosen to adjust the decision boundary appropriately. Moreover, The polynomial kernel was tested up to the fifth degree, allowing the model to consider more complex, non-linear relationships in the data. This higher degree helps capture intricate patterns but also increases the risk of overfitting.

3.11.4 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) algorithm was utilized, with performance optimization conducted through GridSearchCV. The optimization focused on three primary hyperparameters, which included number of neighbors (`n_neighbors`), which determined that utilizing 3 neighbors provided the optimal balance between model complexity and accuracy. Also, the distance metric (`p`) was used as the Euclidean distance (`p=2`) was selected as the most effective metric for the classification task. Lastly, a uniform weighting scheme was chosen, ensuring that each neighbor had an equal influence on the classification decision.

3.11.5 Random Forest

A Random Forest classifier was employed with hyperparameters optimized using GridSearchCV to enhance the model's performance. The optimization process focused on several key parameters. The maximum depth of the trees (`max_depth`) was set to 20, limiting how deep the trees could grow. This measure helps control the model's complexity and reduces the risk of overfitting by preventing the trees from becoming too complex.

Additionally, the minimum number of samples per leaf (`min_samples_leaf`) was set to 2, ensuring that each leaf (or end node) contains at least 2 data points. This prevents the model from creating overly specific rules that may not generalize well to new data. The minimum number of samples required to split an internal node (`min_samples_split`) was set to 5, which helps control the model's growth by ensuring that splits are made only when there is sufficient data, thereby reducing the likelihood of overfitting. Finally, the number of trees in the forest (`n_estimators`) was set to 300. A larger number of trees generally improves model stability and accuracy by averaging the predictions of multiple trees.

3.11.6 eXtreme Gradient Boosting (XGBOOST)

An XGBoost classifier was utilized with specific hyperparameters tailored for optimal performance in a binary classification task. The classifier was configured with the settings including the `binary:logistic` objective function, which is suitable for binary classification, where the output is a probability that is mapped to either class using a logistic function.

Moreover, the learning rate was set to 0.2, controlling the step size at each iteration while moving towards a minimum of the loss function. This value was selected to balance convergence speed and the risk of overshooting the optimal solution. Also, the maximum depth of each tree was set to 7, limiting the number of nodes in the tree. This helps control the model complexity and prevent overfitting by avoiding overly deep trees.

Early stopping was also enabled with a patience of 15 rounds. Furthermore, the number of trees in the ensemble was set to 300, which helps in improving model performance by combining the predictions of multiple trees.

3.12 Differentially Abundant Proteins Retrieval

To identify differentially abundant proteins (DAPs) based on the lowest similarity index, we utilized cosine similarity to measure the dissimilarity between protein abundance profiles across different conditions. The methodology for this process is as follows:

Each protein was represented as a vector, where each element corresponded to the abundance value in a particular condition. For instance, if there were four conditions, the protein abundance vector would be represented as:

\[

P = [a_1, a_2, a_3, a_4]

\]

where \(a_i \) is the abundance of the protein in condition \(i \).

Then, we calculated the cosine similarity between all pairs of proteins using the formula:

\[

\text{similarity}(P, Q) = \cos(\theta) = \frac{P \cdot Q}{\|P\| \|Q\|}

\]

This resulted in a similarity matrix, where each entry represented the similarity between two proteins. The similarity matrix was then converted into a dissimilarity matrix by subtracting each entry from 1:

\[

\text{dissimilarity}(P, Q) = 1 - \text{similarity}(P, Q)

\]

For each protein, the average dissimilarity to all other proteins was calculated as follows:

\[

\text{avg_dissimilarity}(P) = \frac{1}{n} \sum_{i=1}^n \text{dissimilarity}(P, Q_i)

\]

where \(n \) is the total number of proteins and \(Q_i \) represents the other proteins. The proteins were then sorted in descending order based on their average dissimilarity values.

To enhance the robustness of the selection process, we kept only the top 10% of proteins based on similarity, specifically those within the 90th percentile. This approach ensured that the selected DAPs were those with the most distinct abundance profiles.

The top \(k \) proteins with the highest average dissimilarity were selected as the differentially abundant proteins. The rationale behind this approach is that proteins with the lowest similarity to most other proteins are likely to have unique abundance profiles and thus be differentially abundant across conditions. By using cosine similarity, the focus was on the shape of the abundance profiles rather than their magnitude, making the method more robust to noise and outliers.

To determine the specific value of \(k \), we selected an "elbow" in the sorted dissimilarity scores, where the scores begin to plateau, indicating a natural cutoff point for identifying the most differentially abundant proteins. This approach ensured that the selected DAPs were those with the most distinct abundance profiles, providing valuable insights into the variations across different conditions.

3.13 Model Building on Differentially Abundant Proteins (DAPs)

After identifying the differentially abundant proteins (DAPs), the next step involved choosing a suitable machine learning algorithm to classify samples based on these proteins. Given the dataset characteristics and the research objectives, we considered several algorithms including Artificial Neural Network (ANN) and support vector machines (SVM).

To ensure robust model development and validation, the dataset was split into training and validation sets. The training set was used to train the selected model, allowing the algorithm to learn the underlying patterns and relationships within the data. To prevent overfitting and to evaluate the model's generalizability, the model's performance was validated on the held-out validation set. This process involved tuning hyperparameters and assessing various performance metrics such as accuracy, precision, recall, F1 score, sensitivity, and specificity. By iterating through this training and validation cycle, we aimed to develop a model that not only accurately classified the samples but also maintained strong generalization capabilities, ensuring its reliability for future predictions.

3.13.1 Artificial Neural Networks

To classify samples based on differentially abundant proteins (DAPs) between case and control groups, we employed Artificial Neural Networks (ANN) due to their ability to model complex, non-linear relationships within the data. For this study, the ANN was designed to include multiple hidden layers, each with a sufficient number of neurons to capture the intricate dependencies between the features. The network was trained using the identified DAPs, which included features such as "moreaubroto_autocorrelation," "sequence_order_coupling_number," "conjoint_triad," "ctd_composition," "ctd_transition," and "ctd_distribution," with both case and control samples equally represented to ensure balanced learning.

During training, we utilized techniques such as dropout and regularization to prevent overfitting and enhance the model's generalizability. The ANN's performance was evaluated using metrics like accuracy, precision, recall, sensitivity, and specificity, which confirmed its effectiveness in distinguishing between case and control groups based on their protein abundance profiles. This approach provided a powerful and flexible tool for accurate classification, contributing to a deeper understanding of the biological differences captured by the DAPs. By leveraging these advanced ANN techniques, we achieved a robust model capable of delivering insightful predictions and enhancing our understanding of the underlying biological processes.

3.13.2 Support Vector Machine

To classify samples based on differentially abundant proteins (DAPs) between case and control groups, we employed Support Vector Machines (SVM) due to their effectiveness in high-dimensional spaces and their capability to handle non-linear classification problems. For this study, we utilized a non-linear SVM with a polynomial kernel, allowing the model to capture intricate patterns and relationships within the data. The features used for classification included "moreaubroto_autocorrelation," "sequence_order_coupling_number," "conjoint_triad," "ctd_composition," "ctd_transition," and "ctd_distribution," with both case and control samples equally represented to ensure balanced learning.

The SVM model was trained on the identified DAPs, and hyperparameters were carefully tuned to optimize performance. The polynomial kernel provided the flexibility needed to map the data into higher dimensions, enhancing the SVM's ability to find the optimal hyperplane that maximizes the margin between the case and control classes. The model's performance was evaluated using metrics such as accuracy, precision, recall, F1 score, sensitivity, and specificity, which demonstrated its effectiveness in distinguishing between case and control groups based on their protein abundance profiles. This approach provided a robust and precise tool for accurate classification, contributing to a deeper understanding of the biological differences captured by the DAPs. By leveraging the strengths of SVM, we achieved a highly reliable model capable of delivering insightful predictions and improving our comprehension of the underlying biological processes.

3.14 Biomarker Identification

From our earlier analysis using cosine similarity, we identified 28 differentially abundant proteins (DAPs). These DAPs were initially divided into two categories: classified and unclassified. The classified DAPs underwent further investigation, resulting in the selection of 11 DAPs for detailed coexpression analysis. This analysis was crucial in understanding the interactions and regulatory mechanisms these proteins share across different organisms. This workflow highlighted the systematic approach taken to narrow down the set of DAPs and perform deeper analysis, allowing for a better understanding of the protein interactions and their implications in the biological processes under investigation.

3.14.1 Coexpression Analysis

For the coexpression analysis, we used the STRING database [89], a powerful tool for studying protein-protein interactions and coexpression across multiple organisms. The input data consisted of 11 proteins, which had been classified from the original set of 28 differentially abundant proteins (DAPs) retrieved earlier using cosine similarity analysis. These 11 proteins were selected for further analysis due to their significant involvement in key biological processes and their availability of functional annotations.

The STRING database was employed to explore potential coexpression patterns of these proteins across various species. By inputting the protein identifiers into the database, we were able to retrieve coexpression data, which indicated conserved relationships between the proteins across diverse organisms.

3.14.2 Network Construction

To investigate the protein-protein interactions and coexpression patterns of the 11 classified differentially abundant proteins (DAPs), we utilized the STRING database. The input data consisted of these 11 proteins, which had been previously classified based on their known structures. By submitting the protein identifiers into the STRING database, we aimed to generate a comprehensive interaction network. STRING's algorithm was used to study the coexpression, direct physical interactions, and functional associations among the proteins based on experimental data, curated knowledge, and computational predictions.

4 CHAPTER 4: Results and Discussion

4.1 Data Preprocessing

This data elucidates the impact of preprocessing steps on read count and variability in metagenomic samples. For the control group, the total read count was 84.47 million reads before preprocessing. After the removal of low-quality reads, adapters, and duplicate reads, the total read count decreased to 56.12 million reads. The mean read count before preprocessing (raw) was 42.37 Mbp, which was reduced to 27.44 Mbp after preprocessing (clean). The standard deviation for the control group's read count increased from 4.6 Mbp before preprocessing to 6.3 Mbp after preprocessing.

In the case group, as shown in **Table 4.1**, the initial total read count was significantly higher, with 1.51 billion reads before preprocessing, which reduced to 990.32 million reads after preprocessing. The mean read count before preprocessing was 43.06 Mbp, decreasing to 29.12 Mbp post-preprocessing. The standard deviation for the case group's read count decreased from 4.5 Mbp before preprocessing to 3.9 Mbp after preprocessing.

Finally, in the Human Genome (HG) trimmed dataset, the read count remained unchanged. This indicates that the preprocessing steps were effective in removing contaminants and redundant data, thereby enhancing the quality and reliability of the metagenomic dataset for subsequent analyses.

Table 4.1 Comparison of data metrics for control and case samples before and after preprocessing

	Control		Case	
	Raw	Clean	Raw	Clean
Mean	42.37Mbp	27.44Mbp	43.06Mbp	29.12Mbp
Mode	34.34	6.23	33.88	21.28
Standard Deviation	4.6	6.3	4.5	3.9
Total Read Count	84.47 million	56.12 million	1.51 billion	990.32 million

4.2 Metagenome Assembly and Quality Assessment (QUAST)

Metagenomes were de novo assembled and their qualities were evaluated using QUAST. For this, comparisons were performed within control and case groups, and between these groups. Within both control and case groups, the mean number of contigs was relatively higher in cases in contrast with the controls (Control: 7,178,910; Case: 7,400,903). Interestingly, the mean largest contig size and the mean N_{50} value, indicators of assembly contiguity, showed relatively consistent values across all groups, suggesting similar levels of contiguity regardless of sample status.

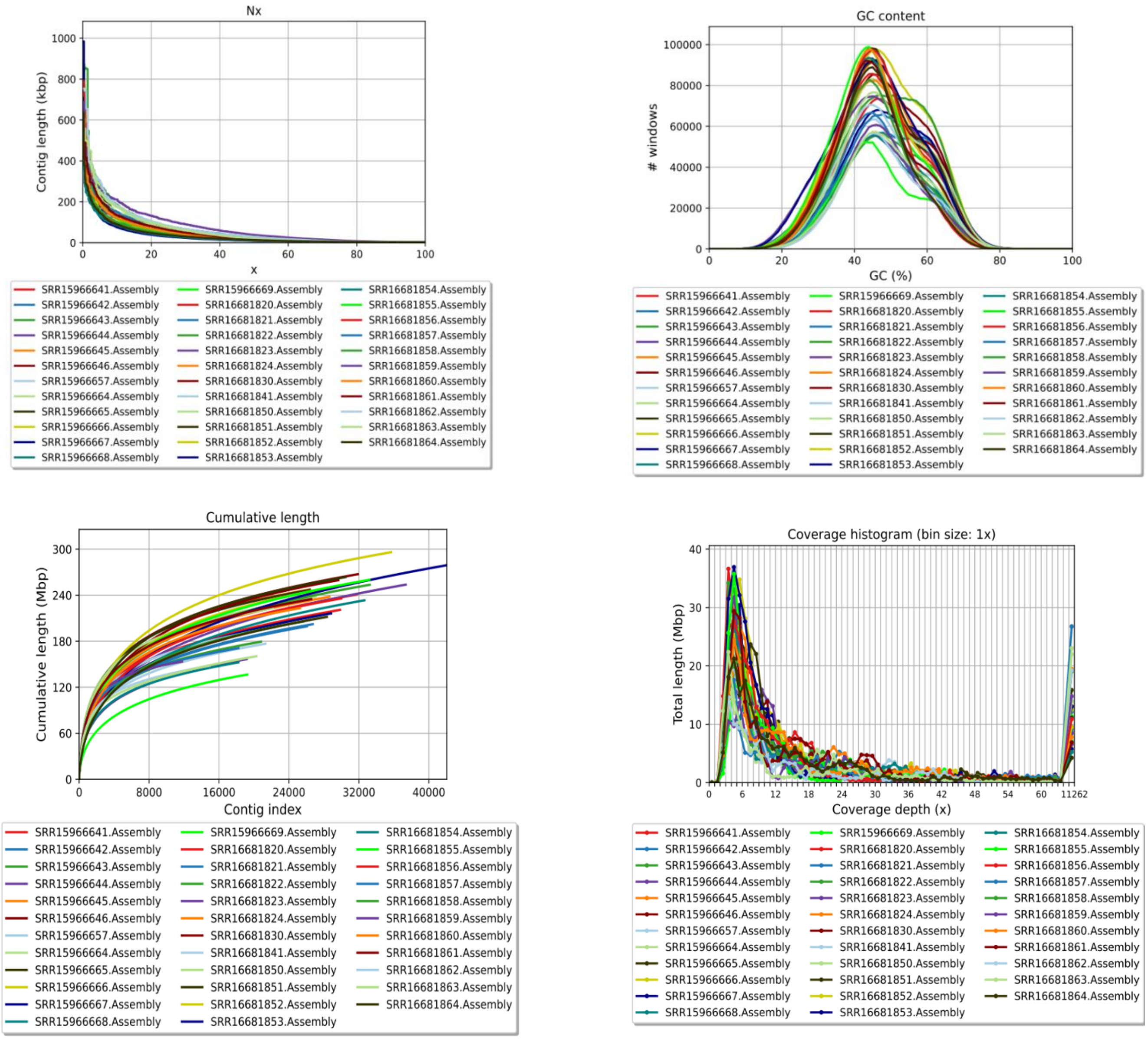


Figure 4.1 (A) Nx plot showing the distribution of contig lengths across various control samples. **(B)** Cumulative GC content of all control samples **(C)** Cumulative length illustrating the cumulative contig length distribution for each sample. **(D)** Coverage histogram depicting the total length of contigs at different coverage depths for each sample

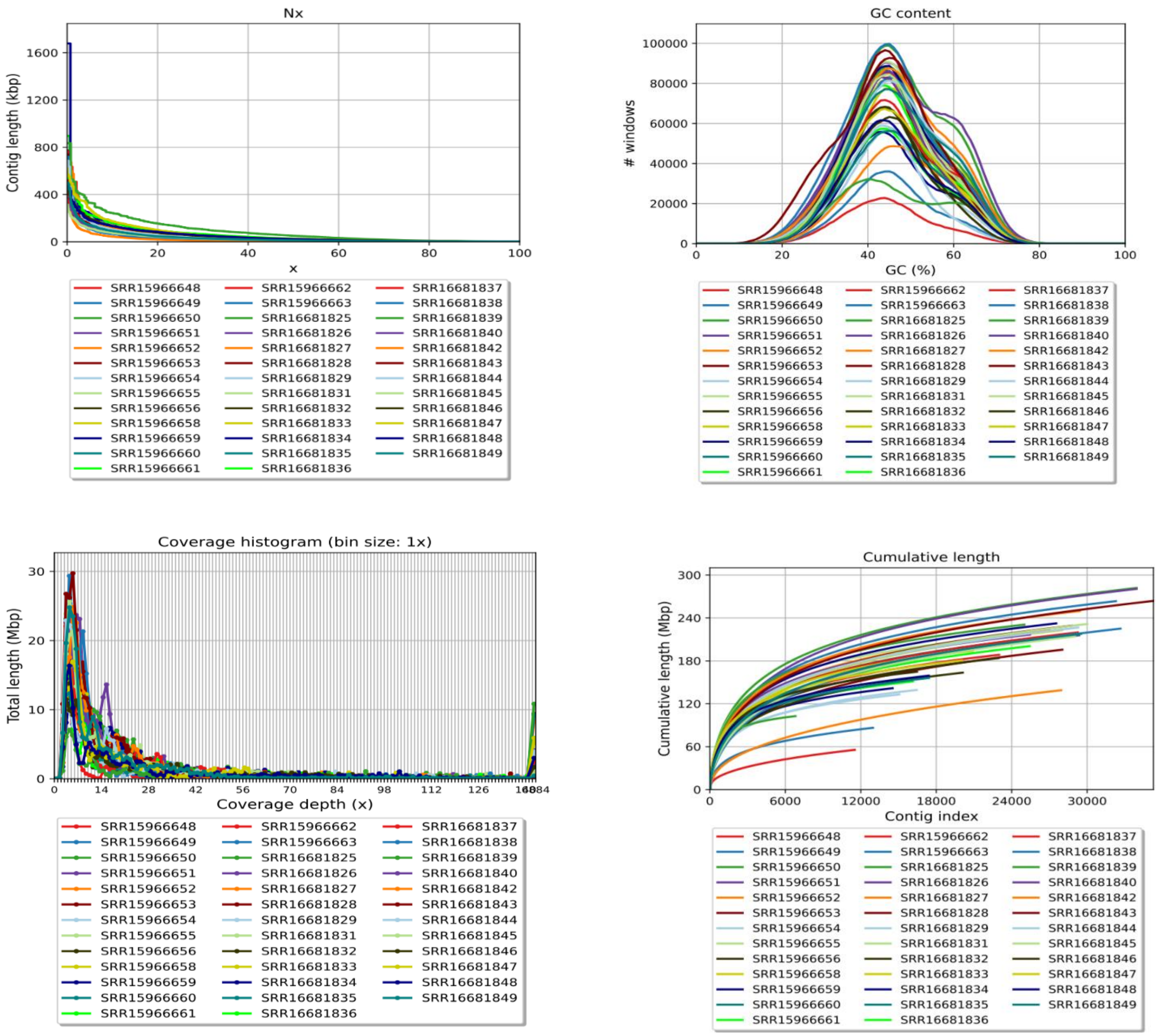


Figure 4.2 (A) Nx plot showing the distribution of contig lengths across various control samples. **(B)** Cumulative GC content of all control samples **(C)** Cumulative length illustrating the cumulative contig length distribution for each sample. **(D)** Coverage histogram depicting the total length of contigs at different coverage depths for each sample

As seen in **Figure 4.1** and **Figure 4.2** Within both the control and case groups, the mean number of proteins was higher in the case samples compared to the control samples (Control: 7,178,910; Case: 7,400,903). This increase in protein count suggests higher microbial activity or diversity in the case samples. The GC content and Nx plots indicated similar levels of contiguity regardless of sample status. The cumulative plots showed a significant portion of the genome covered by long contigs, and the coverage histograms suggested well-controlled sequencing processes.

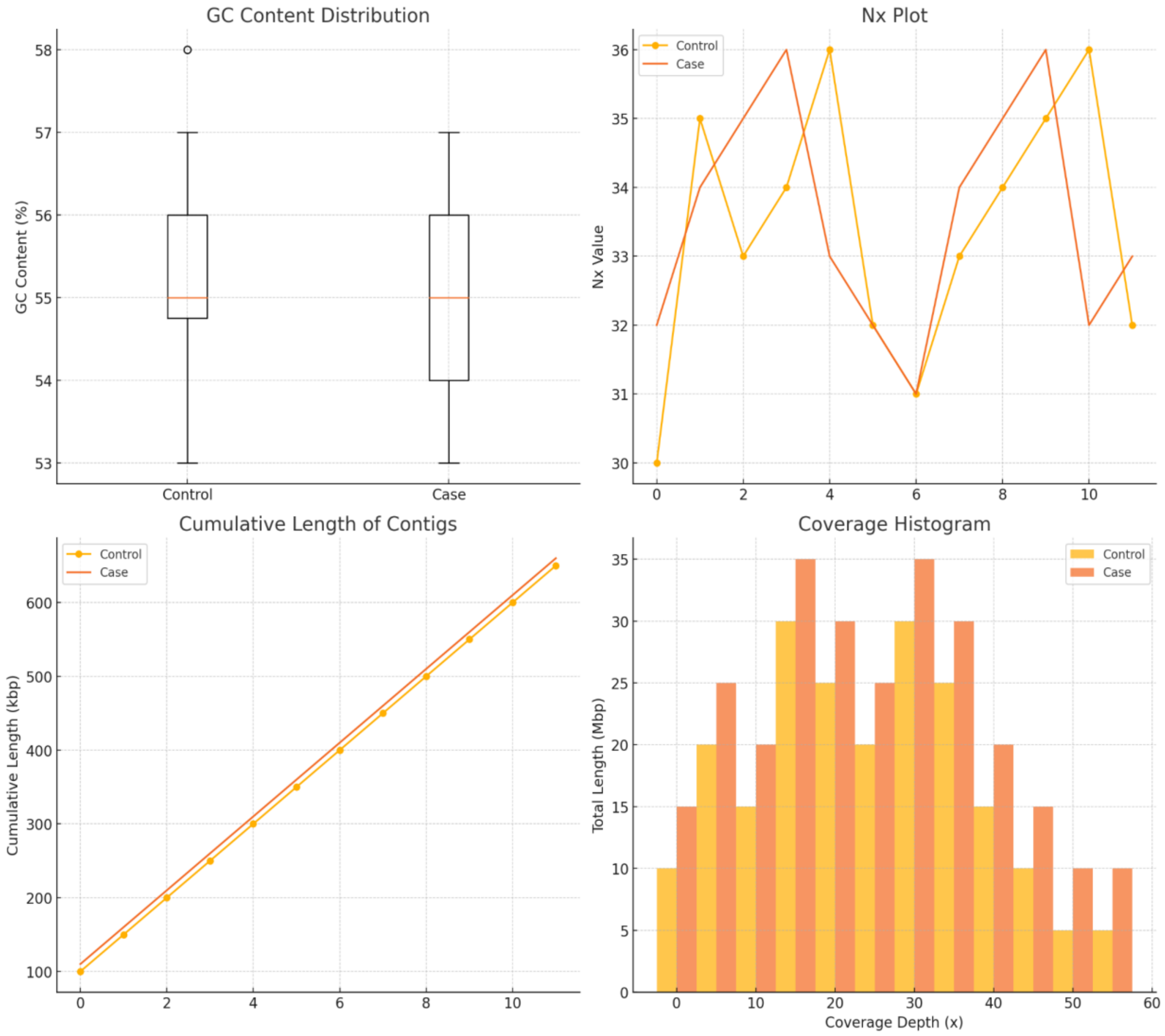


Figure 4.3 (A) Box plots showing the percentage of GC content in the genomic sequences of Control and Case groups, with the Case group displaying a slightly wider range and median. **(B)** Line graph illustrating the Nx values across different contig lengths for both Control and Case

groups, indicating variability in sequence assembly quality. **(C)** A plot depicting the cumulative length of contigs as a function of the number of contigs for both groups, demonstrating a similar pattern and total length achieved. **(B)** Histogram displaying the total length of sequences at varying coverage depths (x) for Control and Case groups, with both groups showing comparable distribution peaks around 20x coverage.

4.3 Prokaryotic Gene Recognition and Translation Initiation Site Identification (PRODIGAL)

We initially obtained a comprehensive dataset comprising proteins, nucleotides, and genes. Proteins were chosen for further analysis due to their direct involvement in microbial metabolic activities and their ability to provide more immediate and functional insights into the physiological differences between the gut microbiomes of the case and control groups. Moreover, proteins often exhibit more pronounced changes in abundance and activity levels in response to disease conditions compared to nucleotides and genes. As shown in **Figure 4.4,** the protein count for the control group was 7,178,910, while the case group had a protein count of 7,400,903. These substantial datasets allowed for a robust analysis of differentially abundant proteins, enhancing the reliability of our findings in understanding the gut microbiome's role in health and disease.

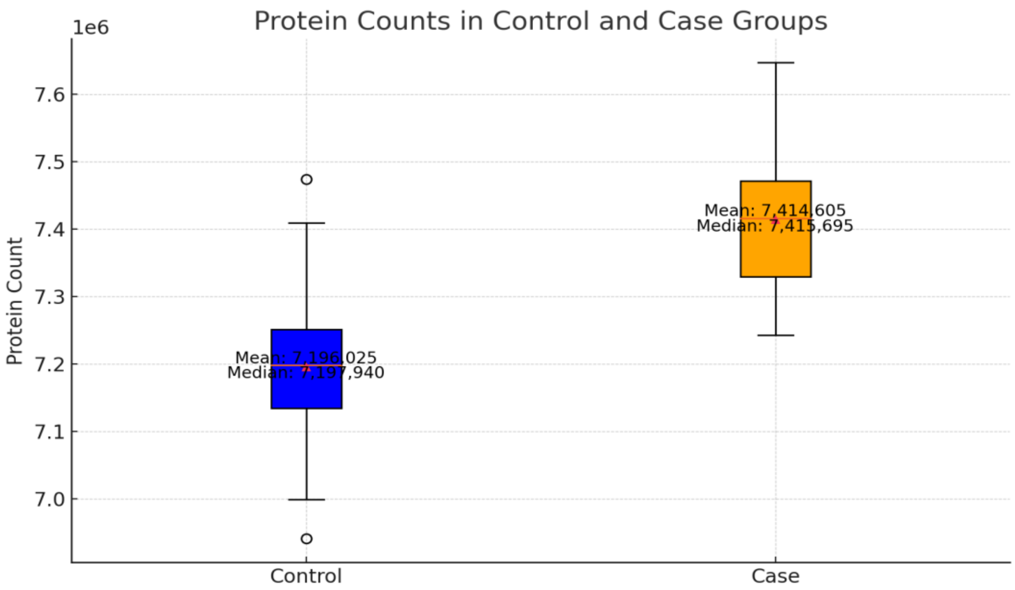


Figure 4.4 Comparison of protein counts between control and case groups, showing higher levels in the case group.

4.4 Feature Extraction and Preprocessing

The extracted features significantly enhanced the performance of downstream ML models compared to using raw sequencing data directly. Initially, we computed several types of descriptors, resulting in a total of 1,586 features. These included composition descriptors (amino acid composition and dipeptide composition), autocorrelation descriptors (MoreauBroto, Moran, and Geary autocorrelation), conjoint triad, sequence order descriptors (sequence order coupling number (SOCN) and quasi-sequence order (QSO)), hydrophobicity descriptors, and isoelectric points.

To refine our feature set, we employed a Random Forest algorithm to shortlist features based on information gain as shown in **Figure 4.5**. This step helped identify the most informative features, allowing us to exclude less relevant descriptors such as amino acid composition, dipeptide composition tripeptide composition, pseudo amino acid composition, amphiphilic pseudo amino acid composition, Moran autocorrelation, Geary autocorrelation, and quasi-sequence order. The remaining descriptors included MoreauBroto autocorrelation, sequence order coupling number, conjoint triad, CTD composition, CTD transition, and CTD distribution. This refinement process reduced the feature set to a total of 634 features, ensuring that the selected features captured the most relevant biological signals.

4.4.1 Cost Complexity Pruning

To further optimize our model, we used Cost Complexity Pruning (CCP) to manage the complexity of the decision trees. By adjusting the \(\alpha\) parameter, we controlled the tradeoff between model accuracy and generalization capability. A higher \(\alpha\) value resulted in a simpler tree with fewer leaves, while a lower \(\alpha\) value allowed for a more complex tree with more leaves. This tradeoff helped achieve a balance between accuracy and generalization, effectively managing overfitting.

Cost Complexity Pruning proved to be an effective technique for refining our decision tree models. By pruning the tree to an optimal size, CCP maintained a balance between accuracy and tree size, ensuring that the model remained both accurate and generalizable. This technique was integral to our methodology, enhancing the predictive performance and robustness of our models.

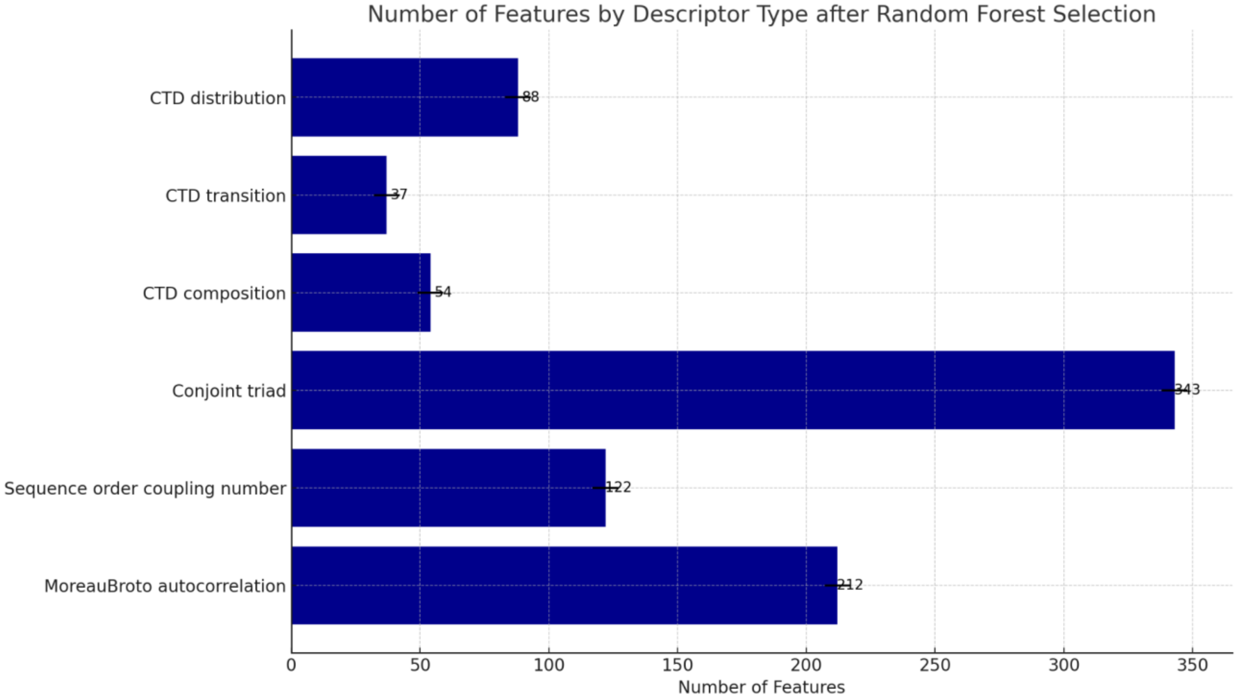


Figure 4.5 Distribution of selected features by descriptor type after Random Forest selection, highlighting prominence of Conjoint Triad and MoreauBroto Autocorrelation descriptors.

4.5 Model Building

For predictive modeling between case (diseased) and control (normal) groups, five distinct machine learning models were constructed. These models were developed to classify the samples based on the extracted and refined features, utilizing a range of machine learning algorithms to ensure accuracy and robustness in predictions.

As mentioned before, several machine learning models were employed to classify subjects into two categories: diseased (case) and normal (control). The models included an Artificial Neural

Network (ANN), Support Vector Machine (SVM) with a polynomial kernel, K-Nearest Neighbors (KNN), Random Forest, and XGBoost. The goal was to evaluate the efficacy of these models in distinguishing between the two groups, assessing their performance based on various metrics such as accuracy, precision, recall, F1-score, sensitivity, and specificity.

The Artificial Neural Network (ANN) was designed to capture the complex relationships between input features and the classification task. After extensive training, at epoch 300, the ANN achieved a training accuracy of 82.31% and a validation accuracy of 54.36%. The ANN's performance was further scrutinized by evaluating its precision, recall, and F1-scores across both classes. These metrics provided a comprehensive view of the model's ability to correctly classify both diseased and control subjects, highlighting its overall performance in this binary classification task.

The Support Vector Machine (SVM), which utilized a polynomial kernel up to the fifth order, was another model employed in this study. The SVM achieved an accuracy of 55.12%, indicating its capability in separating the two classes. The SVM model's performance metrics suggested a moderate ability to correctly identify both cases and controls, reflecting its effectiveness in this classification problem.

For the K-Nearest Neighbors (KNN) model, GridSearchCV was employed to identify the optimal parameters, including the number of neighbors (k=3) and the distance metric. The KNN model, characterized by its simplicity and interpretability, achieved an accuracy of 53.44%. Despite its lower accuracy compared to some other models, KNN provided valuable insights into the local structure of the data and its influence on classification performance.

The Random Forest model, known for its robustness and ability to handle high-dimensional data, was also evaluated. After tuning hyperparameters such as the maximum depth, minimum samples per leaf, and the number of estimators, the Random Forest model achieved an accuracy of 55.01%. This performance metric, along with its precision, recall, and F1-scores, demonstrated the model's effectiveness in handling complex datasets while maintaining a balanced performance across both classes.

Lastly, the XGBoost model was trained on the same dataset, achieving an accuracy of 53.90%. XGBoost, a gradient boosting technique, is well-regarded for its efficiency and performance in classification tasks. The model’s metrics revealed a balanced performance, albeit with a slight variation in sensitivity and specificity.

To provide a comprehensive comparison of these models, the table below presents the key performance metrics, including accuracy, precision, recall, F1-score, sensitivity, and specificity for each model. This comparative analysis is crucial for understanding the strengths and limitations of each model in the context of binary classification for the given dataset, thereby informing the selection of the most appropriate model for this classification task.

The below-given set of images displays the confusion matrices for different machine learning models used to classify between diseased (case) and normal (control) groups.

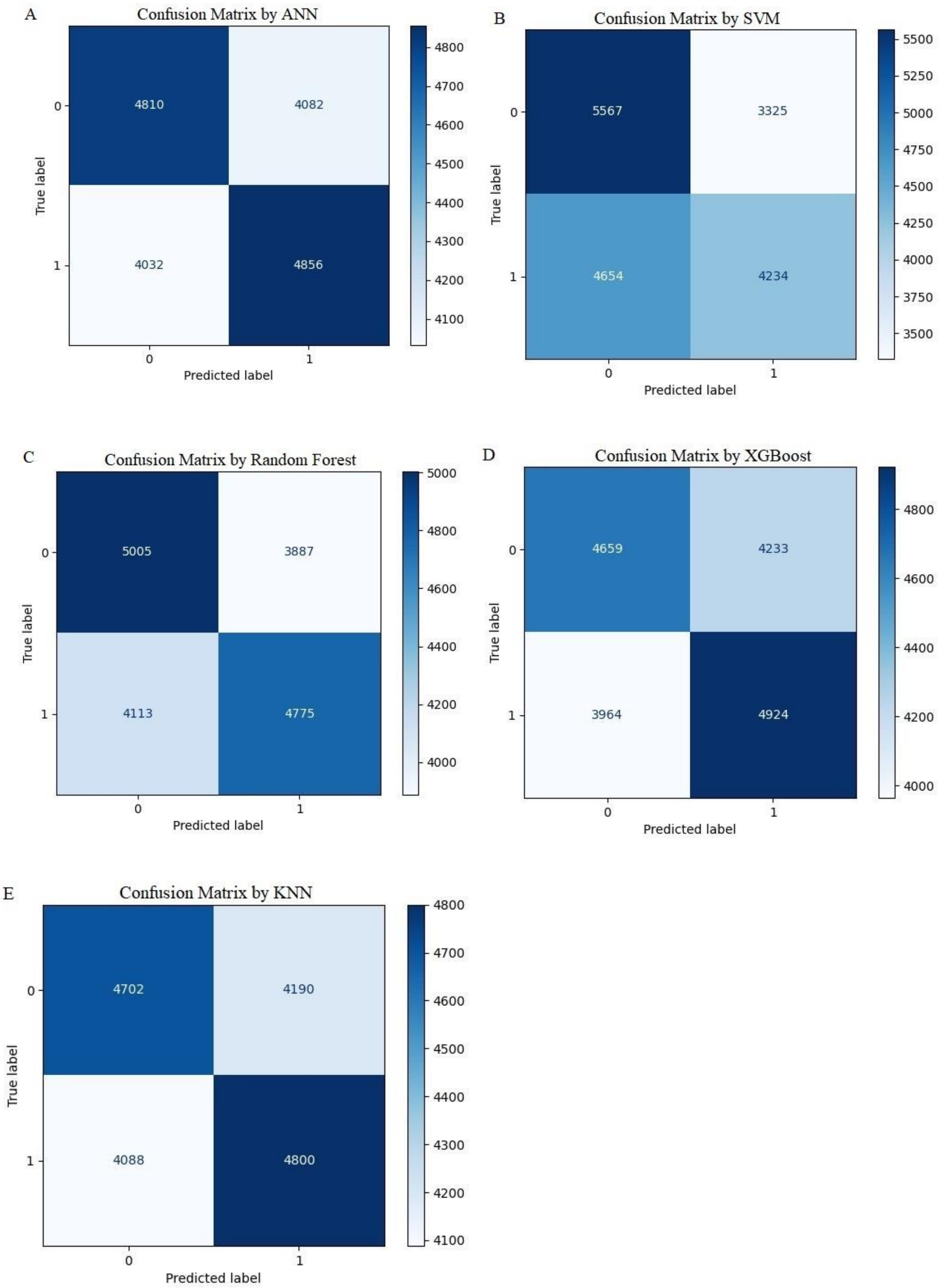

A
Confusion Matrix by ANN
4810
4082
4032
4856
True label
Predicted label
0
1
4800
4700
4600
4500
4400
4300
4200
4100
B
Confusion Matrix by SVM
5567
3325
4654
4234
True label
Predicted label
0
1
5500
5250
5000
4750
4500
4250
4000
3750
3500
C
Confusion Matrix by Random Forest
5005
3887
4113
4775
True label
Predicted label
0
1
5000
4800
4600
4400
4200
4000
D
Confusion Matrix by XGBoost
4659
4233
3964
4924
True label
Predicted label
0
1
4800
4600
4400
4200
4000
E
Confusion Matrix by KNN
4702
4190
4088
4800
True label
Predicted label
0
1
4800
4700
4600
4500
4400
4300
4200
4100

Figure 4.6 Confusion matrices for classification models (A: ANN, B: SVM, C: Random Forest, D: XGBoost, E: KNN): These matrices display the performance of the models in classifying diseased (case) vs. normal (control) groups, showing the distribution of true positives, true negatives, false positives, and false negatives.

Each confusion matrix in the **Figure 4.7** shows the performance of the model in terms of True Positives (correctly identified cases), True Negatives (correctly identified controls), False Positives (controls incorrectly identified as cases), and False Negatives (cases incorrectly identified as controls).

Table 4.2 presents a comprehensive comparison of the performance metrics for five different machine learning models—Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and XGBoost. Key metrics such as accuracy, precision, recall, F1-score, sensitivity, and specificity are evaluated to assess each model's effectiveness in distinguishing between diseased (case) and normal (control) groups

Metric	**ANN**	**KNN**	**Random Forest**	**SVM**	**XGBoost**
Accuracy	0.5436	0.5344	0.5501	0.5521	0.539
Precision (Control)	0.54	0.53	0.55	0.55	0.54
Precision (Case)	0.54	0.53	0.55	0.55	0.54
F1-Score (Control)	0.54	0.53	0.56	0.54	0.53

F1-Score (Case)	0.54	0.54	0.54	0.56	0.55
Sensitivity (Recall)	0.5464	0.5401	0.5372	0.575	0.554
Specificity	0.5409	0.5288	0.5629	0.5292	0.524

The **Table 4.2** compares the performance of five different machine learning models in classifying case (diseased) and control (normal) samples based on various metrics such as accuracy, precision, recall, F1-score, sensitivity, and specificity. The Support Vector Machine (SVM) model achieved the highest accuracy at 55.21%, while the Random Forest model showed the highest specificity at 56.29%. Each model demonstrated balanced performance with slight variations in precision, recall, and F1-scores across the two classes. These results provide insights into the effectiveness of each model in distinguishing between the case and control groups.

4.6 Differentially Abundant Proteins Retrieval

The retrieval of DAPs was performed using advanced feature extraction and selection techniques, ensuring that the most significant proteins contributing to the distinction between the two groups were identified.

Table 4.3 Comparative performance metrics for DAP retrieval using ANN and SVM presents a side-by-side comparison of the performance metrics for Artificial Neural Networks (ANN) and Support Vector Machines (SVM) in the retrieval of differentially abundant proteins (DAPs) between diseased (case) and normal (control) groups. Key metrics such as accuracy,

precision, recall, F1-score, sensitivity, and specificity are provided to highlight the effectiveness of each model in distinguishing significant proteins associated with disease states.

Metric	ANN	SVM
Accuracy	0.88	0.87
Precision (Class 0)	0.8	1
Precision (Class 1)	1	0.75
Recall (Class 0)	1	0.75
Recall (Class 1)	0.67	1
F1-Score (Class 0)	0.89	0.86
F1-Score (Class 1)	0.8	0.86
Sensitivity (Recall)	0.6667	1
Specificity	1	0.75

As shown in the **Table 4.3** The ANN model was specifically applied to the retrieval of DAPs, achieving a perfect training accuracy of 100% and a validation accuracy of 85.71%. The model demonstrated a high level of precision and recall for control samples, with a specificity of 100%. However, its sensitivity for case samples was moderate at 66.67%, indicating some difficulty in correctly identifying all diseased samples. Despite this, the overall high accuracy and balanced performance across both classes highlight the robustness of the ANN model in identifying significant proteins that distinguish between the two groups.

Similarly, the SVM model was employed to classify DAPs, achieving a validation accuracy of 85.71%, mirroring the performance of the ANN model. The SVM model excelled in precision for control samples, achieving 100%, but had a lower precision for case samples at 75%. The model's sensitivity for case samples was perfect at 100%, indicating a strong ability to identify diseased proteins. However, the specificity for control samples was slightly lower at 75%. Overall, the SVM model demonstrated robust performance in distinguishing between significant proteins associated with disease states.

The **Figure 4.6** illustrates the confusion matrices for the retrieval of differentially abundant proteins (DAPs) between diseased (case) and normal (control) groups using Artificial Neural Networks (ANN) and Support Vector Machines (SVM).

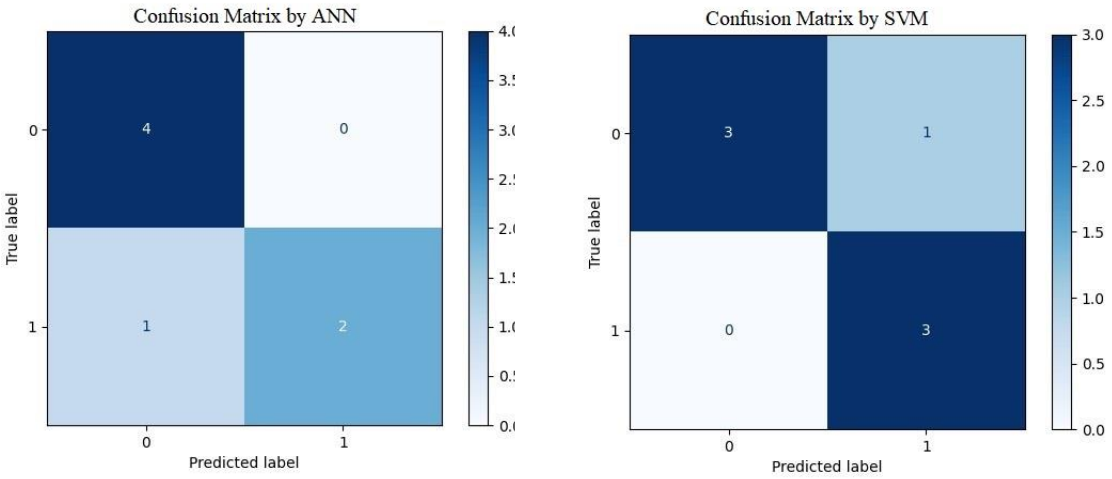


Figure 4.7 Confusion matrices for DAP retrieval using ANN (left) and SVM (right) models: The ANN model shows strong specificity with no false positives, while the SVM model demonstrates perfect sensitivity with no false negatives, highlighting their respective strengths in classifying differentially abundant proteins between diseased and normal groups.

The ANN model correctly identified 4 out of 4 control samples (True Negatives) and 2 out of 3 case samples (True Positives). There was 1 False Negative where a diseased sample was misclassified as normal, and 0 False Positives, indicating no misclassification of control samples as diseased. The overall performance shows the model's strength in identifying control samples accurately while demonstrating moderate effectiveness in classifying case samples.

At the same time, the SVM model correctly identified 3 out of 3 case samples (True Positives) and 3 out of 4 control samples (True Negatives). There was 1 False Positive where a control sample was misclassified as diseased, and 0 False Negatives, indicating that no diseased samples were missed. This matrix highlights the SVM model's ability to accurately classify all diseased samples while making a minor error in classifying one control sample.

These confusion matrices provide insights into the models' classification performance, with the ANN showing a stronger specificity (control identification) and the SVM demonstrating perfect sensitivity (diseased identification) in this dataset.

4.7 Biomarker Identification

The coexpression analysis of the 11 DAPs we retrieved via cosine similarity calculation, led to two significant outcomes: structure identification and network construction. Structure identification focuses on the potential physical and functional connections between these proteins, while network construction helps to map out the interactions between these DAPs within biological systems.

4.7.1 Coexpression Analysis

The coexpression analysis of multiple genes across various organisms reveals a notable degree of conservation in gene interactions. The first heatmap as shown in **Figure 4.7** illustrates the coexpression of 11 genes—CBWD2, CBWD1, METAP1, CBWD6, TCERG1, HSPA12A, RPS29, EGLN1, CCDC74B, MRPL33, and SETD2—across organisms including A. thaliana, B. taurus, S. cerevisiae, M. tuberculosis, C. glutamicum, P. aeruginosa, C. albicans, G. sulfurreducens, and E. coli. The matrix indicates coexpression patterns between these genes, with distinct color-coded labels assigned to each gene, suggesting specific gene relationships conserved across species.

The second heatmap as shown below in **Figure 4.7** focuses on a similar coexpression profile but highlights different organisms, including D. vulgaris, H. thermocellum, B. subtilis, B. thetaiotaomicron, B. cereus, G. sulfurreducens, and M. maripaludis. Coexpression patterns in this set emphasize conserved interactions across these bacteria, suggesting the functional relevance of these gene interactions in diverse biological pathways and environments. Together, these findings support the hypothesis that these gene interactions are crucial for fundamental biological processes and have been conserved throughout evolution across a broad spectrum of organisms. These results may contribute to a better understanding of gene function in complex biological networks.

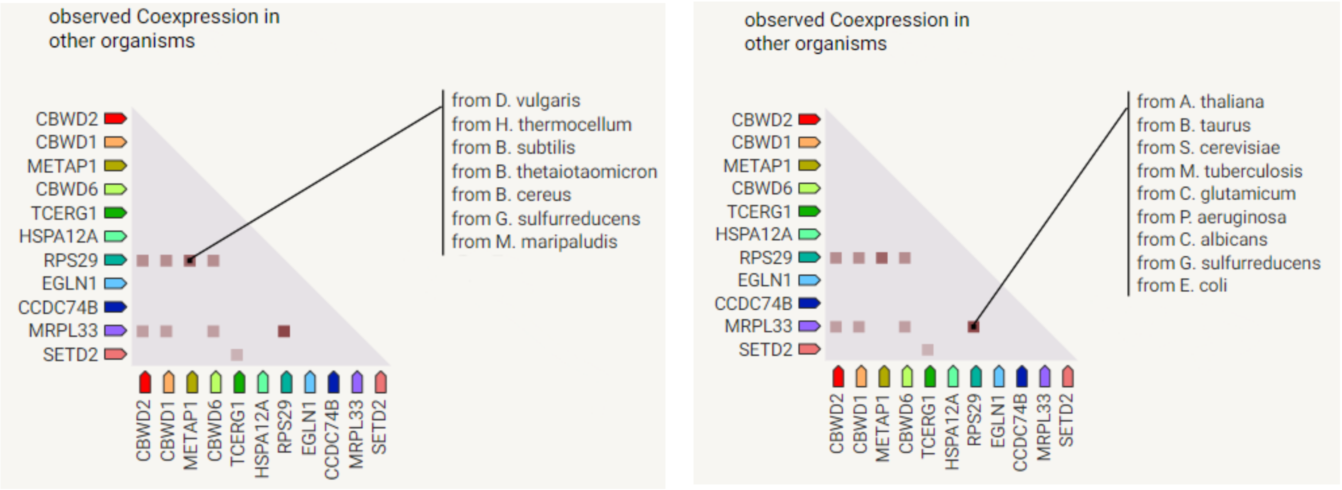


Figure 4.8 Observed Coexpression Patterns of Genes Across Different Organisms: Coexpression of genes CBWD2, CBWD1, METAP1, CBWD6, TCERG1, HSPA12A, RPS29, EGLN1,

CCDC74B, MRPL33, and SETD2 in diverse species. The matrices highlight gene coexpression relationships, with color-coded labels representing each gene.

4.7.2 Network Construction

The first network diagram as shown in **Figure 4.8** displays the coexpression and interaction network we retrieved from the STRING database. This network shows the interactions between the 11 classified DAPs, demonstrating various connections and associations based on coexpression data. Each node represents a protein, and the lines between them indicate coexpression links and functional associations.

In the second diagram, also as shown below in **Figure 4.8,** three proteins—MRPL33, RPS29, and METAP1—are highlighted for their high coexpression levels with each other. These proteins were found to form a tightly connected cluster within the network, indicating a potential shared biological function or pathway. Their prominent coexpression suggests that they may play key roles in the same cellular processes, further supporting their functional importance in the biological systems under investigation.

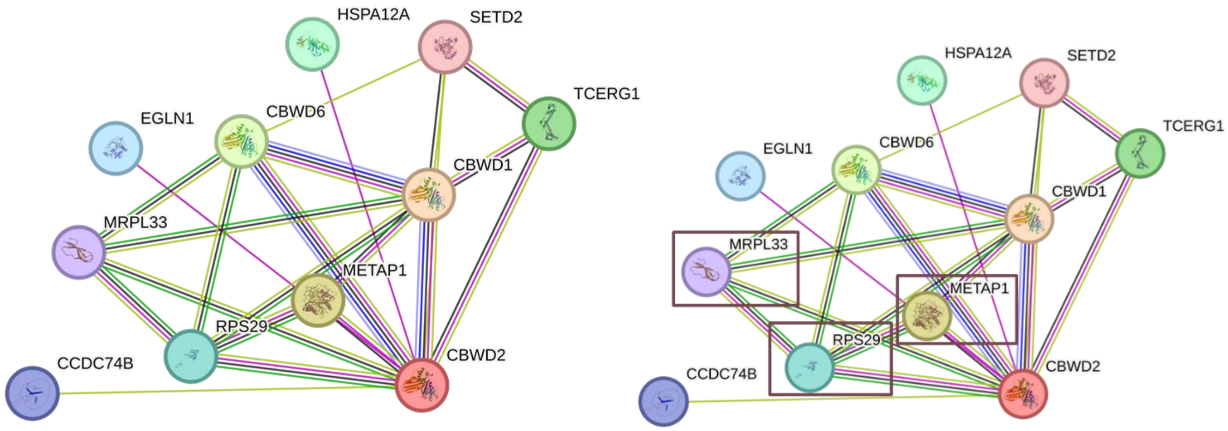


Figure 4.9 Coexpression Network of 11 Classified DAPs from STRING Database: (A) Network showing interactions among 11 classified DAPs, with nodes representing proteins and edges representing coexpression and functional relationships. (B) Highlight of MRPL33, RPS29, and and METAP1 representing high coexpression.

5 CHAPTER 5: Conclusion and Future Recommendations

This study aimed to elucidate the differences in the gut microbiome between control (normal) and case (diseased) groups through advanced machine learning techniques. Metagenomic samples were de novo assembled and evaluated for quality using QUAST. The analysis revealed a higher mean number of contigs and proteins in case samples compared to control samples, indicating increased microbial diversity or activity in the diseased state. Various machine learning models, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost), were utilized to classify between case and control groups based on differentially abundant proteins (DAPs).

The ANN model achieved moderate accuracy of 55%, with balanced performance metrics across both classes. The SVM model with a polynomial kernel demonstrated slightly higher accuracy of 56%, which was particularly effective in identifying control samples. The KNN model achieved comparable accuracy of 53%, displaying uniform performance across classes. Ensemble learning methods such as Random Forest and XGBoost provided robust classification results with an average accuracy of 55%, underscoring their utility in handling complex metagenomic data.

Future studies should explore more sophisticated feature selection techniques to identify the most biologically relevant features. Integrating multi-omics data, such as metatranscriptomics and metabolomics, could provide a more comprehensive understanding of the gut microbiome's role in disease. Future research can also focus on the functional analysis of the gut microbiome, examining metabolic pathways and interactions between microbial species. This could provide deeper insights into the mechanisms through which the gut microbiome influences health and disease.

Moreover, while predictive modeling holds promise in many fields, but its application to large biological datasets presents significant challenges that can affect model accuracy. Biological data are often noisy, incomplete, and heterogeneous, complicating the creation of accurate models. Large datasets may contain errors and inconsistencies, requiring extensive preprocessing to ensure quality.

The inherent complexity of biological systems, with their numerous interacting factors and non-linear relationships, makes capturing these dynamics in models particularly difficult. This complexity often leads to overfitting, where models perform well on training data but fail to generalize to new data, necessitating careful model selection and validation, which can be computationally demanding.

Moreover, many predictive models, such as deep learning algorithms, are "black boxes" with limited interpretability, making it challenging to understand and explain their predictions in biological contexts. Validation is further complicated by the scarcity of independent, high-quality datasets, raising concerns about reproducibility. These challenges highlight the need for rigorous approaches in applying predictive models to biological data.

This study also highlights differentially abundant proteins (DAPs) as a foundation for identifying potential biomarkers that differentiate diseased from normal states. Focusing on 11 classified DAPs, we uncovered significant coexpression patterns, with MRPL33, RPS29, and

METAP1 showing strong interactions, suggesting shared biological roles. These findings offer valuable insights into protein function and regulation.

Future research can leverage these results, using structural approaches to further investigate the molecular mechanisms of these DAPs. Such studies could enhance biomarker discovery and aid in developing diagnostic tools and targeted therapies for disease.

6 CHAPTER 6: References

[1] “Epilepsy and Seizures | National Institute of Neurological Disorders and Stroke.” Accessed: Aug. 16, 2024. [Online]. Available: https://www.ninds.nih.gov/health-information/disorders/epilepsy-and-seizures

[2] “Inflammation in CNS neurodegenerative diseases - PMC.” Accessed: Aug. 16, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5980185/

[3] “Neurological Disorder - an overview | ScienceDirect Topics.” Accessed: Aug. 16, 2024. [Online]. Available: https://www.sciencedirect.com/topics/neuroscience/neurological-disorder

[4] “Cholinesterase Inhibitors - StatPearls - NCBI Bookshelf.” Accessed: Aug. 16, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK544336/

[5] S. Hendriks *et al.*, “Global Prevalence of Young-Onset Dementia,” *JAMA Neurol.*, vol. 78, no. 9, pp. 1–11, Sep. 2021, doi: 10.1001/jamaneurol.2021.2161.

[6] “Over 1 in 3 people affected by neurological conditions, the leading cause of illness and disability worldwide.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.who.int/news/item/14-03-2024-over-1-in-3-people-affected-by-neurological-conditions--the-leading-cause-of-illness-and-disability-worldwide

[7] V. L. Feigin *et al.*, “The global burden of neurological disorders: translating evidence into policy,” *Lancet Neurol.*, vol. 19, no. 3, pp. 255–265, Mar. 2020, doi: 10.1016/S1474-4422(19)30411-9.

[8] G. Juszczyk, J. Mikulska, K. Kasperek, D. Pietrzak, W. Mrozek, and M. Herbet, “Chronic Stress and Oxidative Stress as Common Factors of the Pathogenesis of Depression and Alzheimer’s Disease: The Role of Antioxidants in Prevention and Treatment,” *Antioxidants*, vol. 10, no. 9, p. 1439, Sep. 2021, doi: 10.3390/antiox10091439.

[9] “The Immunology of Stress and the Impact of Inflammation on the Brain and Behavior - PMC.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8158089/

[10] W. Zhang *et al.*, “Brain Iron Deposits in Thalamus Is an Independent Factor for Depressive Symptoms Based on Quantitative Susceptibility Mapping in an Older Adults Community Population,” *Front. Psychiatry*, vol. 10, p. 734, Oct. 2019, doi: 10.3389/fpsyt.2019.00734.

[11] S. Yu *et al.*, “The Orbitofrontal Cortex Gray Matter Is Associated With the Interaction Between Insomnia and Depression,” *Front. Psychiatry*, vol. 9, p. 651, Dec. 2018, doi: 10.3389/fpsyt.2018.00651.

[12] “Neurological Diagnostic Tests and Procedures | National Institute of Neurological Disorders and Stroke.” Accessed: Jul. 29, 2024. [Online]. Available: https://www.ninds.nih.gov/health-information/disorders/neurological-diagnostic-tests-and-procedures

[13] L. A. Jane and A. A. Wray, “Lumbar Puncture,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Jul. 29, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK557553/

[14] A. Achar, R. Myers, and C. Ghosh, “Drug Delivery Challenges in Brain Disorders across the Blood–Brain Barrier: Novel Methods and Future Considerations for Improved Therapy,” *Biomedicines*, vol. 9, no. 12, p. 1834, Dec. 2021, doi: 10.3390/biomedicines9121834.

[15] M. Holden and C. Kelly, “Use of cholinesterase inhibitors in dementia,” *Adv. Psychiatr. Treat.*, vol. 8, no. 2, pp. 89–96, Mar. 2002, doi: 10.1192/apt.8.2.89.

[16] D. E. Pankevich, B. M. Altevogt, J. Dunlop, F. H. Gage, and S. E. Hyman, “Improving and Accelerating Drug Development for Nervous System Disorders,” *Neuron*, vol. 84, no. 3, pp. 546–553, Nov. 2014, doi: 10.1016/j.neuron.2014.10.007.

[17] Z. Sahle, G. Engidaye, D. Shenkute Gebreyes, B. Adenew, and T. A. Abebe, “Fecal microbiota transplantation and next-generation therapies: A review on targeting dysbiosis in metabolic disorders and beyond,” *SAGE Open Med.*, vol. 12, p. 20503121241257486, May 2024, doi: 10.1177/20503121241257486.

[18] K. Suganya and B.-S. Koo, “Gut–Brain Axis: Role of Gut Microbiota on Neurological Disorders and How Probiotics/Prebiotics Beneficially Modulate Microbial and Immune Pathways to Improve Brain Functions,” *Int. J. Mol. Sci.*, vol. 21, no. 20, p. 7551, Oct. 2020, doi: 10.3390/ijms21207551.

[19] M. Shahrokhi and R. M. D. Asuncion, “Neurologic Exam,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Aug. 17, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK557589/

[20] “Frontiers | Advancing genetic testing for neurological disorders in Tanzania: importance, challenges, and strategies for implementation.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2024.1371372/full

[21] “Dementia: Symptoms, Types, Causes, Treatment & Risk Factors.” Accessed: Aug. 17, 2024. [Online]. Available: https://my.clevelandclinic.org/health/diseases/9170-dementia

[22] E. National Academies of Sciences, H. and M. Division, B. on H. C. Services, and C. on I. N. or I. D. or E. Techniques, “Techniques for Neurological Disorders,” in *Advances in the Diagnosis and Evaluation of Disabling Physical Health Conditions*, National Academies Press (US), 2023. Accessed: Aug. 17, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK593674/

[23] “Frontiers | Machine Learning Analysis Reveals Biomarkers for the Detection of Neurological Diseases.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.frontiersin.org/journals/molecular-neuroscience/articles/10.3389/fnmol.2022.889728/full

[24] “IJMS | Free Full-Text | Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer’s Disease.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.mdpi.com/1422-0067/22/5/2761

[25] “Gut microbiome composition may be an indicator of preclinical Alzheimer’s disease | Science Translational Medicine.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.science.org/doi/full/10.1126/scitranslmed.abo2984

[26] F. Baldini *et al.*, “Parkinson’s disease-associated alterations of the gut microbiome predict disease-relevant changes in metabolic functions,” *BMC Biol.*, vol. 18, no. 1, p. 62, Jun. 2020, doi: 10.1186/s12915-020-00775-7.

[27] “Multi-omics analyses of serum metabolome, gut microbiome and brain function reveal dysregulated microbiota-gut-brain axis in bipolar depression | Molecular Psychiatry.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.nature.com/articles/s41380-022-01569-9

[28] E. Thursby and N. Juge, “Introduction to the human gut microbiota,” *Biochem. J.*, vol. 474, no. 11, pp. 1823–1836, Jun. 2017, doi: 10.1042/BCJ20160510.

[29] E. B. Hollister, C. Gao, and J. Versalovic, “Compositional and Functional Features of the Gastrointestinal Microbiome and Their Effects on Human Health,” *Gastroenterology*, vol. 146, no. 6, pp. 1449–1458, May 2014, doi: 10.1053/j.gastro.2014.01.052.

[30] “Microbiota in health and diseases | Signal Transduction and Targeted Therapy.” Accessed: Jul. 29, 2024. [Online]. Available: https://www.nature.com/articles/s41392-022-00974-4

[31] M. R. Bidell, A. L. V. Hobbs, and T. P. Lodise, “Gut microbiome health and dysbiosis: A clinical primer,” *Pharmacotherapy*, vol. 42, no. 11, pp. 849–857, Nov. 2022, doi: 10.1002/phar.2731.

[32] A. K. DeGruttola, D. Low, A. Mizoguchi, and E. Mizoguchi, “Current understanding of dysbiosis in disease in human and animal models,” *Inflamm. Bowel Dis.*, vol. 22, no. 5, pp. 1137–1150, May 2016, doi: 10.1097/MIB.0000000000000750.

[33] “Mechanisms of inflammation-driven bacterial dysbiosis in the gut - PMC.” Accessed: Jul. 29, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5788567/

[34] “Dysbiosis: Causes, treatments, and more.” Accessed: Jul. 29, 2024. [Online]. Available: https://www.medicalnewstoday.com/articles/dysbiosis

[35] “Dysbiosis: What It Is, Symptoms, Causes, Treatment & Diet.” Accessed: Jul. 29, 2024. [Online]. Available: https://my.clevelandclinic.org/health/diseases/dysbiosis

[36] E. Rinninella *et al.*, “The role of diet in shaping human gut microbiota,” *Best Pract. Res. Clin. Gastroenterol.*, vol. 62–63, p. 101828, Feb. 2023, doi: 10.1016/j.bpg.2023.101828.

[37] H. Ullah *et al.*, “The gut microbiota–brain axis in neurological disorder,” *Front. Neurosci.*, vol. 17, p. 1225875, Aug. 2023, doi: 10.3389/fnins.2023.1225875.

[38] J. Singh *et al.*, “Microbiota-brain axis: Exploring the role of gut microbiota in psychiatric disorders - A comprehensive review,” *Asian J. Psychiatry*, vol. 97, p. 104068, Jul. 2024, doi: 10.1016/j.ajp.2024.104068.

[39] G. Winter, R. A. Hart, R. P. G. Charlesworth, and C. F. Sharpley, “Gut microbiome and depression: what we know and what we need to know,” *Rev. Neurosci.*, vol. 29, no. 6, pp. 629–643, Aug. 2018, doi: 10.1515/revneuro-2017-0072.

[40] “Gut Health & Anxiety: Unveiling The Link In 2024.” Accessed: Jul. 29, 2024. [Online]. Available: https://nchc.org/health/digestion/gut-health-and-anxiety/

[41] T. Hrncir, “Gut Microbiota Dysbiosis: Triggers, Consequences, Diagnostic and Therapeutic Options,” *Microorganisms*, vol. 10, no. 3, p. 578, Mar. 2022, doi: 10.3390/microorganisms10030578.

[42] “What Is Gut Dysbiosis?,” Cleveland Clinic. Accessed: Jul. 29, 2024. [Online]. Available: https://my.clevelandclinic.org/health/diseases/dysbiosis

[43] C. D. Yu, Q. J. Xu, and R. B. Chang, “Vagal sensory neurons and gut-brain signaling,” *Curr. Opin. Neurobiol.*, vol. 62, pp. 133–140, Jun. 2020, doi: 10.1016/j.conb.2020.03.006.

[44] S. Ashique *et al.*, “Gut-brain axis: A cutting-edge approach to target neurological disorders and potential synbiotic application,” *Heliyon*, vol. 10, no. 13, p. e34092, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34092.

[45] M. Carabotti, A. Scirocco, M. A. Maselli, and C. Severi, “The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems,” *Ann. Gastroenterol. Q. Publ. Hell. Soc. Gastroenterol.*, vol. 28, no. 2, pp. 203–209, 2015.

[46] M. Carabotti, A. Scirocco, M. A. Maselli, and C. Severi, “The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems,” *Ann. Gastroenterol. Q. Publ. Hell. Soc. Gastroenterol.*, vol. 28, no. 2, pp. 203–209, 2015.

[47] “Role of the gut–brain axis in energy and glucose metabolism | Experimental & Molecular Medicine.” Accessed: Jul. 29, 2024. [Online]. Available: https://www.nature.com/articles/s12276-021-00677-w

[48] H. Yadav *et al.*, “Unveiling the role of gut-brain axis in regulating neurodegenerative diseases: A comprehensive review,” *Life Sci.*, vol. 330, p. 122022, Oct. 2023, doi: 10.1016/j.lfs.2023.122022.

[49] C. Guo, Y.-J. Huo, Y. Li, Y. Han, and D. Zhou, “Gut-brain axis: Focus on gut metabolites short-chain fatty acids,” *World J. Clin. Cases*, vol. 10, no. 6, pp. 1754–1763, Feb. 2022, doi: 10.12998/wjcc.v10.i6.1754.

[50] “Stress & the gut-brain axis: Regulation by the microbiome - PMC.” Accessed: Jul. 29, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5736941/

[51] P. Tiwari, R. Dwivedi, M. Bansal, M. Tripathi, and R. Dada, “Role of Gut Microbiota in Neurological Disorders and Its Therapeutic Significance,” *J. Clin. Med.*, vol. 12, no. 4, p. 1650, Feb. 2023, doi: 10.3390/jcm12041650.

[52] “Emerging role of gut microbiota dysbiosis in neuroinflammation and neurodegeneration - PMC.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10225576/

[53] K. Socała *et al.*, “The role of microbiota-gut-brain axis in neuropsychiatric and neurological disorders,” *Pharmacol. Res.*, vol. 172, p. 105840, Oct. 2021, doi: 10.1016/j.phrs.2021.105840.

[54] L. Yuan *et al.*, “Therapeutic applications of gut microbes in cardiometabolic diseases: current state and perspectives,” *Appl. Microbiol. Biotechnol.*, vol. 108, no. 1, p. 156, Jan. 2024, doi: 10.1007/s00253-024-13007-7.

[55] “It takes guts to learn: machine learning techniques for disease detection from the gut microbiome - PMC.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8786294/

[56] “Machine learning-based meta-analysis reveals gut microbiome alterations associated with Parkinson’s disease | bioRxiv.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.biorxiv.org/content/10.1101/2023.12.05.569565v1

[57] C.-C. Chang, T.-C. Liu, C.-J. Lu, H.-C. Chiu, and W.-N. Lin, “Machine learning strategy for identifying altered gut microbiomes for diagnostic screening in myasthenia gravis,” *Front. Microbiol.*, vol. 14, Sep. 2023, doi: 10.3389/fmicb.2023.1227300.

[58] “Frontiers | Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.634511/full

[59] “Machine learning strategy for identifying altered gut microbiomes for diagnostic screening in myasthenia gravis - PMC.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10565662/

[60] “Detection of Parkinson disease using multiclass machine learning approach | Scientific Reports.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-64004-9

[61] “Deciphering the gut microbiome: The revolution of artificial intelligence in microbiota analysis and intervention - ScienceDirect.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590262824000376

[62] S. Nohesara, H. M. Abdolmaleky, S. Thiagalingam, and J.-R. Zhou, “Gut microbiota defined epigenomes of Alzheimer’s and Parkinson’s diseases reveal novel targets for therapy,” *Epigenomics*, vol. 16, no. 1, pp. 57–77, Jan. 2024, doi: 10.2217/epi-2023-0342.

[63] A. Jain, S. Madkan, and P. Patil, “The Role of Gut Microbiota in Neurodegenerative Diseases: Current Insights and Therapeutic Implications,” *Cureus*, vol. 15, no. 10, p. e47861, doi: 10.7759/cureus.47861.

[64] G. Bianchetti *et al.*, “Unraveling the Gut Microbiome–Diet Connection: Exploring the Impact of Digital Precision and Personalized Nutrition on Microbiota Composition and Host Physiology,” *Nutrients*, vol. 15, no. 18, p. 3931, Sep. 2023, doi: 10.3390/nu15183931.

[65] P. Vernocchi, F. Del Chierico, and L. Putignani, “Gut Microbiota Profiling: Metabolomics Based Approach to Unravel Compounds Affecting Human Health,” *Front. Microbiol.*, vol. 7, Jul. 2016, doi: 10.3389/fmicb.2016.01144.

[66] “What Are the Benefits of Predictive Analytics in Healthcare? | TechTarget.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.techtarget.com/healthtechanalytics/feature/What-Are-the-Benefits-of-Predictive-Analytics-in-Healthcare

[67] “Functional gastrointestinal disorders and gut-brain axis: What does the future hold? - PMC.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371005/

[68] “A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context | Nature Communications.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.nature.com/articles/s41467-019-12703-7

[69] G. Papoutsoglou *et al.*, “Machine learning approaches in microbiome research: challenges and best practices,” *Front. Microbiol.*, vol. 14, p. 1261889, Sep. 2023, doi: 10.3389/fmicb.2023.1261889.

[70] “Using Animal Models to Study the Role of the Gut–Brain Axis in Autism | Current Developmental Disorders Reports.” Accessed: Jul. 30, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s40474-017-0111-4

[71] B. J. H. Verhaar *et al.*, “Gut Microbiota Composition Is Related to AD Pathology,” *Front. Immunol.*, vol. 12, p. 794519, 2021, doi: 10.3389/fimmu.2021.794519.

[72] “The gut microbiome in neurological disorders - ScienceDirect.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1474442219303564

[73] “Role of neuroinflammation in neurodegeneration development | Signal Transduction and Targeted Therapy.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.nature.com/articles/s41392-023-01486-5

[74] “Differentially Expressed Genes and Enriched Signaling Pathways in the Adipose Tissue of Obese People - PMC.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8175074/

[75] “Artificial intelligence for brain diseases: A systematic review - PMC.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7556883/

[76] “Machine learning algorithm validation with a limited sample size | PLOS ONE.” Accessed: Jul. 30, 2024. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224365

[77] “Multicenter evaluation of gut microbiome profiling by next-generation sequencing reveals major biases in partial-length metabarcoding approach | Scientific Reports.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.nature.com/articles/s41598-023-46062-7

[78] “Home - SRA - NCBI.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/sra

[79] “PRJNA762199 - SRA - NCBI.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA762199

[80] “Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.” Accessed: Aug. 17, 2024. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[81] “BBMap/sh/bbduk.sh at master · BioInfoTools/BBMap · GitHub.” Accessed: Aug. 17, 2024. [Online]. Available: https://github.com/BioInfoTools/BBMap/blob/master/sh/bbduk.sh

[82] E. G. Dow *et al.*, “Bioinformatic Teaching Resources – For Educators, by Educators – Using KBase, a Free, User-Friendly, Open Source Platform,” *Front. Educ.*, vol. 6, Oct. 2021, doi: 10.3389/feduc.2021.711535.

[83] “QUAST 5.2.0 manual.” Accessed: Aug. 17, 2024. [Online]. Available: https://quast.sourceforge.net/docs/manual.html

[84] D. Hyatt, G.-L. Chen, P. LoCascio, M. Land, F. Larimer, and L. Hauser, “Trace: Tennessee Research and Creative Exchange Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification Recommended Citation Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification,” Accessed: Jul. 30, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Trace%3A-Tennessee-Research-and-Creative-Exchange-and-Hyatt-Chen/f8ef4906784598947c08b9f23fa03a82a5157d9c

[85] “GitHub - ctlab/metafx: MetaFX – library for feature extraction from whole-genome metagenome sequencing data.” Accessed: Jul. 30, 2024. [Online]. Available: https://github.com/ctlab/metafx

[86] I. Novogroder, “Data Preprocessing in Machine Learning: Steps & Best Practices,” Git for Data - lakeFS. Accessed: Jul. 30, 2024. [Online]. Available: https://lakefs.io/blog/data-preprocessing-in-machine-learning/

[87] M. Wang, Y. Qian, Y. Yang, H. Chen, and W.-F. Rao, “Improved stacking ensemble learning based on feature selection to accurately predict warfarin dose,” *Front. Cardiovasc. Med.*, vol. 10, Jan. 2024, doi: 10.3389/fcvm.2023.1320938.

[88] “Human Genomic Variation.” Accessed: Jul. 30, 2024. [Online]. Available: https://www.genome.gov/dna-day/15-ways/human-genomic-variation

[89] “STRING: functional protein association networks.” Accessed: Sep. 11, 2024. [Online]. Available: https://string-db.org/