

A MACHINE LEARNING FRAMEWORK for the DETECTION of ALZHEIMER'S DISEASE USING GUT FUNGAL ITS SEQUENCING DATA

**Thesis Submitted
In Partial Fulfillment of the Requirements for
the Degree of**

**MASTER OF TECHNOLOGY
in
BIOINFORMATICS
by**

**MAYANK SINGH
(2K24/BIO/01)**

**Under the Supervision of
PROF. YASHA HASIJA
Delhi Technological University**

Department of Biotechnology



**DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India**

May, 2026

ACKNOWLEDGEMENT

I would like to thank my supervisor **Prof. Yasha Hasija** for her guidance, continuous support and patience during this research. Her wisdom and foresight had a significant influence on the course of this thesis and they have been a great source of inspiration for my studies.

I would like to thank the **Department of Biotechnology, Delhi Technological University** for providing the necessary infrastructure and academic environment. I am also thankful to the faculty and technical staff of the Department for their constant support.

I would like to express my gratitude to the **DTU Central Library** for the serene environment and access to literature and academic information that was essential for this thesis.

I'm grateful to the wider scientific community as this research was entirely computational in nature and heavily relied on public dataset and openly shared scientific resources. Without the Open-Source Data Sharing initiative around the world, this research would not have been possible.

A handwritten signature in black ink, appearing to read 'Mayank', with a horizontal line drawn through it.

Candidate's Signature



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daultapur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, Mayank Singh hereby certify that the work presented in the thesis entitled "A Machine Learning Framework for the Detection of Alzheimer's Disease Using Gut Fungal ITS Sequencing Data", submitted in partial fulfilment of the requirements for the award of the Degree of Master of Technology, in the Department of Biotechnology, Delhi Technological University is an authentic record of my own work carried out during the period from January, 2026 to May, 2026 under the supervision of Prof. Yasha Hasija.

The work presented in the thesis has not been submitted by me for the award of any other degree at this or any other Institute.

A handwritten signature in black ink, which appears to read "Mayank Singh", is written over a horizontal line.

Candidate's Signature



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that Mayank Singh (24/BIO/01) has carried out the project work presented in this thesis entitled "A Machine Learning Framework for the Detection of Alzheimer's Disease Using Gut Fungal ITS Sequencing Data", from Department of Biotechnology, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

WM
29.05.26

Prof. Yasha Hasija
(Supervisor)

Head of Department

Department of Biotechnology

Delhi Technological University

Date: 28/5/26

ABSTRACT

Alzheimer's disease is a common neurodegenerative disease. The microbiota-gut-brain axis has emerged as a paradigm in the understanding of AD, and attention has been focused mainly on the bacterial microbiome, whereas the immunologically potent gut mycobiome (fungal community) has been largely overlooked. Most research on the mycobiome associated with AD has been descriptive with little predictive clinical application. To overcome this methodological limitation, this study aimed to create a supervised machine learning model to distinguish between AD patients and cognitively normal controls based on non-invasive gut fungal Internal Transcribed Spacer (ITS) sequencing data.

A public dataset of 120 individuals (60 AD patients and 60 healthy controls) was used for the research. Raw ITS amplicon sequencing data was processed in a standardized, bioinformatics pipeline using QIIME 2 and the DADA2 algorithm for high resolution Amplicon Sequence Variant (ASV) generation and the UNITE database for taxonomic assignment. Four machine learning models: Logistic Regression, Support Vector Machine, Histogram-based Gradient Boosting and Random Forest were optimized and evaluated using a rarefied, high-dimensional ASV feature matrix.

The results showed that the Random Forest classifier had the highest classification accuracy (0.75) and ROC-AUC (0.85) on an independent test set. Linear models on the other hand, did not work very well, showing the complexity and the non-linearity of the microbial interactions. The Random Forest model was shown to be stable by rigorous 5-fold cross validation. In addition, the feature importance analysis showed that certain genera of fungi such as *Candida*, *Purpureocillium*, and *Oligophagozoma* were the most discriminative predictive biomarkers in AD. The strong predictive value of *Candida* is in line with the immunological mechanisms of the gut-brain axis and inflammatory pathogenesis. Finally, this study moves the mycobiome analysis from an ecological description to a predictive clinical informatic, making fungal profiles of the gut a new potential non-invasive diagnostic tool for Alzheimer's disease.

LIST OF PUBLICATION

IEEE Global Symposium on Emerging and Communication Technologies (GSEACT 2026)

Fwd: Acceptance Notification for Paper ID – 1485 – 1st International Conference on Global Symposium on Emerging and Communication Technologies (GSEACT 2026) (IEEE Conference Record No. 68539), Last Date: 14th April 2026 Inbox x

From: **Microsoft CMT** <noreply@msr-cmt.org>

Date: Wed, 8 Apr, 2026, 13:36

Subject: Acceptance Notification for Paper ID – 1485 – 1st International Conference on Global Symposium on Emerging and Communication Technologies (GSEACT 2026) (IEEE Conference Record No. 68539), Last Date: 14th April 2026

To: Yasha Hasija <yashahasija@dtu.ac.in>

Dear Yasha Hasija,

Delhi Technological University

Paper ID: 1485

Paper Title: Machine Learning-Based Classification of Alzheimer's Disease Using Gut Fungal ITS Sequencing Data

Conference Record Number: 68539

We are delighted to inform you that your paper has been accepted for presentation at the 1st International Conference on Global Symposium on Emerging and Communication Technologies (GSEACT 2026).

Conference Dates: 6/11/2026 – 6/12/2026

Mode: Fully Online

TABLE OF CONTENTS

ACKNOWLEDGEMENT	2
CANDIDATE’S DECLARATION.....	3
CERTIFICATE BY THE SUPERVISOR.....	4
ABSTRACT.....	5
LIST OF PUBLICATION	6
TABLE OF CONTENTS	7
LIST OF TABLES	10
LIST OF FIGURES	11
LIST OF ABBREVIATIONS.....	12
CHAPTER 1 – INTRODUCTION	14
1.1 Background of Alzheimer’s Disease	14
1.2 Current Diagnostic Techniques and Their Limitations	14
1.3 Gut-Brain Axis	16
1.4 Gut Mycobiome and Neurological Disorders.....	17
1.5 Machine Learning in Biomedical Research	19
1.6 Research Gap.....	20
1.7 Aim and Objectives	21
CHAPTER 2 - LITERATURE REVIEW	22
2.1 Introduction	22
2.2 Pathophysiological Hallmarks of Alzheimer's Disease	22
2.3 The Microbiota-Gut-Brain Axis in Neurodegeneration.....	24
2.4 Bacterial Microbiome in Alzheimer's Disease	25
2.5 The Emerging Role of Gut Fungi in Neurodegeneration	26
2.5.1 Composition and Immunological Role of Gut Fungi	26
2.5.2 Fungal Dysbiosis	27

2.5.3 The Brain Infection Hypothesis.....	28
2.6 Methodological Challenges in Mycobiome Profiling.....	28
2.6.1 Sequencing Strategies.....	28
2.6.2 Low Biomass and the Contamination Risk.....	29
2.7 Computational Pipelines: From Raw Amplicon Reads to Taxonomy	30
2.8 Machine Learning for High-Dimensional Omics Data	30
2.8.1 Logistic Regression and Support Vector Machines	31
2.8.2 Ensemble Methods: Random Forest and Gradient Boosting	32
2.9 Identification of the Research Gap	33
CHAPTER 3 - METHODOLOGY	34
3.1 Data Availability	34
3.2 Overall Workflow of the Study	34
3.3 Dataset Description	34
3.4 Computational Environment	35
3.5 Denoising and Taxonomic Profiling via QIIME-2	37
3.5.1 Import of Raw Data	37
3.5.2 Primer Trimming Using Cutadapt.....	37
3.5.3 Denoising Using DADA2.....	37
3.5.4 Taxonomic Assignment.....	38
3.5.5 Rarefaction and Feature Table Generation.....	39
3.6 Feature Matrix Construction.....	39
3.7 Machine Learning Models	39
3.7.1 Logistic Regression (LR).....	40
3.7.2 Support Vector Machine (SVM)	40
3.7.3 Histogram-Based Gradient Boosting (HGB).....	41
3.7.4 Random Forest (RF)	41
3.8 Model Training Strategy	42
3.8.1 Dataset Splitting	42
3.8.2 Cross-Validation	43
3.9 Model Evaluation Metrics	44

3.10 Feature Importance and Interpretability	45
CHAPTER 4 – RESULTS AND DISCUSSION	46
4.1 Introduction	46
4.2 Sequencing and Quality Control Results.....	46
4.2.1 Read Processing and Denoising.....	46
4.2.2 Rarefaction	47
4.3 Taxonomic Profiling Results	47
4.4 Machine Learning Classification Results	48
4.5 Confusion Matrix Analysis.....	49
4.6 ROC Curve and Cross-Validation Stability Analysis	51
4.7 Feature Importance and Biological Interpretation	52
4.7.1 The Role of <i>Candida</i> in Neuroinflammation	53
4.7.2 <i>Purpureocillium</i> and <i>Oligophagozoma</i>	54
4.7.3 <i>Debaryomyces</i>	54
4.8 Comparative Discussion with Previous Studies	55
4.9 Clinical and Biological Implications	55
4.10 Conclusion.....	56
CHAPTER 5 - CONCLUSION.....	57
5.1 Summary	57
5.2 Major Findings	57
5.3 Biological Implications	59
5.4 Technical Limitations	60
5.5 Future Scope	60
REFERENCES	62
LIST OF PUBLICATION	68
PLAGIARISM VERIFICATION	69

LIST OF TABLES

Table No.	Title
Table 1.1	Symptoms and Stages of Alzheimer's disease
Table 1.2	Comparison of AD Diagnostic Techniques
Table 1.3	Major Fungal Genera Associated with Neurological Disorders
Table 3.1	Dataset Summary
Table 3.2	Software Configuration
Table 4.1	Classification Performance of Supervised Machine Learning Models

LIST OF FIGURES

Figure No.	Title
Fig 2.1	Pathology of Alzheimer's disease (amyloid plaques and tau tangles disrupting neuronal function).
Fig 2.2	Gut-brain axis communication
Fig 3.1	Overall Study Workflow
Fig 3.2	Random Forest architecture showing the bagging process and ensemble voting mechanism.
Fig 3.3	Machine Learning Pipeline.
Fig 4.1	Confusion matrices for Random Forest, Histogram-based Gradient Boosting, Logistic Regression, and Support Vector Machine classifiers evaluated on the test set.
Fig 4.2	Receiver Operating Characteristic (ROC) curve for the Random Forest classifier using 5-fold cross-validation.
Fig 4.3	Top fungal features contributing to Alzheimer's disease classification identified by Random Forest feature importance analysis.

LIST OF ABBREVIATIONS

AD:	Alzheimer's Disease
APP:	Amyloid Precursor Protein
ASV:	Amplicon Sequence Variant
AUC:	Area Under Curve
BBB:	Blood-Brain Barrier
CARD9:	Caspase Recruitment Domain Family Member 9
CNS:	Central Nervous System
CSF:	Cerebrospinal Fluid
CV:	Cross-Validation
DADA2:	Divisive Amplicon Denoising Algorithm
DAMPs:	Damage Associated Molecular Patterns
ENA:	European Nucleotide Archive
ENS:	Enteric Nervous System
GBA:	Gut-Brain Axis
HGB:	Histogram-Based Gradient Boosting
IBD:	Inflammatory Bowel Disease
IL-1 β :	Interleukin-1 β
IL-6:	Interleukin-6
IL-17:	Interleukin-17
IL-23:	Interleukin-23
ITS:	Internal Transcribed Spacer
LPS:	Lipopolysaccharides
LR:	Logistic Regression
MAMPs:	Microbial Associated Molecular Patterns
MCI:	Mild Cognitive Impairment
MDI:	Mean Decrease in Impurity

ML: Machine Learning
MRI: Magnetic Resonance Imaging
NFTs: Neurofibrillary Tangles
NGS: Next Generation Sequencing
OTU: Operational Taxonomic Unit
PET: Positron Emission Tomography
PRRs: Pattern Recognition Receptors
QIIME: Quantitative Insights Into Microbial Ecology
RF: Random Forest
ROC: Receiver Operating Characteristic
ROS: Reactive Oxygen Species
SCFAs: Short-Chain Fatty Acids
SH: Species Hypothesis
SVM: Support Vector Machine
Th17: T-helper 17
TNF- α : Tumour Necrosis Factor α
WMS: Whole Metagenome Shotgun Sequencing
WSL: Windows Subsystem for Linux

CHAPTER 1 – INTRODUCTION

1.1 Background of Alzheimer's Disease

Alzheimer's disease (AD) is the most common neurodegenerative disorder and is the primary cause of dementia globally in older adults (Knopman et al., 2021). Clinically, AD is characterized as a progressive loss of cognitive functions (Knopman et al., 2021; Zhang et al., 2022) that starts insidiously with the most pronounced short-term memory impairments, and then gradually worsens to significant language, spatial orientation, executive function, and behavioral control impairments. In its advanced stages, patients may completely lose their motor function and independence, making them completely reliant on care and assistance (Zheng & Wang, 2025).

Alzheimer's disease is characterized by two main abnormal protein deposits: amyloid-beta plaques in the extracellular space and neurofibrillary tangles (NFTs), which are aggregates of hyperphosphorylated tau protein, within the brain (Knopman et al., 2021). According to the amyloid cascade hypothesis, the accumulation of amyloid beta peptides, especially the very toxic amyloid beta-42 isoform, occurs because beta-secretase and gamma-secretase enzymes abnormally break down amyloid precursor protein (APP) into overabundant small fragments of amyloid beta (Selkoe & Hardy, 2016). These aggregates create insoluble plaques causing impaired communication between synapses and initiating localized inflammatory responses. At the same time, hyperphosphorylation of tau protein, which normally serves to maintain stability of microtubules in neurons - leads to disconnection and aggregation into NFTs within the neuronal soma. This disruption of the cytoskeleton ultimately causes axonal transport dysfunction, which in turn results in synaptic dysfunction, neuronal death, and macroscopic brain atrophy (Zheng & Wang, 2025).

The widespread burden of AD is becoming a crisis. Given the expected further growth in global life expectancy, the prevalence of AD is growing at an accelerating rate, posing an unprecedented threat to global healthcare systems and the global economy. It has an enormous economic burden, not just on the direct costs of medical treatment and drug use, but also on the massive indirect costs for informal care and lost productivity.

1.2 Current Diagnostic Techniques and Their Limitations

Although much has been learned about the pathogenesis of AD, early, accurate, and

readily available diagnosis is still a huge and challenging clinical hurdle. At present, the diagnosis of AD is largely dependent on neurocognitive testing and on advanced biomarker evaluations (Knopman et al., 2021).

Use of structural imaging, like Magnetic Resonance Imaging (MRI), is used to identify medial temporal lobe and hippocampal atrophy. MRI is particularly good at excluding other causes of dementia (such as tumors or stroke), but is generally only able to see structural changes when they are quite extensive and have already caused irreversible damage to the neurons. Positron Emission Tomography (PET) scans provide the more direct assessment of the AD pathology by using specific radiotracers to visualize the quantities of the amyloid and tau. PET is, however, very costly, requires a specialized radioactive facility and is not widely available in clinical practice, especially in resource-limited or rural settings (Knopman et al., 2021).

Table 1.1: Symptoms and Stages of Alzheimer’s Disease

Stage	Clinical Symptoms	Neuropathological Correlates
Mild (Early)	Short-term memory loss, difficulty finding words, mild confusion, impaired judgement.	Initial accumulation of amyloid beta plaques; entorhinal cortex and hippocampal involvement.
Moderate (Mid)	Disorientation, behavioral changes (agitation, paranoia), severe memory gaps, wandering.	Widespread tau tangles; neuroinflammation; expansion of atrophy to the cerebral cortex.
Severe (Late)	Complete loss of speech, inability to recognize, loss of basic motor functions (swallowing, walking).	Severe critical atrophy; profound neuronal loss across multiple brain regions.

The CSF biomarker analysis (amyloid beta-42, t-tau, and p-tau) has high diagnostic specificity and sensitivity. However, CSF collection is an invasive procedure with risks of infection and discomfort to the patient, and is not likely to be performed for routine screening or for repeated longitudinal monitoring. Recent advances in blood-based markers, such as plasma p-tau181 and p-tau217, hold great promise but these are still in the process of making their way into clinical practice, and are frequently only available with highly specialized assays (Zheng & Wang, 2025).

Thus, new, non-invasive and cost-effective diagnostic approaches are urgently needed to be discovered and validated. Peripheral biomarkers which could reflect central nervous system (CNS) pathology without the need of invasive interventions could revolutionize the early detection and management of Alzheimer's disease.

1.3 Gut-Brain Axis

The neurocentric approach of Alzheimer's disease has been challenged in recent years by the new paradigm of microbiota-gut-brain axis (GBA) (Forbes et al., 2019). The GBA is a complex bidirectional communication system between the enteric nervous system (ENS) and gastrointestinal tract and the central nervous system (CNS) (Cryan et al., 2019). This complex crosstalk occurs via three main mechanisms: neural, immunologic and metabolic signaling.

Neural Signaling: The vagus nerve is the main anatomical pathway from the gut to the brain. The gut microflora can activate the vagus nerve and regulate its communication to the brainstem influencing higher cognitive and emotional centers (Cryan et al., 2019).

Immune Signaling: The gut microbiome is essential for modulating and programming local and systemic immunity (Forbes et al., 2019). An imbalance in the microbial community structure, called dysbiosis which can result in the disruption of the structural integrity of the intestinal epithelial barrier and the occurrence of a phenomenon referred to as "leaky gut" (Yuan et al., 2023). This increased permeability permits translocation of microbial components into the systemic circulation, like lipopolysaccharides (LPS) from bacteria and polysaccharides of the fungal cell walls. These antigens, when present in the bloodstream, cause a systemic inflammatory response that leads to the production of pro-inflammatory cytokines (such as IL-1 β , IL-6 and TNF- α). These cytokines may cross the blood-brain barrier (BBB) or influence its permeability which in turn can activate resident microglia and astrocytes in the brain (Rutsch et al., 2020). Chronic neuroinflammation is directly associated with prolonged microglial activation, which in turn prevents the microglia from clearing amyloid plaques and worsens tau pathologies in AD (Rutsch et al., 2020).

Metabolic Signaling: Gut microbes generate a lot of metabolites that are biochemically active, such as neurotransmitter precursors, secondary bile acids and short-chain fatty acids (SCFAs). The SCFAs, particularly butyrate, have strong anti-inflammatory activity and play an essential role in the maintenance of BBB and intestinal barrier integrity (Cryan et al., 2019). In Alzheimer's disease, levels of beneficial microbes (SCFAs) are often found to be reduced, which cause systemic inflammation and neurodegeneration (Liang et al., 2024).

Table 1.2: Comparison of AD Diagnostic Techniques

Diagnostic Technique	Advantages	Limitations
Magnetic Resonance Imaging (MRI)	High-resolution structural imaging; rules out non-AD pathologies.	Expensive; detects late-stage structural atrophy rather than early molecular changes.
Positron Emission Tomography (PET)	Direct visualization of amyloid and tau burdens; early pathological detection.	Extremely costly; involves radioactive exposure; limited geographic accessibility.
Cerebrospinal Fluid (CSF) Analysis	High sensitivity and specificity for core AD biomarkers (amyloid beta, tau).	Highly invasive (lumbar puncture); risk of patient discomfort and complications; difficult to repeat.
Microbiome/Mycobiome Profiling (Proposed)	Completely non-invasive (fecal sample); reflects systemic immune/metabolic changes.	Still in investigational stages; complex data analysis required; lacks clinical standardization.

1.4 Gut Mycobiome and Neurological Disorders

Despite the extensive study of the bacterial component of the gut microbiome (the bacteriome), the fungal component of the gut microbiome (the mycobiome) is relatively unexplored (Forbes et al., 2019). Although it is a much smaller proportion, fungi present about 0.1% of the total microbial biomass in the human gastrointestinal tract, which is quite small compared to the bacterial biomass. However, because fungal cells are physically larger than bacterial cells, their impact, metabolic output and ability to alter host immunity are far disproportionate (Chin et al., 2020).

As to the mycobiome, it is very dynamic, with *Candida*, *Saccharomyces* and *Malassezia*, as well as other resident genera, being the most common, and the phyla that dominate the ecosystem of the mycobiome are mainly Ascomycota and Basidiomycota (Yetgin, 2024).

Fungi have a more complex cell wall structure than bacteria, which is unique and rich in potent immunogenic polysaccharides, such as beta-glucans, chitin and mannans (Forbes et al., 2019). These molecules are recognised by pattern recognition receptors (PRRs) on host innate immune cells, such as Dectin-1 receptor. Fungal beta-glucans activate the production of pro-inflammatory cytokines

through binding to Dectin-1, and stimulate T-helper 17 (Th17) cell responses (Yuan et al., 2023) (Shao et al., 2025).

On the other hand, if the gut microbial imbalance (dysbiosis) occurs and the opportunistic pathobionts (like *Candida* species) overgrow, these immune pathways can be hyper-activated. This continuous immune activation causes disruption of the intestinal barrier, leading fungal antigens and inflammatory mediators to enter the systemic circulation, ultimately crossing the BBB and triggering neuroinflammation (Yetgin, 2024).

There has been recent evidence that explicitly connects fungal dysbiosis with Alzheimer's disease. Internal transcribed spacer (ITS) based sequencing studies have identified distinct fungal signatures in the feces of patients with AD. For example, *Candida tropicalis* and *Schizophyllum commune* were enriched in AD patients, whereas beneficial fungal diversity was decreased, as shown by Ling et al. (2021). Likewise, increased prevalence of *Botrytis* and *Cladosporium*, the two dominant fungi, was observed in those with Mild Cognitive Impairment (MCI) a prodromal phase of AD (Nagpal et al., 2020), which significantly correlated with an increase in tau biomarkers in CSF (Nagpal et al., 2020) (Kumar et al., 2025).

In addition, studies that are currently being discussed have employed nested-PCR and immunohistochemistry for the detection of fungal DNA (*candida* and *malassezia*) and physical fungal hyphae directly in the brain parenchyma and blood vessels of post-mortem AD patients (Pisa et al., 2015). Although direct brain infection has been a big debate as there is the risk of contamination of the environment during autopsies, the strong peripheral effect of gut mycobiome on systemic inflammation and neurodegeneration has become undeniable.

Table 1.3: Major Fungal Genera Associated with Neurological Disorders

Fungal Genus	Relevance to Host Physiology and Disease	Reported Neurological Association
<i>Candida</i>	Opportunistic pathobiont; strongly activates Th17 immune pathways via cell wall (beta-glucans); capable of driving systemic inflammation	Enriched in AD; positively correlated with CSF tau levels.
<i>Malassezia</i>	Lipid-dependent fungi associated with skin conditions; can provoke systemic antifungal immune responses.	Frequently observed in PD; DNA fragments detected in post mortem brain.
<i>Saccharomyces</i>	Commensal yeast; produces anti-inflammatory metabolites; helps maintain fungal-bacterial homeostasis.	Often depleted in neurodegenerative diseases; loss compromises immune tolerance.
<i>Cladosporium</i>	Environmental fungus transiently present in the gut; associated with respiratory allergies and localized inflammation.	Enriched in Mild Cognitive Impairment (MCI); correlates with early tau pathology.

1.5 Machine Learning in Biomedical Research

Biomedical research has become a very data-intensive field due to new technologies called next generation sequencing (NGS). Sequencing of the microbiome and mycobiome leads to large, high dimensional datasets called omics data. However, these datasets typically used by traditional univariate statistical analyses, which are often insufficient to fully describe the intricate, non-linear, and highly intertwined relationships among thousands of various microbial taxa (Basgaran et al., 2025).

As a result, increasing attention is being paid to Machine Learning (ML) to overcome the "curse of dimensionality". ML includes several different types of computational algorithms that can learn intricate patterns from a set of data and subsequently make a very precise prediction or classification. In supervised learning, which is one of the types of ML, models are trained using labeled data (such as fungal abundances vs. disease status from "AD patients" and "Healthy Controls") to map input features to the output label (disease status) (Velasquez & Oudah, 2023).

Algorithms like Logistic Regression, Support Vector Machines (SVM), Random

Forests, and Gradient Boosting (Breiman, 2001; Cortes & Vapnik, 1995; Dreiseitl & Ohno-Machado, 2002; Ke et al., 2017) are highly effective in the disease prediction of microbiome.

Bioinformatics has unique needs for ensemble techniques, which are often very robust to noise, can naturally model complex non-linear relationship and can provide “feature importance” scores. Furthermore, feature importance enables researchers to gain insights into the biology by uncovering exactly which microbial taxa are influencing the model's decisions (Basgaran et al., 2025).

1.6 Research Gap

While the microbiota-gut-brain axis has been recognized as a new paradigm in neurodegeneration, the bulk of the current research literature is focused on the bacterial microbiome. Technical difficulties related to its low biomass and complex bioinformatics requirements are the main reason the gut mycobiome is largely ignored.

Moreover, there is only descriptive literature available on the gut mycobiome in Alzheimer's disease (Ling et al., 2021; Nagpal et al., 2020; Sait et al., 2021). These studies are typically restricted to the identification of fungi that are enriched or depleted (differential abundance) and not whether the fungal profile has any actual clinical value as a predictive diagnostic tool.

Compared to bacterial 16S rRNA sequencing data, the application of ML models for supervised classification of fungal Internal Transcribed Spacer (ITS) amplicon sequencing data is a methodological gap which has been unexplored. Currently, there is no computational framework to assess the potential of gut fungal community profile to diagnose Alzheimer's disease computationally. To fill this gap, this thesis is focused on the development and assessment of a supervised machine learning approach to classify Alzheimer's disease cases and cognitively normal controls using non-invasive gut fungal ITS sequencing data.

1.7 Aim and Objectives

Aim

To develop and evaluate a machine learning framework for the classification of Alzheimer's disease using non-invasive gut fungal ITS sequencing data.

Objectives

1. To process raw, paired-end ITS amplicon sequencing data from public repositories using a standardized bioinformatics pipeline (QIIME 2).
2. To remove chimeras, denoise sequences and create high resolution Amplicon Sequence Variant (ASV) feature tables using DADA2 algorithm.
3. To perform precise taxonomic identification of the fungal ASVs using the UNITE fungal reference database.
4. To train, optimize and cross-validate multiple supervised machine learning models (Logistic Regression, Support Vector Machine, Random Forest, Histogram-based Gradient Boosting) for binary classification.
5. To compare classifiers systematically and comprehensively, using a range of evaluation metrics: Accuracy, Precision, Recall, F1-score and ROC-AUC.
6. To gain the biological interpretability, by performing feature importance analyses to identify the most important fungal taxonomy to predict diseases.

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction

Alzheimer's disease (AD) has been classically understood in a strictly neurocentric manner, which has emphasized the protein misfolding and neuronal degradation in the central nervous system (CNS) (Forbes et al., 2019). In the last decade, however, there has been a paradigm shift in the study of neurodegenerative diseases with the discovery of the microbiota-gut-brain axis (GBA) (Yetgin, 2024).

Although the gut microbial community (bacterial microbiome) has been well documented in the context of systemic inflammation and neurodegeneration, the fungal community (gut mycobiome) is a relatively uncharted territory. In this chapter, authors review the basic pathophysiology of AD, discuss rapidly accumulating evidence that gut mycobiome dysbiosis is associated with cognitive decline, and highlight methodological and computational advances, such as internal transcribed spacer (ITS) sequencing and supervised machine learning, that are necessary for moving mycobiome research from descriptive observation to predictive clinical application.

2.2 Pathophysiological Hallmarks of Alzheimer's Disease

The most prevalent type of dementia in the world is Alzheimer's disease: clinically defined as progressive, irreversible cognitive decline with loss of memory and independence (Knopman et al., 2021; Kumar et al., 2025). The principal neuropathological features of AD are extracellular deposition of amyloid-beta plaques and the intracellular accumulation of neurofibrillary tangles (NFTs) which are comprised of hyperphosphorylated tau proteins (Dominguez-Gortaire et al., 2025a; Zheng & Wang, 2025).

For a long time, the amyloid cascade hypothesis has been the major paradigm in the study of the etiology of AD. It proposes that excessive production of amyloid beta-42 (Selkoe & Hardy, 2016) from the amyloid precursor protein (APP), through abnormal proteolytic processing by beta-secretase and gamma-secretase, leads to the accumulation of insoluble amyloid beta peptides (amyloid beta-42) which are highly neurotoxic (Chen et al., 2017). The formation of these peptides into oligomers and the subsequent dense plaques disturb synaptic communication and start a cascade of local neurotoxicity. At the same time, hyperphosphorylation of tau, a critical protein that helps to bind to the microtubules, causes it to lose its ability to bind to the microtubules and self-assemble into insoluble tangles inside the neuron's soma. This

interference with intracellular transport inevitably results in synaptic starvation, marked neuronal apoptosis, and macroscopic brain atrophy (Knopman et al., 2021).

In addition to proteinopathy, chronic neuroinflammation is now known to be a key and self-propagating mechanism of AD pathogenesis rather than just a by-product. Microglia and astrocytes are resident innate immune cells in the CNS that have a complex dual role. In a normal brain, microglia are protective cells that constantly patrol the microenvironment and remove cell debris and early aggregations of amyloid. Chronic exposure to Amyloid beta plaques and damage associated molecular patterns (DAMPs), however, leads to chronic activation of microglia in a pro-inflammatory state (Heneka et al., 2015). In this state, they produce an ongoing release of pro-inflammatory cytokines (e.g., tumor necrosis factor α [TNF- α], interleukin-1 β [IL-1 β] and interleukin-6 [IL-6]) and reactive oxygen species (ROS) (Naik et al., 2025), which establishes a very neurotoxic environment that exacerbates tau hyperphosphorylation and resulting neuronal death (Dominguez-Gortaire et al., 2025b). An increased awareness of the factors that contribute to this chronic microglial activation has prompted researchers to seek other peripheral sites of inflammation, outside the brain, as the origin of these factors.

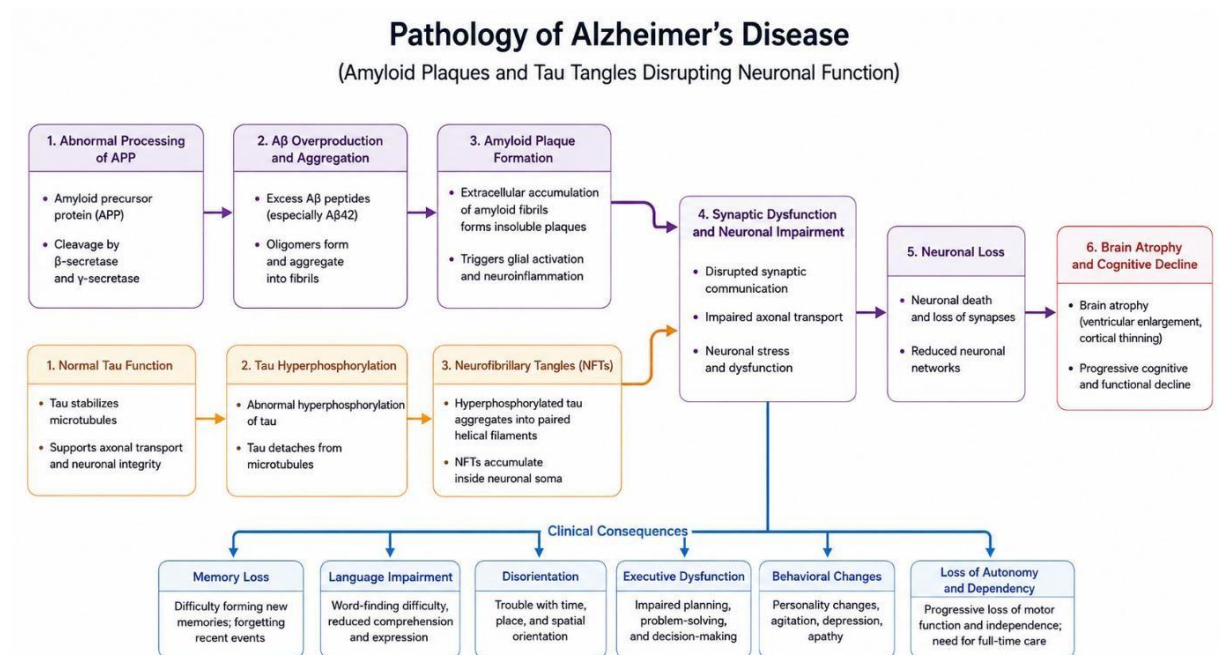


Fig 2.1: Pathology of Alzheimer's disease (amyloid plaques and tau tangles disrupting neuronal function).

2.3 The Microbiota-Gut-Brain Axis in Neurodegeneration

The microbiota-gut-brain axis (GBA) has broken the walls of the brain, which traditionally had been regarded as an immune-privileged organ isolated from the rest of the body. The GBA is an extremely intricate bidirectional communication system between the enteric nervous system (ENS) and the intestinal microenvironment on one hand, and the CNS on the other, physically and biochemically linked (Cryan et al., 2019). This crosstalk is regulated by three main pathways: neural, metabolic and immune signaling.

The neural pathway is most strongly associated with the vagus nerve, connecting the gut wall directly to the brainstem. Metabolically, the gut microbes living in the gut break down dietary fibers into a wide range of metabolites that are neuroactive and immunomodulatory, such as short-chain fatty acids (SCFAs) like butyrate, acetate and propionate (Rutsch et al., 2020). Butyrate is crucial for colonocyte energy supply, regulating the intestinal epithelial barrier, and having systemic anti-inflammatory properties (Van Hul et al., 2024).

Perhaps the most important in the context of neurodegeneration is the immune pathway. A healthy and diverse microbiome helps maintain barrier integrity of the intestinal epithelia and mucosal tolerance. But if gut dysbiosis sets in, with the decrease in microbial diversity and growth of pathogenic microbes, this balance is thrown off. The dysbiosis results in a lower concentration of the barrier fortifying SCFAs and leads to a higher intestinal permeability (also known as “leaky gut”) (Yuan et al., 2023). This leaky barrier allows microbial associated molecular patterns (MAMPs), including bacterial lipopolysaccharides (LPS) and fungal beta-glucans to enter the systemic circulation (Heneka et al., 2015). These antigens when they get into the blood stream, trigger a strong systemic inflammatory response. This leads to peripheral cytokines that cross the BBB or change its permeability, which in turn cause chronic microglial activation that worsens AD pathology (Heneka et al., 2015; Rutsch et al., 2020).

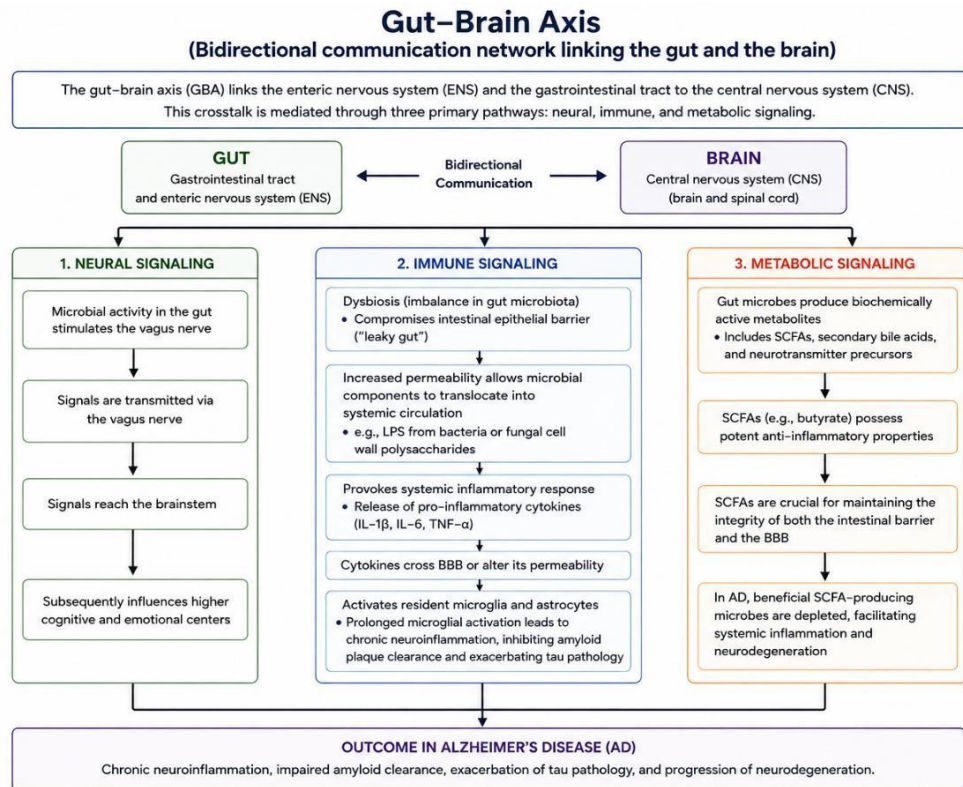


Fig 2.2: Gut-brain axis communication

2.4 Bacterial Microbiome in Alzheimer's Disease

Most studies examining the GBA in Alzheimer's Disease have only examined the bacterial microbiome (bacteriome). The taxonomic differences in the gut bacteriome between AD patients and cognitively normal individuals, repeatedly observed in both cross-sectional and longitudinal studies, have been identified.

One of the earliest and most clear-cut demonstrations of this phenomenon was provided by (Vogt et al., 2017) who showed that the diversity of microbes in AD patients was reduced, with a significant decrease in the phylum Firmicutes and an increase in Bacteroidetes.

Specifically, there is a consistent depletion of families of beneficial bacteria that produce short-chain fatty acids (SCFAs), including *Faecalibacterium* and *Roseburia* genera of *Lachnospiraceae* and *Ruminococcaceae* families (Liang et al., 2024). In contrast, *Escherichia* and *Shigella*, which are strong producers of pro-inflammatory LPS are overrepresented (Vogt et al., 2017).

Not only are these bacterial changes a late-stage result of dementia, but they seem to be a part of the very process of the disease in the pre-clinical stages. Recently, (Ferreiro et al., 2023; Jung et al., 2022) independently discovered that gut

microbiome composition is already significantly altered in preclinical stages of AD and is tightly linked with early high level of amyloid-beta and tau biomarkers, well before the onset of cognitive symptoms. This indicates that dysbiosis of gut flora is an early upstream process that may contribute to the early steps in neuroinflammation, and thus the microbiome is a good primary diagnostic target for new biomarkers.

2.5 The Emerging Role of Gut Fungi in Neurodegeneration

Although there is a lot of information about the bacteria that inhabit the gut, the fungal flora of the human gastrointestinal tract has largely been ignored (Cesaro et al., 2025). The gut mycobiome makes up a fraction of the total microbial biomass, but its biological effect is enormous in relation to its relative abundance (Chin et al., 2020; Forbes et al., 2019).

2.5.1 Composition and Immunological Role of Gut Fungi

The human gut mycobiome is a dynamic ecosystem, with a much lower diversity than the human gut bacteriome, and with high inter-individual variability which is linked to diet, environment, and host genetics (Zhang et al., 2022). It consists mainly of Ascomycota and Basidiomycota, the most commonly detected resident genera are *Candida*, *Saccharomyces* and *Malassezia* (Tiew et al., 2020; Yetgin, 2024).

There are cell wall structures in fungi which are highly immunogenic and are not found in bacteria. The fungal cell wall contains complex polysaccharides such as beta-glucans, chitin and mannans (Forbes et al., 2019; Richard & Sokol, 2019). They are potent MAMPs that are sensed by certain pattern recognition receptors (PRRs) found on host innate immune cells. Most importantly, is Dectin-1, a C-type lectin receptor that is mostly found on dendritic cells and macrophages (Gaspar et al., 2025).

When fungal beta-glucans bind to Dectin-1, it triggers a significant intracellular signaling pathway involving CARD9 (Caspase Recruitment Domain Family Member 9) (Liu et al., 2022). The CARD9 pathway directly activates the secretion of pro-inflammatory cytokines

(e.g., IL-6, IL-23, and IL-1 β) and is the main pathway for T-helper 17 (Th17) cell differentiation (Gaspar et al., 2025; Liu et al., 2022). The overactivation of the Th17 response due to overgrowth of fungal pathobiont in the gut may result in dramatic, chronic systemic inflammation, while a tightly controlled Th17 response is essential for defense against fungal mucosal infections. In addition, there is significant cross-talk between fungi and bacteria, with the loss of competitive commensal bacteria (such as result of antibiotic use or by age-related dysbiosis) often allowing the opportunistic growth of fungi and thereby disrupting inter-kingdom homeostasis and critically weakening the intestinal epithelial barrier (Zhang et al., 2022).

2.5.2 Fungal Dysbiosis

The study of the precise role of the gut mycobiome in AD is still in its infancy, but the currently available evidence suggests that there is an important connection. A study that was very impactful in the literature was done by Ling et al. (2021) where they were comparing the faecal fungal profile of Chinese patients suffering from AD and healthy controls. They found a fungal dysbiosis in the AD cohort, with a significant enrichment of opportunistic pathobionts *Candida tropicalis* and *Schizophyllum commune*. Other commensal fungi, however, such as *Rhodotorula mucilaginosa*, were depleted. Interestingly, these mycobiome shifts were linked with elevated systemic inflammatory markers, thus offering the first human proof for the potential role of gut fungal dysbiosis as an active participant in systemic immune dysregulation seen in AD (Ling et al., 2021).

Another example of the importance of fungi in the early stages of disease comes from (Nagpal et al., 2020), who did a pilot study of gut mycobiome in people with Mild Cognitive Impairment (MCI), which often becomes full AD. The authors discovered a unique mycobiome signature for patients with MCI, with higher levels of *Botrytis*, *Kazachstania*, and *Cladosporium*. Additionally, these fungi also showed strong positive correlations with the CSF levels of central proteinopathy markers, like total tau and phosphorylated tau, in these patients, indicating a direct relationship between these gut fungi and early central proteinopathy (Hadrach et al., 2024; Nagpal et al., 2020).

The ways in which fungi may affect neurodegeneration go beyond the systemic inflammatory response. (Otoo et al., 2008) reported that the amyloidogenic proteins on the cell surfaces (such as Als adhesins) of certain fungal species are highly conserved. Amyloid cross-seeding has been found in in vitro experiments, in which amyloidogenic proteins from one species can also nucleate and promote the aggregation of amyloid proteins from another species (Subedi et al., 2022). This leads into the intriguing possibility that systemic exposure to fungal amyloids that

leak through this leaky gut might cross-seed the host Amyloid beta or alpha-synuclein, thus initiating or promoting brain protein aggregation.

2.5.3 The Brain Infection Hypothesis

Another, controversial, theory on the pathogenesis of fungal involvement in the CNS proposes that fungi can pass the GBA and actually invade the CNS. Using advanced immunohistochemistry and nested PCR, directly visualized fungal cells and hyphae in the neurons and blood vessels of the post-mortem brain tissue from AD patients (Phuna & Madhavan, 2022; Pisa et al., 2015).

(Alonso et al., 2018) conducted follow-up studies that revealed the presence of multiple fungal DNA in the AD brain such as *Candida*, *Malassezia*, *Cladosporium* and *Saccharomyces*, which are located with bacterial antigens (Agarwal et al., 2026). They suggested a "polymicrobial pathogenesis" model where continuous low-grade infection with bacteria in the brain causes ongoing neuroinflammation which contributes to the pathology of AD.

But these results have stirred up much controversy among scientists. The brain is a low biomass environment. The molecular detection of ubiquitous environmental and human commensal fungi (such as *Malassezia* spp. and *Candida* spp.) in post mortem tissues is easily affected by contamination during autopsies, tissue processing, and from laboratory reagents (kitome) (Eisenhofer et al., 2019; Ko et al., 2024). Although the role of fungal invaders within the brain is highly controversial, the effects of the gut mycobiome on the immune system and gut brain axis is much more well-established and biologically plausible in models of disease pathogenesis.

2.6 Methodological Challenges in Mycobiome Profiling

There are many methodological hurdles to adopt mycobiome as a reliable biomarker for transitioning from theoretical associations to using it in practice.

2.6.1 Sequencing Strategies: Trade-offs Between ITS Amplicons and WMS

Most fungal profiling is done with targeted amplicon sequencing. While the 16S rRNA gene is the usual target of bacterial studies, mycobiome studies focus on the Internal Transcribed Spacer (ITS) region of the nuclear ribosomal RNA, and has been officially designated as the DNA barcode marker for fungi (Schoch et al.,

2012). A high degree of sequence variability in ITS region, between 18S, 5.8S and 28S rRNA gene, allows for better taxonomic resolution, even up to species level (Schoch et al., 2012).

The ITS region is split into two parts as ITS1 and ITS2. The selection of the subregion that is amplified can significantly affect the biological inferences drawn (Huseyin et al., 2017). The ITS1 sequence is extremely variable in length and sequence conservation, and so standard PCR can selectively amplify shorter fragments of the ITS1, leading to a bias in the data towards the taxa with shorter insertions. ITS2, on the other hand, is a more conserved sequence and more generally suited for broad taxonomic classification, producing more stable global alignments (Nilsson et al., 2019).

Whole Metagenome Shotgun Sequencing (WMS) overcomes the impact of primer bias by sequencing all DNA in a sample. However, WMS is essentially not feasible for the normal studies of the human gut mycobiome because of the "0.1 percent problem" (Donovan et al., 2018). Since fungi only make up a very small fraction of the gut biomass, the bacterial and host signals completely outnumber the fungal signal. Despite extremely high and costly sequencing, WMS often fails to capture sufficient numbers of reads to generate accurate genomes and identify low-abundance taxa (Donovan et al., 2018; Tedersoo et al., 2018).

However, targeted ITS sequencing is still the gold standard, albeit not without its drawbacks, in the absence of this.

2.6.2 Low Biomass and the Contamination Risk

Environmental and reagent contamination are the most insidious problem that exists in mycobiome research, and especially studies of low biomass environments such as cerebrospinal fluid or brain tissue. Microbial DNA has been found everywhere in the laboratory environment, in DNA extraction kits, and in PCR water (Salter et al., 2014). However, this "kitome" signal is swamped in high-biomass bacterial samples. But when amplifying the faintest hints of fungal DNA found in the gut, or for the most subtle fungal signals in the brain, these contaminants can easily dominate the true signal and researchers can be led to mistakenly conclude that a disease is represented by environmental molds.

To overcome this, modern bioinformatics pipelines require rigorous decontamination steps. These include statistical methods like those from the Decontam package that are able to detect and subtract features that are more abundant in negative extraction controls, or features whose relative abundance decreases with decreasing DNA concentration, compared to true biological samples

(Davis et al., 2018; Eisenhofer et al., 2019). The strict compliance of such controls is a prerequisite to guarantee the validity of any machine learning model created using the mycobiome data.

2.7 Computational Pipelines: From Raw Amplicon Reads to Taxonomy

There is a lot of processing power needed to convert raw sequencing reads into quantitative biological information. The microbiome platform QIIME 2 (Quantitative Insights Into Microbial Ecology), which is open source, reproducible and highly modular, for microbiome data science used for this project (Bolyen et al., 2019).

The sequences of amplicon were traditionally partitioned into Operational Taxonomic Units (OTUs) using an arbitrary genetic similarity limit (usually 97%). This method masked species level differences and frequently resulted in a large and heterogeneous group of organisms. Today, clustering using OTU is completely superseded by exact sequence inference based on algorithms such as DADA2 (Callahan et al., 2016). DADA2 creates a rich, high-probabilistic error model directly from the quality score of the particular sequencing run. It provides an explicit discrimination between true biological sequences and sequencing errors as well as highly precise Amplicon Sequence Variants (ASVs). In addition, DADA2 is able to accurately identify and remove chimeric sequences - PCR artifacts resulting from two different DNA templates being incorrectly joined while being amplified (Callahan et al., 2016).

After the ASV generation, sequences need to be taxonomically annotated. This is done through a Naive Bayes classifier that was pretrained with fungal ITS data from the UNITE database (Nilsson et al., 2019). UNITE is a curated reference database to solve an important problem: dark taxa - fungi that have been sequenced but not cultured and therefore not formally named in Linnaean taxonomy. The UNITE groups sequences into stable Species Hypotheses (SHs) with high genetic similarity and assigns identifiers that enables tracking of specific, unculturable fungal lineages across different studies worldwide (Kõljalg et al., 2013).

2.8 Machine Learning for High-Dimensional Omics Data

The high throughput ITS sequencing data is extremely high dimensional, with thousands of sparse features (ASVs) across a relatively small number of human

samples. Traditional univariate statistical methods (e.g., T-tests, ANOVA or simple correlational methods) are poorly designed to analyze this kind of data (Mohammadi & Karwowski, 2024). They do not reflect the multi-variate, non-linear and complex interaction between different organisms within the gut microbiome (Dominguez-Gortaire et al., 2025b).

In order to harness this data for predictive clinical use, supervised Machine Learning (ML) has become the popular choice of researchers. Supervised ML in the context of microbiome studies aims to train an algorithm to mathematically map a high-dimensional feature vector (relative abundance of fungal ASVs) to a specific diagnostic output label (e.g., Alzheimer's disease versus Healthy Control). ML can move microbiome research beyond merely describing observed patterns to predictive diagnostic screening by uncovering deeply buried multivariate patterns (Basgaran et al., 2025; Velasquez & Oudah, 2023). Multiple algorithmic architectures need to be assessed in order to ensure strong classification.

2.8.1 Logistic Regression and Support Vector Machines

Logistic Regression (LR) is the commonly used linear baseline model in biomedical classification problems (Dreiseitl & Ohno-Machado, 2002). LR learns the probability of a data sample belonging to a specific class by fitting a linear equation to the data and using a sigmoid function to confine the output to the range between zero and one. LR is easy to interpret and computationally simple, but it makes the strict assumption of linear relationships between microbial features and disease state, which can often be violated in complex biological systems.

The more powerful mathematically correct method of classification is Support Vector Machines (SVM) (Cortes & Vapnik, 1995). The main idea behind an SVM is to find the hyperplane in N-dimensional space that separates the data points of different classes as much as possible and maximizes the distance between them. The real strength of SVM is the use of the kernel trick. An SVM can be used to implicitly transform non-linearly separable data into a higher dimensional feature space where classes can be linearly separated. Although efficient, SVMs are susceptible to the scaling of the data and have the need to tune the hyperparameters well or they will overfit on datasets with many features and a small sample size ($p \gg n$).

2.8.2 Ensemble Methods: Random Forest and Gradient Boosting

A number of ensemble learning architectures that integrate the predictions of several models have been shown to be the most capable architectures to process microbiome omics data (Nandal et al., 2025). Random Forest (RF) is one of the most important ensemble algorithms made up of a large number of decision trees (Breiman, 2001). In order to achieve diversity among the trees, RF uses a technique called bootstrap aggregation (bagging), which involves growing trees on a random subset of the data. In addition, for each split in a tree, the algorithm only chooses a random subset of the features to consider. The final classification is made by the majority of all trees (Breiman, 2001). This stochastic construction significantly decreases the model variance and makes RF much less sensitive to overfitting, noise, and more able to naturally model the complex and non-linear microbial interactions without explicit feature engineering, as explained by (Breiman, 2001).

The other alternative ensemble method in this class is the Gradient Boosting method based on the histogram (HGB) (Ke et al., 2017). HGB builds "weak" decision trees sequentially, as opposed to RF that constructs independent trees concurrently. The individual trees are trained to compute and rectify the residual errors left by the ensemble of all previous trees, and optimize a loss function using gradient descent. In modern implementations, the data from the features are grouped into discrete bins to significantly speed up computation time and lower memory consumption, this is the basis of the histogram method. Even though gradient boosting can be used to produce very high accuracy predictive models, it is also more likely to be prone to overfitting than the bagging approach used in Random Forest in small biomedical datasets (Ke et al., 2017).

Crucially, tree-based ensemble algorithms, such as RF and HGB, enable the retrieval of "Feature Importance". Mean decrease in node impurity (e.g., Gini importance) can be computed over all trees to rank the influence of particular fungal ASVs on the predictive accuracy of the model in a quantitative manner. This fills the gap between mathematical 'black-box' predictions and biologically interpretable biomarker discovery.

2.9 Identification of the Research Gap

The existing literature suggests a number of important knowledge gaps regarding the microbiota-gut-brain axis in the context of AD. Firstly, the bacterial microbiome is extensively studied, and the immunologically potent gut mycobiome is vastly under-studied. Although present in low numbers, fungi have unique components of their cell wall (beta-glucans) that are highly effective at triggering inflammation and microglial activation in the central nervous system, the pathways involved in the pathogenesis of AD.

Second, existing research that has examined the gut mycobiome and cognitive decline has a descriptive nature (Ling et al., 2021; Nagpal et al., 2020). They note the evolution of fungal taxonomy and make correlations with tau or inflammatory markers, but they do not consider the potential for the high-dimensional fungal signatures in any specific clinical value as predictive diagnostic tools.

Lastly, there is a large gap in computational biology methodology. In the field of classification of neurodegenerative diseases using bacterial 16S rRNA data, advanced machine learning frameworks have shown promise and been successfully applied (Basgaran et al., 2025; Velasquez & Oudah, 2023), but this novel task of classifying Alzheimer's disease based on high-resolution, fungal ITS Amplicon Sequence Variant (ASV) data is entirely unexplored.

It is urgently needed to create a standardized computational pipeline combining a high-quality sequence processing implemented in QIIME 2 with powerful cross-validated ensemble machine learning models. Filling this void will move mycobiome study from descriptive ecological studies to medical informatics, an innovative, non-invasive computational approach that will allow to diagnose Alzheimer's disease and select very specific fungal markers with a role in the process of neurodegeneration.

CHAPTER 3 - METHODOLOGY

3.1 Data Availability

This chapter outlines the extensive computational tool that was designed to categorize Alzheimer's disease (AD) from the data generated by gut fungal Internal Transcribed Spacer (ITS) sequencing. This method integrates two computational fields: high-throughput microbiome bioinformatics and supervised machine learning.

This work is conducted in silico using public, anonymous, genomic data. Raw paired-end ITS sequencing data used in this study were previously generated by Ling et al. (2021) and obtained from the European Nucleotide Archive (ENA) (PRJNA678832). As these data are publicly accessible and de-identified, this is a computational study that does not require primary institutional ethical board approval (Ling et al., 2021).

3.2 Overall Workflow of the Study

The end-to-end analytical pipeline developed for this thesis spans from raw, unstructured sequence reads to a machine learning classification model that is biologically interpretable. The workflow is divided into two main phases: Phase 1 (Bioinformatics Processing) and Phase 2 (Machine Learning Modeling).

The Phase 1 consists of importing raw FASTQ data into an Ubuntu (WSL) environment, mapping the data using a manifest file, quality control of the data, adapter trimming, denoising with DADA2, taxonomy assignment based on the UNITE database, and rarefaction to construct an Amplicon Sequence Variant (ASV) feature table. This high dimensional ASV matrix is used for training, optimizing and evaluating four different machine learning classifiers and is used to generate feature importance scores which are used to identify fungal biomarkers for prediction.

3.3 Dataset Description

The data used for this study is one of the largest human gut mycobiome cohorts available to date with regards to Alzheimer's disease. The data is from fecal sample sequencing from a Chinese cohort of 120 individuals (60 patients with clinically diagnosed Alzheimer's disease and 60 cognitively normal healthy controls) (Ling et al., 2021).

The original study excluded participants who had received systemic or antibiotic therapy in the past month prior to sampling, and who were not taking any medication to treat gastrointestinal disorders, to minimize confounding variables that could significantly impact the gastrointestinal mycobiota. In addition, patients with major gastrointestinal disorders were excluded (e.g., Inflammatory Bowel Disease (IBD)). The extracted DNA was amplified targeting the highly variable ITS2 region of the fungal ribosomal RNA gene, and then sequenced on an Illumina MiSeq platform with paired-end sequencing (Ling et al., 2021).

Table 3.1: Dataset Summary

Parameter	Value / Description
Data Repository	European Nucleotide Archive (ENA)
Project Accession	PRJNA678832
Total Cohort Size	120
Alzheimer's Disease (AD) Samples	60
Cognitively Normal Control Samples	60
Sequencing Target	Fungal Internal Transcribed Spacer 2 (ITS2)
Sequencing Platform	Illumina MiSeq

3.4 Computational Environment

Reproducibility is a main principle of contemporary computational biology. A standardized virtualized environment was introduced to guarantee data processing and machine learning pipelines were completely reproducible. All bioinformatics processing was performed with the use of QIIME 2 platform (Bolyen et al., 2019) and predictive modeling was performed in python.

Table 3.2: Software Configuration

Component	Specification / Version
Operating System	Ubuntu 22.04 LTS (via Windows Subsystem for Linux - WSL2)
Bioinformatics Framework	QIIME 2 (version 2025.10)
Programming Language	Python 3.10.11
Machine Learning Library	scikit-learn
Data Manipulation Libraries	Pandas, NumPy



Fig 3.1: Overall Study Workflow

3.5 Denoising and Taxonomic Profiling via QIIME-2

All raw fungal ITS reads were processed within the QIIME 2 ecosystem (Bolyen et al., 2019). QIIME 2 utilizes ‘.qza’ (artifact) files to encapsulate data, metadata, and provenance (a detailed log of every command and software version), ensuring strict data integrity.

3.5.1 Import of Raw Data

Raw paired-end FASTQ files (forward and reverse reads for all 120 samples) were directly downloaded to the Ubuntu environment. To import the data, a manifest.tsv file was prepared to make an explicit link between each sample ID and the forward and reverse FASTQ files, and to prevent any sample mixing. The data was loaded into the program in order to create the initial artifact (paired-end-demux.qza). To make informed decisions for downstream trimming and denoising, a demultiplexing summary visualization (demux.qzv) was generated. This essential visualization revealed the number of reads per sample and also the distribution of the quality scores by base position, which is essential to avoid blind quality estimation.

3.5.2 Primer Trimming Using Cutadapt

Before sequence denoising, it is crucial to eliminate PCR primer sequences. These sequences can be retained, causing false ASVs, incorrect taxonomic assignments, and artificially high diversity measures. The q2-cutadapt plugin was run on paired-end-demux.qza to search and cut primer sequences to create trimmed-seqs.qza.

3.5.3 Denoising Using DADA2

The most important step in the bioinformatics pipeline is the conversion of raw reads to Amplicon Sequence Variants (ASVs) using the divisive amplicon denoising algorithm DADA2 (Callahan et al., 2016)

One of the most critical issues in studying mycobiome is the great variation in length of fungal ITS, compared to the relatively short length of bacterial 16S rRNA amplicons.

Implementing strict uniform length truncation of ITS sequences may result in the exclusion of biologically short amplicons and affect the merging of paired-end sequences. Therefore, the q2-dada2 plugin was specifically designed to have the

truncation lengths set to zero to eliminate the heterogeneity of sequence length in the natural sequence.

The characteristic 3' sequence quality degradation of Illumina sequencing was overcome without any hard truncation by applying rigorous Expected Error (EE) filtering. The reads were filtered using maximum error limits of 2.0 for forward reads and a more relaxed forward error limit of 5.0 for the inherently noisier reverse reads.

After quality filtering, DADA2 drew up a parametric error model specific to the sequencing run, corrected erroneous sequences, linked paired-end sequences and chimera removal. The resulting output included a very curated ASV feature matrix (table.qza) and the representative sequences (rep-seqs.qza).

The trimmed-seqs.qza file was input to the q2-dada2 plugin, which created many important output files: The contents of this file include the essential feature matrix (Samples $\times$ ASV read counts):

- table.qza: The core feature matrix (Samples $\times$ ASV read counts).
- rep-seqs.qza: The error-corrected representative DNA sequences of all the ASVs.
- stats.qza: Denoising statistics, showing reads that were retained and lost.
- base-transition-stats.qza: Error model diagnostics that display the misread frequencies for bases.

The denoising statistics was plotted using stats.qzv to formally show the data quality and support any data read exclusions.

3.5.4 Taxonomic Assignment

These sequences are not named, but DADA2 is able to identify unique sequences. These anonymous ASVs are assigned to known fungal lineages in a taxonomic assignment map. This was accomplished with the use of a Naive Bayes machine learning based classifier, which was pre-trained on the UNITE database (Version 10.0). UNITE is the worldwide standard for fungal ITS identification (Nilsson et al., 2019). It is unique in tackling the problem of the unculturable and unnamed fungi using dynamic Species Hypotheses (SHs). Each experimental ASV is compared to the UNITE reference sequences in the classifier, which provides taxonomy at the highest confidence level available (Kingdom, Phylum, Class, Order, Family, Genus, Species). ASVs that were not confidently assigned to genus or species were also kept at the highest confident taxonomic level to avoid data loss.

3.5.5 Rarefaction and Feature Table Generation

One of the common problems in the high-throughput sequencing is the high variation in sequencing depth (total number of reads) in various samples. Overcorrection of the microbial diversity will be created by higher sequencing depth, as it will give the impression of greater diversity when left untreated.

The ASV feature table was rarefied using the rarefy function in the qiime feature-table package so as to normalize the dataset. Rarefaction is a type of subsampling that reduces all samples to a common, predetermined depth of sequencing, with no replacement of any samples. This leads to excess reads being thrown away from deeply sequenced samples, but guarantees that each sample is statistically equal, which is essential for powerful machine learning applications.

3.6 Feature Matrix Construction

The feature table was then exported from the QIIME 2 feature table to a tab-separated values (TSV) file for downstream machine learning analysis in Python, after rarefaction. This data was a sparse, high-dimensional matrix with rows as samples (120) and columns as fungal amplicon sequence variants (ASVs). The number of individuals of each ASV was entered as a matrix value for each sample.

The feature matrix was also highly sparse because of the nature of the data in mycobiome sequencing, wherein many of the features were only present in a subset of the samples, leading to the vast majority of the matrix entries being zeros.

There was no explicit feature engineering or previous statistical feature filtering prior to model training. To enable the machine learning algorithms to learn directly from the complete fungal community profile, all detected ASVs were kept as input features. This was done to reduce possible investigator bias and to not exclude potentially important fungal taxa based on low abundance that could still have biological or predictive relevance for the classification of Alzheimer's disease.

3.7 Machine Learning Models

This study sets up the detection of Alzheimer's disease classification as a binary classification problem to assess the predictive capacity of the gut mycobiome. The four supervised machine learning architectures were selected to ensure that the linear and non-linear predictive boundaries were evaluated. The four different supervised machine learning architectures were used to provide a comprehensive evaluation of both linear and non-linear predictive boundaries, using the scikit-learn library.

3.7.1 Logistic Regression (LR)

The popular baseline (linear) model was Logistic Regression. It is in biomedical applications where it has been shown to be both easy to compute and easy to interpret (Dreiseitl & Ohno-Machado, 2002). The algorithm is a model of the log-odds of a sample belonging to the positive class (Alzheimer's disease) as a linear combination of fungal ASV abundances. The predicted probability $P(Y = 1 | X)$ is the output of a sigmoid function:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Typical logistic regression is very prone to severe overfitting because of the curse of dimensionality ($p \gg n$, where p is the number of features and n is the number of samples) of the mycobiome data. In order to counter this, an L2 penalty (Ridge regularization) was added to the loss function. This regularization reduces the importance of the less informative fungal ASVs, thereby making the model depend on the strongest signals of fungi.

3.7.2 Support Vector Machine (SVM)

SVM are powerful mathematically sound models, which aim to obtain the optimal decision boundary separating the two clinical classes to the maximum. This goal is to build a hyperplane in the high-dimensional ASV feature space that has the largest possible geometric margin between the AD samples and the cognitively normal controls (Cortes & Vapnik, 1995).

If the data are linearly separable, the hyperplane of the SVM is given by:

$$w^T x + b = 0$$

This is where 'w' is the normal to the hyperplane, 'x' is the input feature vector (ASV abundances) and 'b' is the bias. The input space was already very high-dimensional, and in this study, a linear kernel was used. The SVM is specifically designed for the most challenging samples for classification - the support vectors which are nearest to the decision boundary and it is particularly well suited for subtle differences between biological samples.

3.7.3 Histogram-Based Gradient Boosting (HGB)

While Random Forests build trees independently and in parallel, Gradient Boosting algorithms build decision trees sequentially (Breiman, 2001; Ke et al., 2017). A new shallow decision tree (weak learner) is intentionally learned to correct the errors that the ensemble of all learned shallow decision trees (weak learners) puts together (Ke et al., 2017).

The optimized implementation inspired by LightGBM, Histogram-based Gradient Boosting Classifier was used in this study. The HGB algorithm quantizes continuous biological feature data to integer-based bins (histograms). This greatly speeds up the time taken to consider node splits and dramatically cuts down on the model's memory consumption. Gradient boosting is extremely powerful in fitting complex data, is more aggressive than Random Forest and riskier when dealing with small datasets.

3.7.4 Random Forest (RF)

Random Forest is a very sophisticated ensemble learning technique which is regularly achieving state-of-the-art results on the microbiome omics data. The Random Forest algorithm is able to model complex, non-linear biological interactions by building many independent decision trees in training, which are then combined into a single decision at the final stage of the algorithm (Breiman, 2001).

The strength of the RF algorithm comes from two stochastic processes: bagging (bootstrap aggregating) and the random subspace method. The first is that each individual tree is trained using a random sub-sample of the data with replacement. Second, each node split in a tree only processes a random sub-sample of available ASV features to identify the best split according to the criteria of Gini impurity reduction. To make the final prediction on the classification, a majority vote is taken over all the trees in the ensemble. This architecture is most used to reduce model variance and offers a huge resistance to overfitting when dealing with a relatively small clinical cohort.

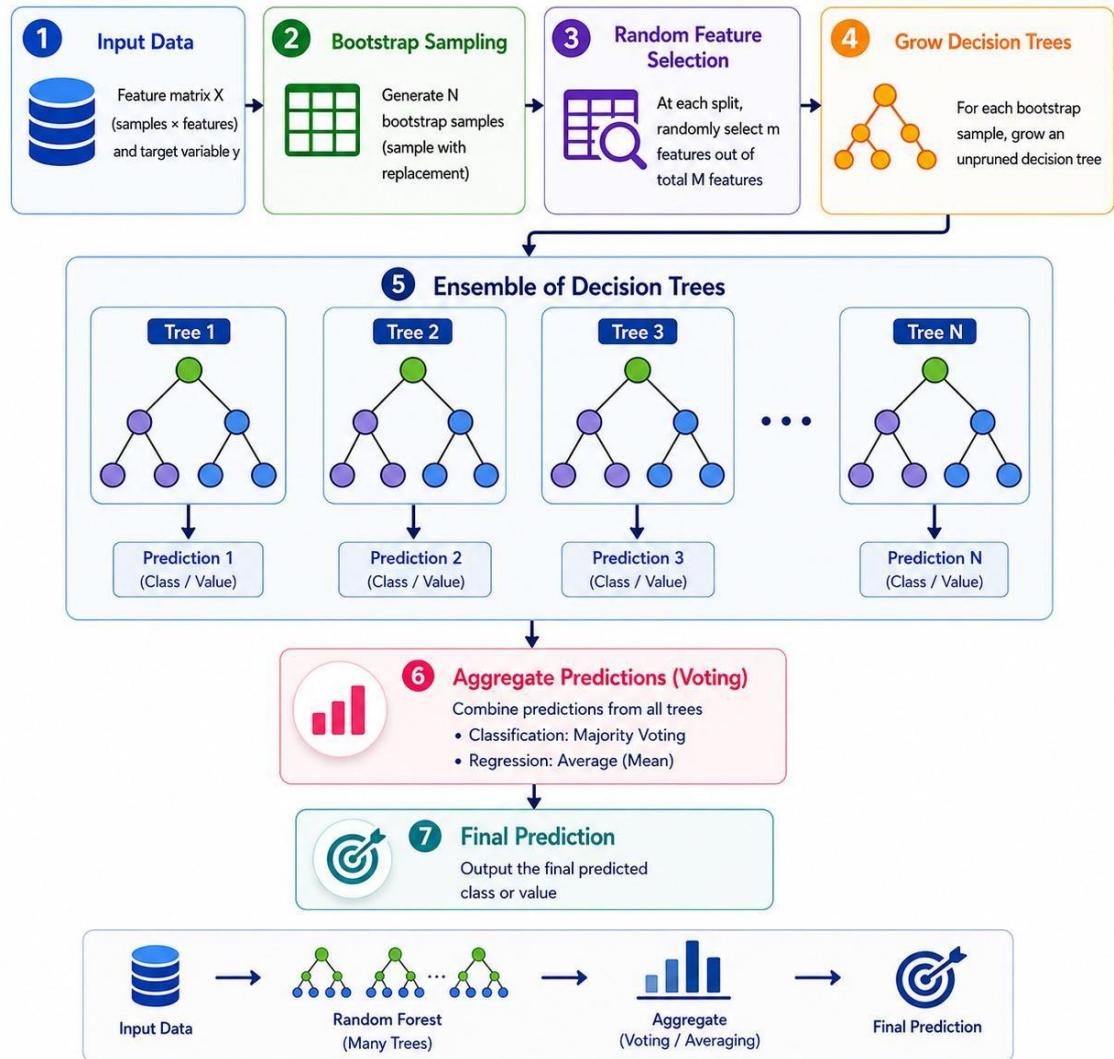


Fig 3.2: Random Forest architecture showing the bagging process and ensemble voting mechanism.

3.8 Model Training Strategy

A strict training and validation procedure was adopted to ensure generalizability of the machine learning models and to avoid data leakage.

3.8.1 Dataset Splitting

The normalized ASV feature matrix was divided into independent training set (80% n=96) and isolated hold out test set (20% n=24). The test set was completely held back and isolated throughout the model training and served as an ultimate unbiased

assessor of the real-world model performance. Stratified sampling was used to ensure the same distribution of the AD patients and healthy controls in both datasets. In addition, a fixed random state (random_state=42) was globally set, which ensures a completely reproducible split of the data and a completely reproducible stochasticity in the algorithms used in later studies.

3.8.2 Cross-Validation

The overall sample size of $N=120$ is moderately sized, so the models are quite sensitive to variations in the manner in which the data is partitioned. In order to thoroughly prove stability and reliability of classifiers, k-fold Stratified Cross-Validation ($k=5$) was followed on the training set.

The training set was split into five equally-sized folds. The model was trained on 4 folds and tested on the last one-fold, until each fold had been used for testing once. This helps prevent the "lucky" train-test split from skewing the performance reported in the model, and ensures that the reported performance is indicative of real, underlying patterns in biology.

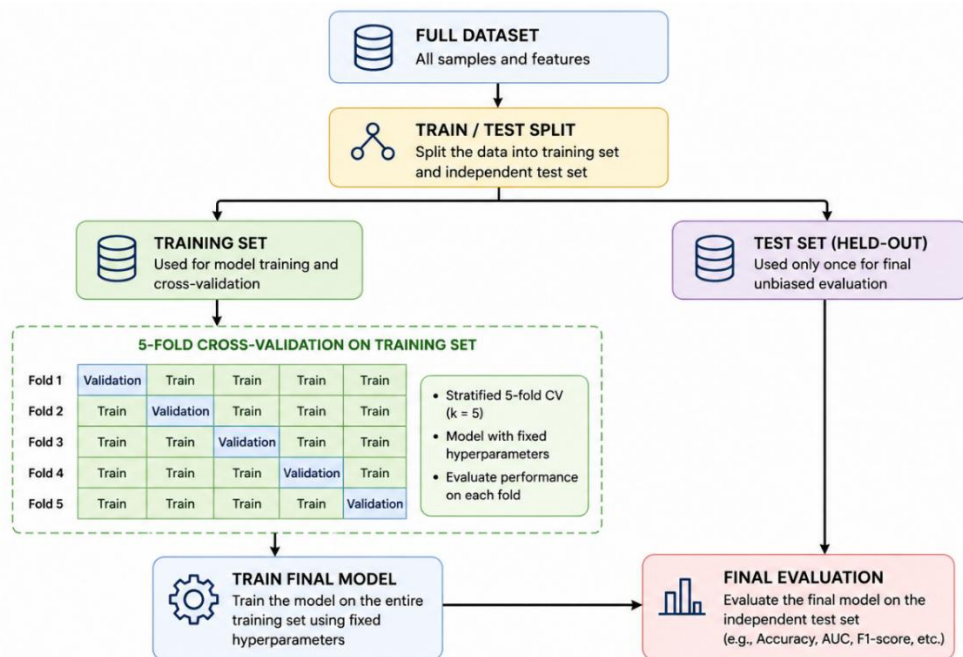


Fig 3.3: Machine Learning Pipeline

3.9 Model Evaluation Metrics

The predictive models were comprehensively evaluated with a set of common metrics of machine learning given by the confusion matrix, including True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) (Sokolova & Lapalme, 2009). True Positive - a patient correctly identified as having Alzheimer's disease, and False Negative - a patient who was incorrectly classified as a healthy control in this clinical context.

1. Accuracy: The accuracy is the percentage of the total number of predictions made, that were correct (Sokolova & Lapalme, 2009). Intuitive but may be misleading using imbalanced data sets; our data set is balanced 60/60, so this is not a concern.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision (Positive Predictive Value): Percentage of positive identifications that were positive. The percentage of false alarms is considered low if it is a high level of precision (Sokolova & Lapalme, 2009).

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. Recall (Sensitivity): The percentage of true positive cases that were correctly classified. When used in a medical screening environment, there may be an emphasis on having as many as possible (recall) patients who are not sick (Sokolova & Lapalme, 2009).

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. F1-Score: The harmonic mean of Precision and Recall. It offers a fair evaluation of the performance of a model, especially when the cost of false positives is the same as the cost of false negatives (Sokolova & Lapalme, 2009).

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. Receiver Operating Characteristic - Area Under Curve (ROC-AUC):

The ROC curve is a plot of the True Positive Rate (Recall) against the False Positive rate at different classification thresholds. The AUC is the area under this curve that the AUC computes for two dimensions. The closer to 1.0, the closer the AUC is to a perfect classifier, and the closer to 0.5, the closer the AUC is to a random guess. The ROC-AUC was chosen as the primary evaluation metric for this dissertation as it evaluates the model's discriminative threshold agnostic of rigid class boundaries.

3.10 Feature Importance and Interpretability

A model that predicts Alzheimer's disease with high accuracy is computationally impressive, but it lacks clinical utility if the biological basis of its decision remains a black box. Therefore, the final methodological step was to extract biological interpretability from the models.

The natural property of the Random Forest classifier of measuring the relative contribution of features in the bagging process made it a natural choice for extracting Mean Decrease in Impurity (MDI) scores. The MDI will be calculated as the reduction in entropy/gini (uncertainty) across all the decision trees in the forest caused by a given fungal ASV.

Features with the highest importance scores were remapped to their taxonomic annotation from their alphanumeric ASV identifiers (ASVs) in the UNITE database. This is a key step in converting mathematical weights into biological meaning, specifically identifying which fungal genera, like *Candida* or *Purpureocillium*, are the main contributors to the pathogenesis of the gut-brain axis, and thus make them potential targets for future in vivo biomarker validation studies.

CHAPTER 4 – RESULTS AND DISCUSSION

4.1 Introduction

The main aim of this study was to assess the ability of gut mycobiome to differentiate Alzheimer's disease (AD) patients from cognitively normal participants. This work takes traditional, descriptive microbiome ecology a step further by utilizing high throughput fungal Internal Transcribed Spacer (ITS) amplicon sequencing data and a machine learning approach. The outputs of the bioinformatics processing pipeline are thoroughly analyzed, as well as the community structures of the taxa. Four different classifiers - Random Forest (RF), Histogram-based Gradient Boosting (HGB), Logistic Regression (LR) and Support Vector Machine (SVM) - are evaluated extensively in terms of their performance. Additionally, this chapter elucidates the biological relevance of the discriminative fungal features identified by the best model and situates them in the context of the existing literature on the microbiota-gut-brain axis.

4.2 Sequencing and Quality Control Results

In this study, the raw sequencing data analyzed were paired end FASTQ files from fecal samples of a well-established clinical cohort of 120 individuals (60 AD patients and 60 healthy controls). Machine learning classifiers are very sensitive to artifacts in the sequencing data, difference in library sizes and noise, and quality control and integrity of data are crucial prerequisites to achieving reliable machine learning.

4.2.1 Read Processing and Denoising

The raw reads were processed using the QIIME 2 bioinformatics platform. Quality control was rigorously applied after the first import by applying the DADA2 algorithm. A common approach to traditional microbiome studies was to use Operational Taxonomic Unit (OTU) clustering to group sequences based on an arbitrary similarity threshold of 97%, which can obscure slight differences in strains that may be important. However, DADA2 builds a parametric error model specific to the sequencing run, which differentiates true biological sequences from amplification errors to create accurate Amplicon Sequence Variants (ASVs).

In the denoising process, reads were quality filtered according to expected error rates, dereplicated, and chimeric sequences removed by a denoting algorithm, which eliminates sequences resulting from PCR artifacts brought about by the misalignment of multiple biological templates. As expected, with the conversion of raw PE reads to high-quality, non-chimeric ASVs, there was a significant loss of reads. The extensive filtering done on this feature matrix will ensure that the biological entities that are sent to the machine learning models are not technical noise.

4.2.2 Rarefaction

One of the primary difficulties in processing the data from microbiome omics is the considerable variation in the amount of sequencing achieved for each biological sample. If the microbial richness of samples is directly determined by their sequencing depth, then this will create great biases in statistical and machine learning models. To counteract this, the processed ASV feature table was rarefied to a common sample depth, a common normalization method that involves randomly sub sampling without replacement the samples.

In this data, the depth of the rarefaction was set to 10,000 reads per sample. The discarding of excess reads, as occurs with rarefaction, occurred naturally with these samples because they were deeply sequenced, but we were left with a very strong and standard feature matrix of 10,000 reads. Preservation of strong classification signals after post-rarefaction validates that the models are capturing true biological patterns, not merely technical (sequencing depth) variations.

4.3 Taxonomic Profiling Results

The taxonomic composition of the cohort's gut mycobiome was profiled before the application of machine learning. A Naive Bayes classifier pre-trained using the UNITE fungal reference database (Version 10.0) was used for the annotation of the ASVs. UNITE employs a dynamic Species Hypothesis (SH) framework essential for the discovery of fungal lineages that are largely un-cultured and poorly-described in the traditional Linnaean taxonomy.

The fungal community in the gut of the cohort was consistent with the global mycobiome literature, and it was mostly comprised of Ascomycota and Basidiomycota fungi. Mycobiome was highly variable among individuals at genus level. *Candida*, *Saccharomyces* and *Malassezia* were common commensals and opportunistic genera found in both AD and control samples.

But a key conclusion was that there was a large element of ‘dark taxa’. Many of the ASVs were only confidently classified to a higher taxonomic level (e.g., Family or Order) or "Fungi" (a very broad classification). Importantly, in the following machine learning pipeline, these unclassified ASVs were kept as separate numerical features. The models were able to retain predictive value in these uncharacterized fungi without having to reduce the data to known genera, allowing them to harness the usefulness of sequence-level exactness.

4.4 Machine Learning Classification Results

The central hypothesis of this study was that supervised machine learning architectures were capable of using the high dimensional ASV profiles to accurately classify AD. The feature matrix was constructed in a rarefied manner and was divided into an independent training set and an isolated test set in a stratified manner with sizes of 80% and 20% respectively. A set of classifiers, namely Random Forest (RF), Histogram-based Gradient Boosting (HGB), Logistic Regression (LR), Support Vector Machine (SVM) were evaluated on the test set.

A clear hierarchy of algorithmic effectiveness is seen in the comparative evaluation. The RF model easily surpassed all the other models in every metric; when it came to the independent test set, the classification Accuracy of the RF model was 0.75 and the ROC-AUC value was 0.85 which was impressive. The model has good precision (0.76) and recall (0.75), meaning that it is very effective in correctly identifying Alzheimer's cases while preventing unnecessary alerts.

Table 4.1: Classification Performance of Supervised Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Mean CV ROC-AUC
Random Forest	0.75	0.76	0.75	0.75	0.85	0.72
Histogram-based Gradient Boosting	0.63	0.69	0.62	0.59	0.69	0.59
Logistic Regression	0.58	0.59	0.58	0.58	0.59	0.62
Support Vector Machine	0.54	0.54	0.54	0.53	0.54	0.57

By comparison the linear and kernel-based models were very poor. Logistic Regression (LR) and Support Vector Machine (SVM) gave the lowest accuracy of 0.58 and 0.54, respectively, which is only slightly better than a random guess. The low performance of these models demonstrates the significant limitation of linear decision boundaries in dealing with the highly complex, multivariate and non-linear interactions in microbial ecosystems.

The second one, Histogram-based Gradient Boosting (HGB), a very powerful Tree-based Ensemble model, had intermediate performance with an Accuracy of 0.63 and ROC-AUC of 0.69. Although better than the linear models, the performance of HGB was far behind that of RF. Gradient boosting algorithms are constructed one tree at a time trying to minimize the residuals. This was a very aggressive optimization algorithm logic in a very sparse, very high dimensional dataset with a moderate number of samples, harming its ability to generalize on unseen test data.

The reason why the Random Forest algorithm is superior is due to its basic structure. The random feature subspace selection at each node split and the use of bootstrap aggregation (bagging) in RF naturally reduce the variance of the model, and make the data-independence analysis of sparse, high-dimensional matrices simple and efficient, rather than requiring some extreme data reduction methods that risk discarding biologically relevant information.

4.5 Confusion Matrix Analysis

The independent test data set was used to perform a confusion matrix analysis to assess the clinical relevance of model predictions. The various kinds of prediction errors have varying diagnostic implications in the classification of Alzheimer's disease.

The confusion matrix outcomes are defined as follows:

- **True Positive (TP):** AD sample correctly classified as AD.
- **True Negative (TN):** Control sample correctly classified as Control.
- **False Positive (FP – Type I Error):** Control sample incorrectly classified as AD.
- **False Negative (FN – Type II Error):** AD sample incorrectly classified as Control.

The overall predictive behaviour of the evaluated classifiers was presented through confusion matrix analyses; as shown in the figures, the maximum balanced overall predictive behaviour was presented by the Random Forest classifier. Sensitivity of the RF model is good as it successfully diagnosed 8/12 AD samples; specificity of

the model is relatively low with False Positive predictions. The Support Vector Machine classifier had the same number of True Positive classifications, but many more False Positive predictions, resulting in a lower overall specificity.

The Histogram-based Gradient Boosting model was the most specific but was the least sensitive and was unable to correctly classify a large proportion of AD samples. Logistic Regression had moderate but relatively weaker classification performance meanwhile.

Minimizing False Negative is especially relevant from a clinical standpoint in neurodegenerative disease screening when accuracy in identifying true cases of AD can result in a delay in clinical evaluation, treatment, and supportive care during the earlier stages of the disease. In this respect, the RF and SVM classifiers showed the highest sensitivity levels of all the classifiers analyzed. The Random Forest classifier, however, had the least number of False Positive predictions and more balanced classification profile and showed more clinical reliability within the scope of this study.

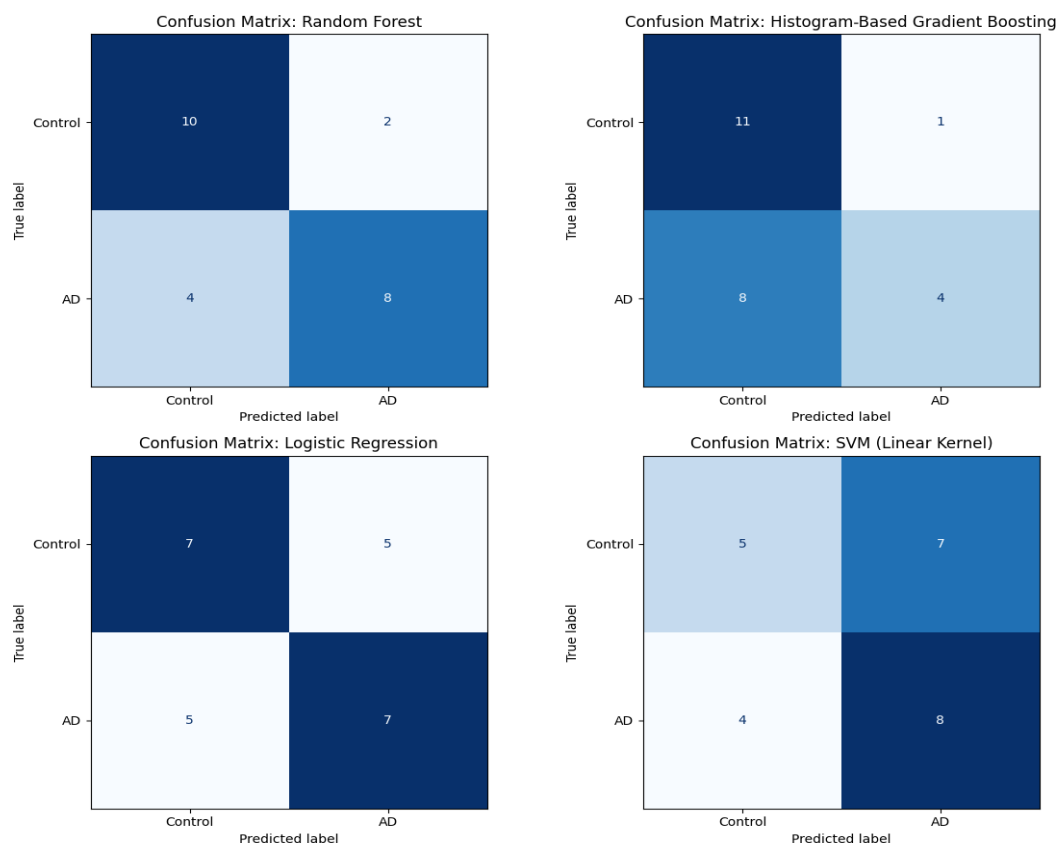


Fig. 4.1: Confusion matrices for Random Forest, Histogram-based Gradient Boosting, Logistic Regression, and Support Vector Machine classifiers evaluated on the test set

4.6 ROC Curve and Cross-Validation Stability Analysis

Single splits of the test set in moderate-sized cohorts can be subject to stochastic variability, and test set metrics can only be used as an estimate of performance. In order to strictly test the generalizability of the model and to make sure that it was not simply fitting the statistical noise, a 5-fold stratified cross validation (CV) was employed on the training data (Kohavi, 1995).

The Receiver Operating Characteristic (ROC) curve is used to assess the continuously changing relationship between the True Positive Rate (Sensitivity) and the False Positive Rate ($1 - \text{Specificity}$) as the classification threshold changes (Fawcett, 2006). This discriminative power is measured by the area under the curve (AUC) with 1.0 indicating perfect separation and 0.5 indicating a non-informative model.

The Mean CV ROC-AUC with 5-fold cross validation was 0.72 for Random Forest which was the most robust architecture. The corresponding ROC plots showed that the RF model had a fairly small standard deviation (shaded area) for the five different folds of data. This tight variation is an important validation parameter and it solidly demonstrates that the model has learned generalizable, underlying biological rules about the AD mycobiome, and not the specific idiosyncrasy of a single training subset.

The other models exhibited poor stability. In spite of its theoretical strength, the Mean CV ROC-AUC for Histogram-based Gradient Boosting was 0.59 and it showed greater variance between folds supporting the hypothesis of its overfitting tendency on sparse omics data. The Mean CV ROC-AUC of Support Vector Machine was the lowest at 0.57 with high variation between folds, which implies that it was not at all suitable for finding a robust biomarker in the un-engineered ASV feature space.

Finally, the cross-validation analyses validate ensemble-based, bagging methods such as Random Forest as being currently the best computational approach to reliably obtain predictive signals from data derived from human gut mycobiome.

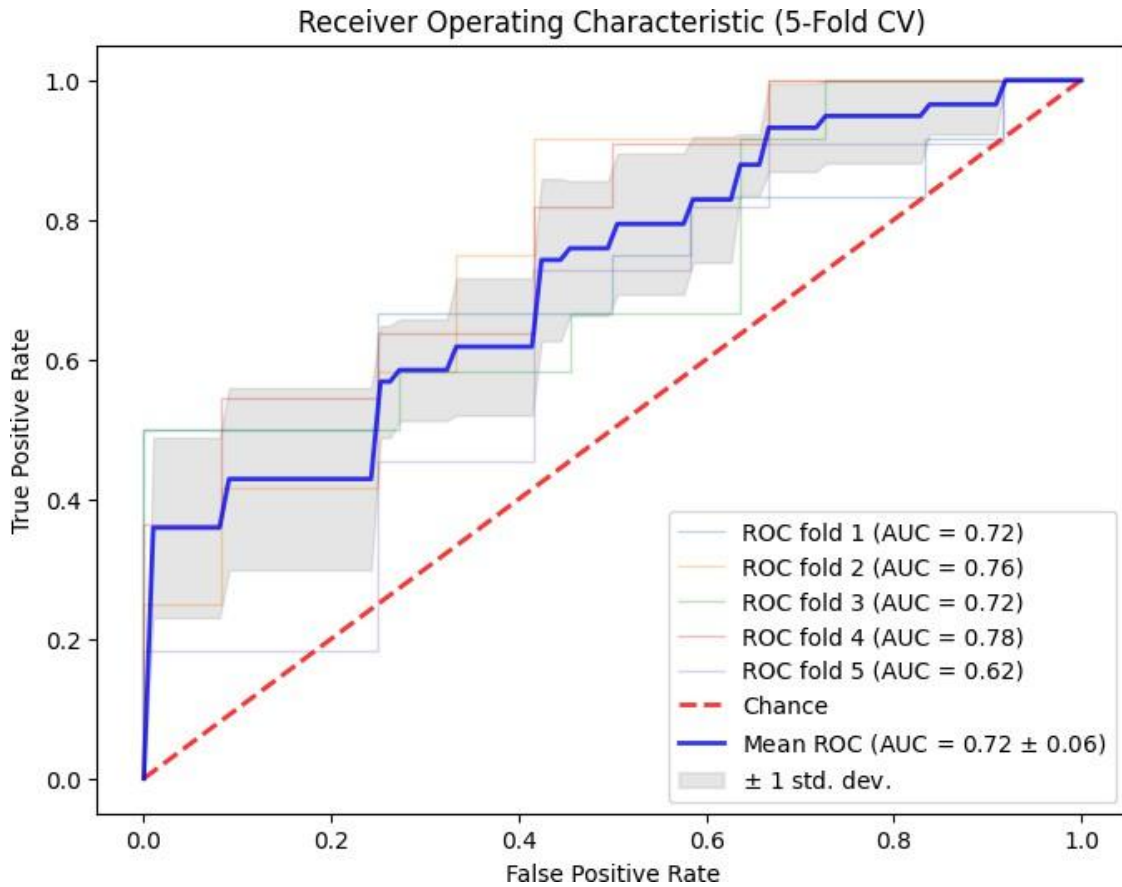


Fig.4.2: Receiver Operating Characteristic (ROC) curve for the Random Forest classifier using 5-fold cross-validation. The blue line represents the mean True Positive Rate vs False Positive Rate across all folds (Mean ROC-AUC = 0.72). The grey shaded area illustrates the standard deviation (± 1 std. dev.), confirming the model's stability despite the moderate sample size.

4.7 Feature Importance and Biological Interpretation

One of the major drawbacks of a lot of medical machine learning applications is the 'black box' effect, in which a machine learning model makes good predictions without providing any mechanism. Tree-based algorithms do this using Feature Importance extraction. The internal Gini impurity reduction scores of the Random Forest model were used to quantitatively assess the most critical ASVs in the classification of AD. The mathematical extraction of these features converts the computational model back to biologically useful hypotheses.

The algorithm gave high importance to particular genera of fungi as seen in the

feature importance analysis. The most discriminative taxa that affected the predictions of the model were genera such as *Candida*, *Purpureocillium*, *Oligophagozoma* and *Debaryomyces*.

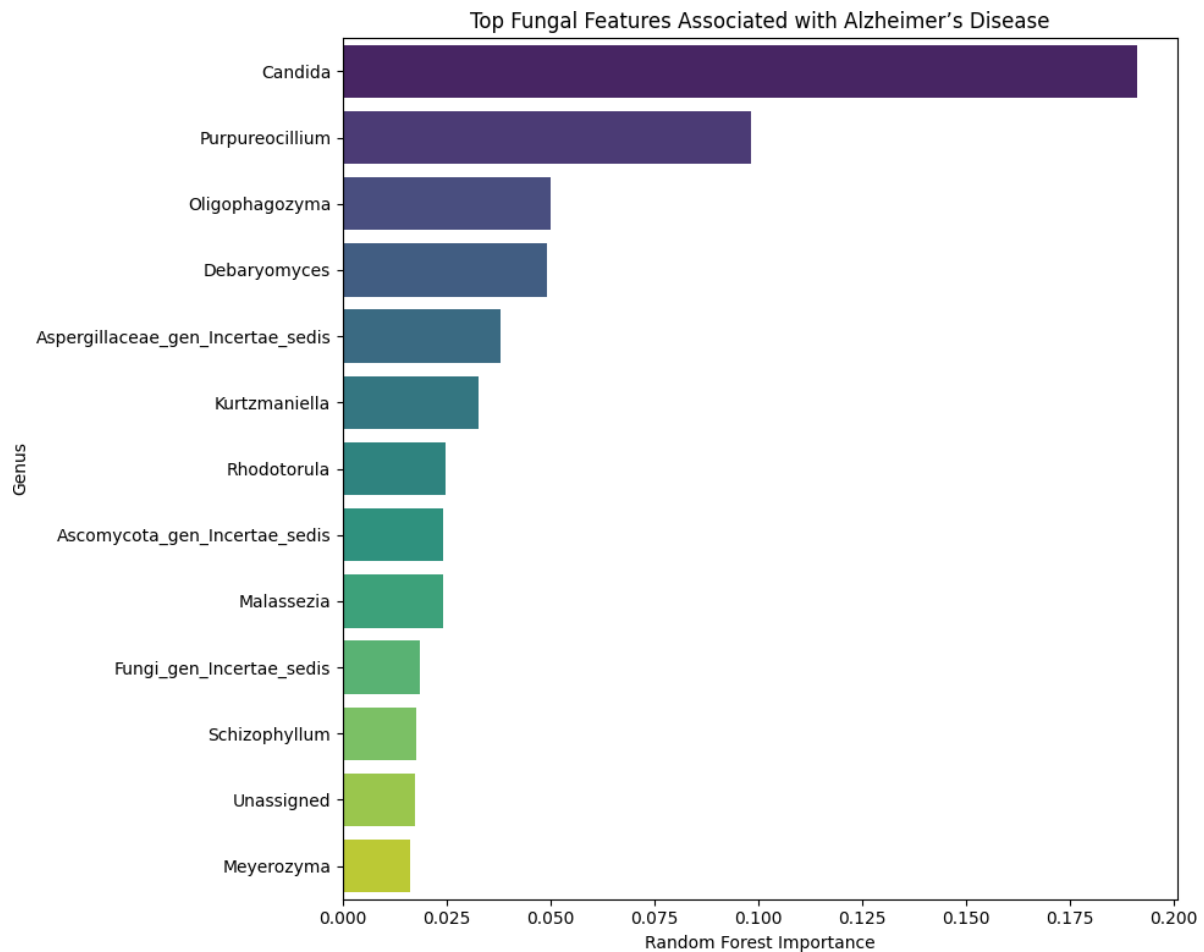


Fig 4.3: Top fungal features contributing to Alzheimer's disease classification identified by Random Forest feature importance analysis.

4.7.1 The Role of *Candida* in Neuroinflammation

Candida's rise to become a key predictive biomarker is in perfect alignment with current pathophysiological knowledge of the gut-brain axis. *Candida* species like *Candida albicans* and *Candida tropicalis* are known as opportunistic pathobionts in human gastrointestinal tract (Yetgin, 2024). Major enrichment of *Candida* has been

reported in previous descriptive studies in AD cohorts.

The mechanism of action between *Candida* and AD is very complex and related to host innate immunity. An immunogenic beta-glucan is present in large quantities in the cell walls of *Candida*. These beta-glucans strongly bind to host macrophages receptors, Dectin-1, to activate a powerful CARD9-mediated signaling pathway when the gut dysbiosis allows the overgrowth of *Candida*. Through this cascade, T-helper 17 (Th17) cells are stimulated to differentiate, leading to the release of large amounts of pro-inflammatory cytokines like IL-17 and IL-23 (Gaspar et al., 2025).

In AD, this is a chronic inflammation of the entire body that is driven by *Candida* and leads to a leaky gut which allows the cytokines to enter the bloodstream and weaken the blood brain barrier. After the barrier is broken, peripheral inflammation triggers resident central nervous system microglia to become chronically neurotoxic, directly leading to deposition of amyloid-beta plaques and tau tangles (Rutsch et al., 2020). Thus, the algorithm's dependence on *Candida* abundance for AD classification does not arise as an artifact, but is a direct reflection of underlying inflammatory pathogenesis.

4.7.2 *Purpureocillium* and *Oligophagozoma*

The discovery capabilities of machine learning are demonstrated by the fact that *Purpureocillium* and *Oligophagozoma* were identified as the most predictive features. *Purpureocillium* is commonly called an environmental fungus and an opportunistic human pathogen capable of causing deep local and systemic infections in immunocompromised patients (Alharbi et al., 2022; Luangsa-ard et al., n.d.). The overrepresentation of its presence in the AD cohort could indicate a general breakdown of immune surveillance or changes in the gut microenvironment in the patients with neurodegenerative disease that allows for the growth of atypical fungi. *Oligophagozoma* is a relatively little-known basidiomycetous yeast, which is an ideal target for future multivariate machine learning studies and mechanistic in vivo studies (Li et al., 2020).

4.7.3 *Debaryomyces*

Debaryomyces is known to interact in complex manner with the host immune system and abundance changes in this species could possibly be linked to the general dietary and metabolic changes that are often observed in the course of dementia (Forbes et al., 2019).

In addition, a large number of unclassified ASVs were assigned significant weight, pointing to the fact that a significant proportion of the predictive diagnostic signal is in "dark taxa" fungi that remain to be formally biologically characterized.

4.8 Comparative Discussion with Previous Studies

This computational framework provided results that are a major step forward from past work on mycobiome. The previous study by (Ling et al., 2021) was able to demonstrate the presence of fecal fungal dysbiosis in AD with the help of univariate statistical. The use of descriptive statistics is limited, though, because knowing that two groups are statistically different ($p < 0.05$) does not necessarily mean that these differences can be used to accurately predict the status of a new patient that has yet to be seen.

Likewise, in a small pilot study of patients with Mild Cognitive Impairment, (Nagpal et al., 2020) showed a positive correlation between fungal taxa (such as *Botrytis* and *Cladosporium*) and cerebrospinal fluid tau markers. Correlational matrices are useful for understanding the biological nature of the data but cannot be used as diagnostic tools.

This study goes beyond these descriptive paradigms. Using an independent test set and stringent 5-fold cross-validation, the Random Forest model statistically demonstrates the discriminative ability of fungal ASV patterns to be computationally used for disease classification. A 0.75 classification accuracy and 0.85 ROC-AUC that was achieved using only fungal ITS sequencing data is an extraordinarily promising result, given the signal of fungal data is very low in biomass and very sparse. It demonstrates that the gut mycobiome, which accounts for less than 0.1% of the total microbial community, has a predictive signal comparable to that of larger mycobiome classification models, bacterial microbiome models (Velasquez & Oudah, 2023).

4.9 Clinical and Biological Implications

In a clinical setting, the successful performance of the Random Forest classifier confirms the idea of a non-invasive, stool-based screening modality for Alzheimer's disease. The gold standard diagnostics (e.g. CSF analysis performed in lumbar puncture and PET scans) are very invasive, expensive and not available for most of the population in the world (Knopman et al., 2021). A standalone mycobiome test is not yet in a place to supplant these definitive modalities, but the optimized machine learning pipeline that analyzes patterns in gut fungal data may be an excellent, readily available "first-line" biomarker. Clinicians may be able to stratify patients

based on risk levels to identify those who need costly secondary neurological testing and thus conserve resources in medical care.

From a biological perspective, the extraction of particular feature importance vectors rephrases the therapeutic strategies. The presence of other taxa such as *Candida* and *Purpureocillium* may be more indicative of the disease state and are not just bystanders, they are active biological factors causing neuroinflammation. This opens the fascinating opportunity for interventions acting on the mycobiome. Precision antifungal therapy targeting only the pro-inflammatory pathobionts in the absence of affecting the overall bacterial microbiome may be a future treatment option. On the other hand, the use of beneficial, neuroprotective yeast strains, such as *Saccharomyces boulardii*, to effectively outcompete opportunistic yeast pathogens, restore inter-kingdom ecological homeostasis and then dampen the systemic immune activation that drives neurodegeneration could also be considered (Ye et al., 2022).

4.10 Conclusion

As shown in this chapter, it is possible to convert the raw fungal ITS sequencing data into usable bioinformatics data with thorough processing by QIIME 2 and to use it as input for supervised machine learning architectures. The Random Forest ensemble algorithm was significantly better than linear models with an ROC-AUC of 0.85 on test data and strict robustness in 5-fold cross-validation. Importantly, the model's predictive capacity was greatly dependent on biologically relevant taxa, especially *Candida*, forming a direct computational link to the known immunological mechanisms of gut-brain axis. The results validate the utility of the gut mycobiome for diagnosis and offers a solid mathematical framework for future longitudinal biomarker discovery.

CHAPTER 5 - CONCLUSION

5.1 Summary

The main goal of this thesis was to develop, implement, and thoroughly test a computational tool that can accurately classify Alzheimer's disease (AD) based on sequencing data from gut fungi Internal Transcribed Spacer (ITS). Bacterial microbiome in Microbiota-Gut-Brain Axis has been well studied, however, gut mycobiome is still an under-investigated field, given its low absolute biomass, complex bioinformatics requirements. This research shifted the mycobiome analysis from ecological description, to a predictive clinical informatics.

A publicly available paired-end ITS2 sequencing dataset of 120 individuals (60 AD patients and 60 cognitively normal controls) was used to perform the study. The bioinformatics part took place in the standardized QIIME 2 platform. The raw sequencing reads were subjected to the stringent quality control, primer trimming with Cutadapt and state-of-the-art denoising with DADA2 algorithm. This method produced high resolution Amplicon Sequence Variants (ASVs), which circumvents the limitations of the traditional and low-resolution Operational Taxonomic Unit (OTU) clustering. The resulting high dimensional ASV feature matrix was then rarefied to a depth of 10,000 reads per sample to ensure that each sample has the same number of reads and to prevent any bias caused by differences in sequencing depth.

For machine learning, the detection of AD was defined as supervised binary classification. Four algorithmic architectures – Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Histogram-based Gradient Boosting (HGB) – were optimized and evaluated. An independent test set evaluation and a solid 5-fold stratified cross-validation protocol were used to evaluate the models for generalization, stability and predictive viability. Finally, biological interpretability was gleaned by assigning the most relevant mathematical characteristics of the algorithm to their UNITE taxonomic annotation, thanks to the ability to bridge the gap between predictive computation and biological pathogenesis.

5.2 Major Findings

The strict testing of the machine learning classifiers, however, resulted in a number of key computational and biological findings that support the central thesis of this thesis:

- **Superiority of Ensemble Learning:** The Random Forest (RF) classifier clearly outperformed all other classifiers in mycobiome-based classification. For independent test set, RF performance was remarkable with an accuracy of 0.75, precision of 0.76, recall of 0.75 and an ROC-AUC of 0.85. In addition, it had the greatest resistance to stochastic variation with a Mean cross-validated ROC-AUC of 0.72. The bagging architecture and the random feature subspace selection inherent in RF was very effective in dealing with the sparsity and extreme dimensionality of the microbiome datasets and avoiding overfitting.
- **Inadequacy of Linear Decision Boundaries:** The simple linear decision boundary models (Logistic Regression and Support Vector Machine) were not very successful, with the maximum accuracy being slightly above random guessing (0.58 and 0.54 respectively). This significant drop in performance clearly illustrates that these ecological changes, which are part of AD dysbiosis, are highly non-linear and multivariate. A single linear hyperplane is not enough to represent the complex and interdependent shifts in microbial communities that occur during neurodegeneration.
- **Identification of Predictive Fungal Taxa:** Feature Importance scores were extracted from the Random Forest model to show that the classification was not due to random technical artifacts, but due to biologically relevant fungal lineages. The presence of particular genera such as *Candida*, *Purpureocillium*, *Oligophagozyma* and *Debaryomyces* was highly weighted in the model. These taxa serve as the mathematical anchors of the predictive model, with the strongest discriminative signals between an Alzheimer's patient and a healthy person.
- **Diagnostic Viability of the Mycobiome:** Gut mycobiome's signal is discriminative enough to result in accurate machine learning classification, even though it represents only <0.1% of the total gut microbial community, highlighting the diagnostic viability of the Mycobiome. The success of the computational application of this framework validates the high fidelity of gut fungal profiles as markers of systemic and neurological health.

5.3 Biological Implications

The biological interpretation of the feature importance in mathematics gives a deep understanding of the etiology of Alzheimer's disease through the microbiota-gut-brain axis. It is very insightful that the *Candida* marker is the strongest predictor in the Random Forest model and it is in perfect harmony with the consensus on peripheral neuroinflammation.

The *Candida* species are ubiquitous opportunistic pathobionts. Their cell walls contain rich amounts of immunogenic polysaccharides, especially beta-glucans and mannans. A healthy gut will allow for the growth of commensal bacteria that compete with *Candida* growth and a healthy intestinal barrier to prevent *Candida* growth. *Candida* may, however, grow uncontrollably in the dysbiosis situation often found in AD patients. The biological significance of this is immense, because the physical overgrowth of *Candida* can cause a huge immune response locally.

The fungal beta-glucans strongly engage the Dectin-1 pattern recognition receptors on the cell surface of resident macrophages and dendritic cells in the intestinal tract. This binding triggers the intracellular signaling cascade of CARD9 which induces differentiation of T-helper 17 (Th17) cells and causes massive local release of pro-inflammatory cytokines, such as Interleukin-17 (IL-17) and Interleukin-23 (IL-23). This chronic inflammation leads to breakdown of tight junction proteins in the intestinal lining causing a "leaky gut."

In the case of breach of the intestinal barrier, these inflammatory cytokines enter the systemic circulation, along with fungal antigens. This peripheral inflammatory storm is then absorbed into the central nervous system (CNS) and causes the permeability of the blood-brain barrier (BBB) to change. Upon arrival of cytokines, resident microglia and astrocytes in the brain change their state from a neuroprotective, phagocytic to a chronically activated, neurotoxic state. This chronic microglial activation is what is responsible for the marked impairment of the brain's ability to remove amyloid beta plaques and what actively promotes hyperphosphorylation of tau proteins which in turn promote neuronal apoptosis and cognitive loss.

Thus, the machine learning model's use of *Candida* and *Schizophyllum* as a classification of AD is a direct mathematical representation of this inflammatory cascade. These fungal-markers are not merely associative but most likely serve as upstream regulators of the systemic immune-dysregulation leading to a terminal neurodegeneration.

5.4 Technical Limitations

The computational framework predicatively showed good results, yet it is essential to recognize the technological and methodological constraints of the study that should help to frame the obtained results.

- 5.4.1 **Moderate Sample Size:** The sample size used was 120 people. It is quite a bit, but is relatively small for training highly complex models. If the deep learning architectures were not used with the risk of overfitting, then they would not be viable. More samples are needed to develop models that can detect less common fungal phenotypes and be clinically accurate.
- 5.4.2 **Cross-Sectional Design:** The biggest weakness of this data and the resulting predictive model is that it is cross-sectional. This data represents one moment in time of the mycobiome at the time of a diagnosed AD. It is therefore not possible to definitively distinguish correlation from causation. The model predicts AD based on fungal signatures, but does not tell us if the dysbiosis of fungi is the primary cause of AD or if it is a secondary effect of the late-stage dementia altered diet, metabolism and motility.
- 5.4.3 **ITS Sequencing and Primer Biases:** ITS2 is a region that is prone to primer bias during amplicon sequencing. Certain "dark taxa" may be systematically underrepresented or not detected at all in the ASV feature table as certain fungal lineages are amplified with greater efficacy than others with universal fungal primers. Moreover, ITS sequencing can generate taxonomic information but cannot give any information about the metabolic or functional activity of the existing fungi.
- 5.4.4 **Sparsity of Fungal Data:** Many fungal ASVs are found in a minority of the cohort, the gut mycobiome is extremely sparsity. This level of sparsity is hard for mathematical stability of the machine learning models, as this can yield different rank orders for features under slight variations in training distributions.

5.5 Future Scope

The intersection of machine learning, gut mycobiome and neurodegeneration is a new and emerging area. Such are the results and constraints of this dissertation, it suggests a number of very advanced research areas:

Deep Learning and Multi-Omics Integration: With growing cohort sizes, the use of traditional ensemble models should be phased out in favor of deep learning, specifically for sparse tabular data. Most important of all, future models should

have a multi-omics framework. In combination with the bacterial 16S data and fungal ITS data, host metabolomic data can be used to map complex inter-kingdom ecological networks. A much more accurate predictive model will be created when the cooperative/antagonistic nature of specific bacteria and fungi (e.g., the depletion of a specific bacterium allows *Candida* to overgrow) is identified.

Longitudinal Cohort Studies: Future designs need to be based on large scale, prospective, longitudinal cohort designs to go beyond correlation to the issue of disease causation.

Machine learning models could be trained on the mycobiome profiles of cognitively normal people over ten years, to predict when they will develop Alzheimer's disease and detect meaningful changes in the fungal communities well before irreversible brain damage takes place in the preclinical stage of the disease.

Whole Metagenome Shotgun Sequencing (WMS): Future studies should use ultra-deep Whole Metagenome Shotgun Sequencing (WMS) along with improved host-DNA depletion methods to avoid primer biases in ITS sequencing. In addition to detecting what fungi are present, WMS will enable detection of what they are doing; for example, functional genomic pathways (the genes that are involved in the synthesis of neurotoxic fungal metabolites) will be used as features for classification.

Therapeutic Interventions: If predictive fungal biomarkers have been identified as causal factors, they can be very specific therapeutic targets. New computed models could simulate and predict the response of patients to precision mycobiome interventions, including targeted and narrow-spectrum antifungals or the delivery of neuroprotective probiotic yeasts, such as *Saccharomyces boulardii*, which could herald novel, non-invasive therapeutic opportunities for the prevention and treatment of Alzheimer's disease.

REFERENCES

- Agarwal, U., Tonk, R. K., & Verma, S. (2026). Pathogens Association with Alzheimer Disease: Emerging Concepts and New Perspectives. *Current Gene Therapy*, 26(2), 1–20. <https://doi.org/10.2174/0115665232389348250722073721>
- Alharbi, M., Alruqaie, N., Alzahrani, A., & Almuneef, M. (2022). Paecilomyces/Purpureocillium Infection in Children, Case Report, and Review of the Literature. *Journal of Fungi*, 8(9), 930. <https://doi.org/10.3390/jof8090930>
- Alonso, R., Pisa, D., Fernández-Fernández, A. M., & Carrasco, L. (2018). Infection of Fungi and Bacteria in Brain Tissue From Elderly Persons and Patients With Alzheimer's Disease. *Frontiers in Aging Neuroscience*, 10, 159. <https://doi.org/10.3389/fnagi.2018.00159>
- Basgaran, A., Lymberopoulos, E., Burchill, E., Reis-Dehabadi, M., & Sharma, N. (2025). Machine learning determines the incidence of Alzheimer's disease based on population gut microbiome profile. *Brain Communications*, 7(2), fcac059. <https://doi.org/10.1093/braincomms/fcac059>
- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., ... Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8), 852–857. <https://doi.org/10.1038/s41587-019-0209-9>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581–583. <https://doi.org/10.1038/nmeth.3869>
- Cesaro, A., Broset, E., Salimando, G. J., Öztekin, S., Mai, D., Bennett, M. L., & Fuente-Nunez, C. de la. (2025). Using synthetic biology to understand the gut-brain axis. *Cell Reports Physical Science*, 6(3). <https://doi.org/10.1016/j.xcrp.2025.102458>
- Chen, G.-F., Xu, T.-H., Yan, Y., Zhou, Y.-R., Jiang, Y., Melcher, K., & Xu, H. E. (2017). Amyloid beta: Structure, biology and structure-based therapeutic development. *Acta Pharmacologica Sinica*, 38(9), 1205–1235. <https://doi.org/10.1038/aps.2017.28>
- Chin, V. K., Yong, V. C., Chong, P. P., Amin Nordin, S., Basir, R., & Abdullah, M.

- (2020). Mycobiome in the Gut: A Multiperspective Review. *Mediators of Inflammation*, 2020, 9560684. <https://doi.org/10.1155/2020/9560684>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>
- Cryan, J. F., O’Riordan, K. J., Cowan, C. S. M., Sandhu, K. V., Bastiaanssen, T. F. S., Boehme, M., Codagnone, M. G., Cussotto, S., Fulling, C., Golubeva, A. V., Guzzetta, K. E., Jaggar, M., Long-Smith, C. M., Lyte, J. M., Martin, J. A., Molinero-Perez, A., Moloney, G., Morelli, E., Morillas, E., ... Dinan, T. G. (2019). The Microbiota-Gut-Brain Axis. *Physiological Reviews*, 99(4), 1877–2013. <https://doi.org/10.1152/physrev.00018.2018>
- Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, 6(1), 226. <https://doi.org/10.1186/s40168-018-0605-2>
- Dominguez-Gortaire, J., Ruiz, A., Porto-Pazos, A. B., Rodriguez-Yanez, S., & Cedron, F. (2025a). Alzheimer’s Disease: Exploring Pathophysiological Hypotheses and the Role of Machine Learning in Drug Discovery. *International Journal of Molecular Sciences*, 26(3), 1004. <https://doi.org/10.3390/ijms26031004>
- Dominguez-Gortaire, J., Ruiz, A., Porto-Pazos, A. B., Rodriguez-Yanez, S., & Cedron, F. (2025b). Alzheimer’s Disease: Exploring Pathophysiological Hypotheses and the Role of Machine Learning in Drug Discovery. *International Journal of Molecular Sciences*, 26(3), 1004. <https://doi.org/10.3390/ijms26031004>
- Donovan, P. D., Gonzalez, G., Higgins, D. G., Butler, G., & Ito, K. (2018). Identification of fungi in shotgun metagenomics datasets. *PloS One*, 13(2), e0192898. <https://doi.org/10.1371/journal.pone.0192898>
- Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, 35(5–6), 352–359. [https://doi.org/10.1016/s1532-0464\(03\)00034-0](https://doi.org/10.1016/s1532-0464(03)00034-0)
- Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., & Weyrich, L. S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, 27(2), 105–117. <https://doi.org/10.1016/j.tim.2018.11.003>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, ROC Analysis in Pattern Recognition, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Ferreiro, A. L., Choi, J., Ryou, J., Newcomer, E. P., Thompson, R., Bollinger, R.

M., Hall-Moore, C., Ndao, I. M., Sax, L., Benzinger, T. L. S., Stark, S. L., Holtzman, D. M., Fagan, A. M., Schindler, S. E., Cruchaga, C., Butt, O. H., Morris, J. C., Tarr, P. I., Ances, B. M., & Dantas, G. (2023). Gut microbiome composition may be an indicator of preclinical Alzheimer's disease. *Science Translational Medicine*, 15(700), eabo2984. <https://doi.org/10.1126/scitranslmed.abo2984>

Forbes, J. D., Bernstein, C. N., Tremlett, H., Van Domselaar, G., & Knox, N. C. (2019). A Fungal World: Could the Gut Mycobiome Be Involved in Neurological Disease? *Frontiers in Microbiology*, 9, 3249. <https://doi.org/10.3389/fmicb.2018.03249>

Gaspar, B. S., Roșu, O. A., Enache, R.-M., Manciușlea Profir, M., Pavelescu, L. A., & Crețoiu, S. M. (2025). Gut Mycobiome: Latest Findings and Current Knowledge Regarding Its Significance in Human Health and Disease. *Journal of Fungi (Basel, Switzerland)*, 11(5), 333. <https://doi.org/10.3390/jof11050333>

Hadrich, I., Turki, M., Chaari, I., Abdelmoula, B., Gargouri, R., Khemakhem, N., Elatoui, D., Abid, F., Kammoun, S., Rekik, M., Aloulou, S., Sehli, M., Mrad, A. B., Neji, S., Feiguin, F. M., Aloulou, J., Abdelmoula, N. B., & Sellami, H. (2024). Gut mycobiome and neuropsychiatric disorders: Insights and therapeutic potential. *Frontiers in Cellular Neuroscience*, 18, 1495224. <https://doi.org/10.3389/fncel.2024.1495224>

Heneka, M. T., Carson, M. J., El Khoury, J., Landreth, G. E., Brosseron, F., Feinstein, D. L., Jacobs, A. H., Wyss-Coray, T., Vitorica, J., Ransohoff, R. M., Herrup, K., Frautschy, S. A., Finsen, B., Brown, G. C., Verkhratsky, A., Yamanaka, K., Koistinaho, J., Latz, E., Halle, A., ... Kummer, M. P. (2015). Neuroinflammation in Alzheimer's disease. *The Lancet. Neurology*, 14(4), 388–405. [https://doi.org/10.1016/S1474-4422\(15\)70016-5](https://doi.org/10.1016/S1474-4422(15)70016-5)

Huseyin, C. E., Rubio, R. C., O'Sullivan, O., Cotter, P. D., & Scanlan, P. D. (2017). The Fungal Frontier: A Comparative Analysis of Methods Used in the Study of the Human Gut Mycobiome. *Frontiers in Microbiology*, 8. <https://doi.org/10.3389/fmicb.2017.01432>

Jung, J. H., Kim, G., Byun, M. S., Lee, J. H., Yi, D., Park, H., Lee, D. Y., & Group, for the K. R. (2022). Gut microbiome alterations in preclinical Alzheimer's disease. *PLOS ONE*, 17(11), e0278276. <https://doi.org/10.1371/journal.pone.0278276>

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, 3149–3157. <https://dl.acm.org/doi/10.5555/3294996.3295074>

Knopman, D. S., Amieva, H., Petersen, R. C., Chételat, G., Holtzman, D. M.,

- Hyman, B. T., Nixon, R. A., & Jones, D. T. (2021). Alzheimer disease. *Nature Reviews. Disease Primers*, 7(1), 33. <https://doi.org/10.1038/s41572-021-00269-y>
- Ko, Y. K., Kim, E., Lee, E.-J., Nam, S. J., Kim, Y., Kim, S., Choi, S.-Y., Kim, H. Y., & Choi, Y. (2024). Enrichment of infection-associated bacteria in the low biomass brain bacteriota of Alzheimer's disease patients. *PLOS ONE*, 19(2), e0296307. <https://doi.org/10.1371/journal.pone.0296307>
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'95*, 1137–1143.
- Köljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., Bates, S. T., Bruns, T. D., Bengtsson-Palme, J., Callaghan, T. M., Douglas, B., Drenkhan, T., Eberhardt, U., Dueñas, M., Grebenc, T., Griffith, G. W., Hartmann, M., Kirk, P. M., Kohout, P., ... Larsson, K.-H. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, 22(21), 5271–5277. <https://doi.org/10.1111/mec.12481>
- Kumar, D., Gupta, S. M., Chauhan, N. S., & Shah, K. (Eds.). (2025). *Alzheimer's Disease: Diagnosis, Molecular Targets, and Therapeutic Modalities*. CRC Press. <https://doi.org/10.1201/9781003602477>
- Li, A.-H., Yuan, F.-X., Groenewald, M., Bensch, K., Yurkov, A. M., Li, K., Han, P.-J., Guo, L.-D., Aime, M. C., Sampaio, J. P., Jindamorakot, S., Turchetti, B., Inacio, J., Fungsin, B., Wang, Q.-M., & Bai, F.-Y. (2020). Diversity and phylogeny of basidiomycetous yeasts from plant leaves and soil: Proposal of two new orders, three new families, eight new genera and one hundred and seven new species. *Studies in Mycology*, 96, 17–140. <https://doi.org/10.1016/j.simyco.2020.01.002>
- Liang, Y., Liu, C., Cheng, M., Geng, L., Li, J., Du, W., Song, M., Chen, N., Yeleen, T. A. N., Song, L., Wang, X., Han, Y., & Sheng, C. (2024). The link between gut microbiome and Alzheimer's disease: From the perspective of new revised criteria for diagnosis and staging of Alzheimer's disease. *Alzheimer's & Dementia*, 20(8), 5771–5788. <https://doi.org/10.1002/alz.14057>
- Ling, Z., Zhu, M., Liu, X., Shao, L., Cheng, Y., Yan, X., Jiang, R., & Wu, S. (2021). Fecal Fungal Dysbiosis in Chinese Patients With Alzheimer's Disease. *Frontiers in Cell and Developmental Biology*, 8. <https://doi.org/10.3389/fcell.2020.631460>
- Liu, X., Jiang, B., Hao, H., & Liu, Z. (2022). CARD9 Signaling, Inflammation, and Diseases. *Frontiers in Immunology*, 13. <https://doi.org/10.3389/fimmu.2022.880879>
- Luangsa-ard, J., Houbraken, J., van Doorn, T., Hong, S.-B., Borman, A. M., Hywel-Jones, N. L., & Samson, R. A. (n.d.). *Purpureocillium*, a new genus for the

medically important *Paecilomyces lilacinus*. Retrieved May 26, 2026, from <https://dx.doi.org/10.1111/j.1574-6968.2011.02322.x>

Mohammadi, H., & Karwowski, W. (2024). Graph Neural Networks in Brain Connectivity Studies: Methods, Challenges, and Future Directions. *Brain Sciences*, 15(1), 17. <https://doi.org/10.3390/brainsci15010017>

Nagpal, R., Neth, B. J., Wang, S., Mishra, S. P., Craft, S., & Yadav, H. (2020). Gut mycobiome and its interaction with diet, gut bacteria and alzheimer's disease markers in subjects with mild cognitive impairment: A pilot study. *eBioMedicine*, 59, 102950. <https://doi.org/10.1016/j.ebiom.2020.102950>

Naik, B., Sasikumar, J., & Das, S. P. (2025). From Skin and Gut to the Brain: The Infectious Journey of the Human Commensal Fungus *Malassezia* and Its Neurological Consequences. *Molecular Neurobiology*, 62(1), 533–556. <https://doi.org/10.1007/s12035-024-04270-w>

Nandal, P., Dahiya, M., Singh, M., Dagur, A., & Kumar, B. (Eds.). (2025). *Progressive Computational Intelligence, Information Technology and Networking*. CRC Press. <https://doi.org/10.1201/9781003650010>

Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., & Tedersoo, L. (2019). Mycobiome diversity: High-throughput sequencing and identification of fungi. *Nature Reviews. Microbiology*, 17(2), 95–109. <https://doi.org/10.1038/s41579-018-0116-y>

Otoo, H. N., Lee, K. G., Qiu, W., & Lipke, P. N. (2008). *Candida albicans* Als adhesins have conserved amyloid-forming sequences. *Eukaryotic Cell*, 7(5), 776–782. <https://doi.org/10.1128/EC.00309-07>

Phuna, Z. X., & Madhavan, P. (2022). A closer look at the mycobiome in Alzheimer's disease: Fungal species, pathogenesis and transmission. *The European Journal of Neuroscience*, 55(5), 1291–1321. <https://doi.org/10.1111/ejn.15599>

Pisa, D., Alonso, R., Juarranz, A., Rábano, A., & Carrasco, L. (2015). Direct visualization of fungal infection in brains from patients with Alzheimer's disease. *Journal of Alzheimer's Disease: JAD*, 43(2), 613–624. <https://doi.org/10.3233/JAD-141386>

Richard, M. L., & Sokol, H. (2019). The gut mycobiota: Insights into analysis, environmental interactions and role in gastrointestinal diseases. *Nature Reviews. Gastroenterology & Hepatology*, 16(6), 331–345. <https://doi.org/10.1038/s41575-019-0121-2>

Rutsch, A., Kantsjö, J. B., & Ronchi, F. (2020). The Gut-Brain Axis: How Microbiota and Host Inflammasome Influence Brain Physiology and Pathology.

- Frontiers in Immunology, 11, 604179. <https://doi.org/10.3389/fimmu.2020.604179>
- Sait, A., Angeli, C., Doig, A. J., & Day, P. J. R. (2021). Viral Involvement in Alzheimer's Disease. *ACS Chemical Neuroscience*, 12(7), 1049–1060. <https://doi.org/10.1021/acscchemneuro.0c00719>
- Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12, 87. <https://doi.org/10.1186/s12915-014-0087-z>
- Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Fungal Barcoding Consortium, & Fungal Barcoding Consortium Author List. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, 109(16), 6241–6246. <https://doi.org/10.1073/pnas.1117018109>
- Selkoe, D. J., & Hardy, J. (2016). The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Molecular Medicine*, 8(6), 595–608. <https://doi.org/10.15252/emmm.201606210>
- Shao, K., Yang, Y., Gong, X., Chen, K., Liao, Z., & Ojha, S. C. (2025). Staphylococcal Drug Resistance: Mechanisms, Therapies, and Nanoparticle Interventions. *Infection and Drug Resistance*, 18, 1007–1033. <https://doi.org/10.2147/IDR.S510024>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Subedi, S., Sasidharan, S., Nag, N., Saudagar, P., & Tripathi, T. (2022). Amyloid Cross-Seeding: Mechanism, Implication, and Inhibition. *Molecules*, 27(6), 1776. <https://doi.org/10.3390/molecules27061776>
- Tedersoo, L., Tooming-Klunderud, A., & Anslan, S. (2018). PacBio metabarcoding of Fungi and other eukaryotes: Errors, biases and perspectives. *The New Phytologist*, 217(3), 1370–1385. <https://doi.org/10.1111/nph.14776>
- Tiew, P. Y., Mac Aogain, M., Ali, N. A. B. M., Thng, K. X., Goh, K., Lau, K. J. X., & Chotirmall, S. H. (2020). The Mycobiome in Health and Disease: Emerging Concepts, Methodologies and Challenges. *Mycopathologia*. <https://doi.org/10.1007/s11046-019-00413-z>
- Van Hul, M., Cani, P. D., Petitfils, C., De Vos, W. M., Tilg, H., & El-Omar, E. M. (2024). What defines a healthy gut microbiome? *Gut*, 73(11), 1893–1908.

<https://doi.org/10.1136/gutjnl-2024-333378>

Velasquez, P., & Oudah, M. (2023). Machine Learning Based Screening Tool for Alzheimer's Disease via Gut Microbiome. In M. Rocha, F. Fdez-Riverola, M. S. Mohamad, & A. B. Gil-González (Eds.), *Practical Applications of Computational Biology and Bioinformatics, 17th International Conference (PACBB 2023)* (pp. 62–72). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-38079-2_7

Vogt, N. M., Kerby, R. L., Dill-McFarland, K. A., Harding, S. J., Merluzzi, A. P., Johnson, S. C., Carlsson, C. M., Asthana, S., Zetterberg, H., Blennow, K., Bendlin, B. B., & Rey, F. E. (2017). Gut microbiome alterations in Alzheimer's disease. *Scientific Reports*, 7(1), 13537. <https://doi.org/10.1038/s41598-017-13601-y>

Ye, T., Yuan, S., Kong, Y., Yang, H., Wei, H., Zhang, Y., Jin, H., Yu, Q., Liu, J., Chen, S., & Sun, J. (2022). Effect of Probiotic Fungi against Cognitive Impairment in Mice via Regulation of the Fungal Microbiota-Gut-Brain Axis. *Journal of Agricultural and Food Chemistry*, 70(29), 9026–9038. <https://doi.org/10.1021/acs.jafc.2c03142>

Yetgin, A. (2024). Exploring the Link Between the Gut Mycobiome and Neurological Disorders. *Advanced Gut & Microbiome Research*, 2024(1), 9965893. <https://doi.org/10.1155/2024/9965893>

Yuan, C., He, Y., Xie, K., Feng, L., Gao, S., & Cai, L. (2023). Review of microbiota gut brain axis and innate immunity in inflammatory and infective diseases. *Frontiers in Cellular and Infection Microbiology*, 13, 1282431. <https://doi.org/10.3389/fcimb.2023.1282431>

Zhang, F., Aschenbrenner, D., Yoo, J. Y., & Zuo, T. (2022). The gut mycobiome in health, disease, and clinical applications in association with the gut bacterial microbiome assembly. *The Lancet Microbe*, 3(12), e969–e983. [https://doi.org/10.1016/S2666-5247\(22\)00203-8](https://doi.org/10.1016/S2666-5247(22)00203-8)

Zheng, Q., & Wang, X. (2025). Alzheimer's disease: Insights into pathology, molecular mechanisms, and therapy. *Protein & Cell*, 16(2), 83–120. <https://doi.org/10.1093/procel/pwae026>



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daultpur, Main Bawana Road, Delhi-110042

PLAGARISM VERIFICATION

Title of the Thesis: **A Machine Learning Framework for the Detection of Alzheimer's Disease Using Gut Fungal ITS Sequencing Data**

Total Pages: **64**

Name of the Student: **Mayank Singh**

Supervisor: **Prof. Yasha Hasija**

Department of Biotechnology, Delhi Technological University, Delhi-110042

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software Used: **Turnitin**, Similarity Index: **10%**, Total Word Count: **17114**

Date: **28/05/2026**

A handwritten signature in black ink, appearing to read "Mayank Singh", written over a horizontal line.

Candidate's Signature

A handwritten signature in black ink, appearing to read "Ym", written over a horizontal line. Below the signature, the date "29.05.26" is written.

Signature of Supervisor