Application of Supervised Learning Classifiers on Gut Microbial data to Predict Parkinson Disease

MSc Research Project
Data Analytics

Madhuri Dhatrak
Student ID: 21208808

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Madhuri Dhatrak
Student ID:	21208808
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Qurrat Ul Ain
Submission Due Date:	14/08/2023
Project Title:	Application of Supervised Learning Classifiers on Gut Microbial data to Predict Parkinson Disease
Word Count:	5998
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Application of Supervised Learning Classifiers on Gut Microbial data to Predict Parkinson Disease

Madhuri Dhatrak
21208808

Abstract

The motor and non-motor symptoms of Parkinson's disease (PD) affect millions of individuals worldwide, making it a significant public health concern on a global scale. If the illness is discovered and treated when still in the prodromal stage, the severity of the ailment may be lessened. This investigation makes use of machine learning predictive technologies to investigate the compositions of the gut microbiomes and its alterations in Parkinson Diseases Patients. Early intervention during the prodromal stage is made possible by detection of these alterations in the gut compositions which may allow for customized treatment approaches and a fresh perspective on the treatment of this condition. This study compares the microbiome diversity between cohorts with healthy and PD-affected individuals and assesses several supervised learning models for predictive classification in order to investigate the relationship between gut microbiota and Parkinson's disease (PD). For people who are affected by this disorder, the goal is to provide hope and better outcomes. These supervised predictive analyses have the potential to revolutionize PD treatment and improve patients' quality of life.

1 Introduction

Millions of individuals worldwide suffer from Parkinson's disease (PD), which has both severe motor and non-motor symptoms and is a critical and urgent global health issue. It belongs to the most widespread group of neurodegenerative diseases. Parkinson's disease (PD), which is the second most common form of neurodegenerative disease, has a profound effect on patients and their families, carers, and the healthcare system as a whole. It is important to emphasize that this illness prevalence has increased over the past 25 years and has resulted in 5.8 million DALYs (disability adjusted life years), an increase of 81% since 2000 [1] . Enhancing PD management and control is crucial, especially given the population's propensity to age.

1.1 Background & Motivation

PD, a complicated condition impacted by a number of environmental, behavioural, and genetic factors, currently has no known cure; the only treatment available is symptomatic relief. Patients with PD may exhibit motor symptoms like as tremors and stiffness in addition to non-motor symptoms such as depression, cognitive decline, and sleep problems.

[1] https://www.who.int/news-room/fact-sheets/detail/parkinson-disease

It is possible that in PD patients have both motor and non-motor symptoms. Researchers have recently paid a lot of attention to the prodromal phase, which occurs prior to the onset of PD. In PD patients non-motor symptoms typically appear years before the conventional motor symptoms during the prodromal period. Constipation, hyposmia, potential REM-sleep behaviour disorder (RBD), sadness, anxiety, and cognitive impairment are among these symptoms. Early intervention during this stage may present opportunities for neuroprotective treatments. However, prodromal PD symptoms are not exclusive to PD and can be brought on by other neurological disorders. As a result, it is unclear that prodromal PD may be determined using only one prodromal marker or bio marker. Alternatively, the co-occurrence of several indicators (bio markers) in a single person may be a more accurate sign of prodromal PD. A study by Longitudinal Aging research Amsterdam Roos et al. (2022) in late middle-aged participants, evaluated the prevalence of constipation, hyposmia, potential RBD, depression, anxiety, and cognitive impairment. Another such study suggests that prodromal stages of the disease, specifically when RBD is emerging, are characterized by gut dysbiosis comparable to that seen in PD. This study gives valuable insight on the complex relationships between gut health and neurological diseases, which has significance for understanding the possible causes of PD and its diagnostic procedures Huang et al. (2023). Recent advances in machine learning and AI have increased the ability to identify disorders in groups at risk at an early stage. Early identification and a precise prediction of the disease were made possible by the predictive potential of these approaches and to better understand the relationship between genetic predisposition, environmental factors, and the initial onset of the PD symptoms. So, a proactive treatment approach is now possible in PD instead of a reactive approach because of these complex predictive models that can generate accurate predictions. This paves the way for PD patients to obtain individualized care and have better health outcomes. This study makes use of the one-of-a-kind prodromal phase potential, which is a bidirectional relation between gut microbiome and PD , to offer improved outcomes and hope to those who are affected with PD Hey et al. (2023).

1.2 Research Question

Based on the above research analysis this paper attempts to address the following research question. **RQ:** "To what extent does the application of supervised machine learning algorithms using gut microbiota data, classify and contribute to the accuracy of predicting Parkinson's disease?" Utilizing machine learning classification models on examined gut microbiome compositions in biological sequence data to determine PD.

1.3 Research Objective and Contribution

To address above question, the following specific sets of research objectives were derived:

1. To investigate the potential relationship between gut microbiota and Parkinson's disease and compare the microbiome diversities of healthy and diseased cohorts.

2. To explore the feasibility of utilizing machine learning classification models in predicting Parkinson's disease based on gut microbiota compositions.

3. To evaluate the predictive accuracy and reliability of supervised learning classifier models in diagnosing Parkinson's disease using gut microbiota as input features.

4. To compare and analyse the performance of different supervised learning classifiers in utilizing gut microbiota data for Parkinson's disease prediction.

The purpose of this project is to apply and analyse supervised learning predictive model that uses gut microbiome data and its role in detecting PD at the prodromal stage, which may allow medical professionals to intervene at an earlier stage and provide patients with more individualized treatment plans. If treatment intervention was focused on the pre-motor phases of the PD, the level of quality life of PD patients would be significantly enhanced and extended. This could be able to slow down or prevent the spread of disease to the brain.

Following is the structure of this technical paper; the various research publications in this area will be covered in depth in the chapter 2. In the chapter 3 on Research methodology discuses about the approach or various steps that is used on this study. In the chapter 4 on Design Specification & Implementation discusses about the framework that is used in predicting PD and in the chapter 5, Evaluation discusses about the outcomes and analysis and final chapter gives the conclusion and discussion.

2 Literature Review

A significant amount of research has been done on the subject of examining the changes in gut bacteria in PD patients. The relevant study in this area can be presented in two different ways: first, by using bioinformatics or statistical analysis to examine the gut microbiota of PD patients, and second, by applying these findings to predictive model like Artificial Intelligence (AL), deep learning (DL) and machine learning (ML) that can assist with early diagnosis, classification, and access PD biomarkers.

2.1 Bioinformatic and statistical methods to examine gut microbes in PD patients

The field of bioinformatics integrates software tools, statistical methods, mathematical concepts, and biotechniques to store, examine, and to integrate enormous and complex data sets.A range of bioinformatic methods like metagenomics, sequencing based approach have been investigated to analyse and assess the gut microbiome's composition.

To begin with the study scientist have explored the complex composition of the gut microbiome of PD patients using models of animals. It revealed the profound influence of the gut microbiome on PD. These gut microbes had the power to regulate microglia activation and aggregation of alpha-synuclein (Syn) which are the two key players in the characteristic motor impairments in PD patients. In germ-free mice, bacteria-generated short-chain fatty acids (SCFAs) could reestablish the disease's characteristics, and in humanized mice with Parkinson's-associated microbiota, motor symptoms worsened. A microbial imbalance known as dysbiosis may be brought on by a variety of things, including changes in physiological processes and pesticide exposure. This finding implied the therapeutic potential of microbiome-based treatment and suggested a compelling connection between gut health and brain disorders. This gives an important findings and sheds light on the complex connection between PD and the gut microbiota Sampson et al. (2016).

Now that the studies revealed the bidirectional connection between the gut microbes and PD the research journey continues on exploring the specific taxonomy levels of bacterial groups and diversity of gut microbiomes which are actually associated with PD. Lachnospiraceae, a type of gut bacteria, is linked to the synthesis of short chain fatty acids (SCFA), which could be the possibility of SCFA deficiency in PD patients. Apart from SCFA changes, emphasis on alterations in bacterial families like Akkermansia, Lactobacillus, and Bifidobacterium has been linked to PD. The identification of altered SCFA in these individuals laid a strong platform for further investigation into their function in gut inflammation, microglial cell activation, and gastrointestinal symptoms associated with PD. But also various studies proved the influence of sex, age, and dietary habits, geographical differences, environmental, behavioural on the microbiome diversities which added layers of complexity to these understanding Hill-Burns et al. (2017).In addition to altered gut microbiomes bioinformatic methods like shotgun metagenomics in multi-cohort studies also shown that in PD patients there are significant disturbances in how these microbe's function. These disturbances are related to the metabolism of key nutrients, essential microbial mobility, community signalling and responses to oxidative stress Boktor et al. (2023).

These investigations on the intricate interactions between PD and gut microbiota (GM) dysbiosis opened up novel opportunities for several innovative diagnostic and treatment possibilities. These findings showed PD patients needed a healthy gut microbiome which opened the door to possible therapies like probiotics and fecal microbial transplantation (FMT). These therapeutic trials showed a two-way link between the activation of dopamine agonist therapies (DATs) and the GM. However, there are inconsistencies in PD-GM studies that necessitate the use of standardized methods and comparisons among various disease stages. The question DAT activation and GM is still in its early stages, with much more to uncover Hey et al. (2023).

The studies mentioned above demonstrate that gut microbiota is a useful tool for PD diagnosis and treatment and that further study is needed in this area. However, analysing huge datasets of information on the human microbiome and identifying patterns to forecast disorders are difficult by using conventional bioinformatic or statistical techniques. Metagenomics uses cutting-edge technology like machine learning and artificial intelligence to overcome these challenges. These prediction models assisted researchers to detect, analyse, and anticipate certain diseases based on the medical history and underlying biological traits of patients. The subsequent section discusses useful information about several recent research that used predictive algorithms to forecast diseases using the gut microbiome data sets.

2.2 Accessing PD biomarkers using machine learning (ML) and deep learning (DL)

The use of AL, DL and ML has recently emerged as one of the most important methods for predicting Parkinson's disease because their capacity to analyse large, complex data sets, finding trends and biomarkers. These prediction models can aid in early diagnosis and categorization, allowing for the development of individualized treatment options.

It proved beneficial to employ ML-based image applications in neuroimaging studies because they made it feasible to automatically identify PD at an early stage so that patients can get treatment to limit disease progression. ML-based SPECT image processing outperforms traditional analysis in identifying PD related degeneration of dopamine. It is equivalent to a skilled visual evaluation and aids radiologists in making more precise diagnoses of PD. Although there are some challenges in images datasets such as inaccurate image extraction of features these methods have produced encouraging results and helped medical practitioners to diagnose Parkinsonism and enhanced the early detection of PD which reduced the error rate in detecting Zhang (2022). Supervised machine learning algorithms like Support Vector Machine (SVM) and k-nearest neighbours (KNN) classified and revealed major variations observed in electroencephalogram (EEG) signals between healthy people and PD patients in almost all areas of the brain and they have the potential to reduce the amount of PD misdiagnoses by identifying the condition and classifying at a premotor stage, facilitating the provision of early therapies like neuroprotective measures Coelho et al. (2023).

Along with extensive research utilizing ML/DL/AI models for processing brain imaging and analysis with large datasets for PD diagnosis, these techniques have shown similarly amazing results when applied to metagenomic data to achieve taxonomy classification. These prediction models provide a number of benefits when used with metagenomic data such as integrating different omics data, pattern recognition by utilizing certain genes or similar taxonomic features to find genetic relationships whereas analysing large and complex biological sequencing data from various environmental samples is challenging with traditional bioinformatics tools. Recent studies developed a deep learning-based classification method for 16S short read sequences that is based on k-mer representation that allowed each taxonomic category i.e., from phylum to genus to produce its own classification model. The findings of the tests confirmed how well suggested pipeline classified bacterial genomes accurately. These techniques have been integrated into popular metagenomic data analysis tools. According to the findings, this method was successfully classifying the data from both 16S shotgun (SG) and amplicon (AMP) devices Fiannaca et al. (2018).

The results from the study Bang et al. (2019) showed accurate prediction and classification of diseases based on human gut microbial data by machine learning algorithms which resulted in major improvements in disease diagnosis, prognosis, and individualized treatment, as well as developments in medical research and practice. Three factors—taxonomy level, classifier preference, and feature selection methods—are used to classify various diseases, including juvenile idiopathic arthritis (JIA), chronic fatigue syndrome (CFS), multiple sclerosis (MS), and colorectal cancer (CRC). ML models have successfully identified microbes connected to these disorders, indicating the possibility of these microbes to serve as biomarkers for detection and categorization. Several characteristics, such as the existence of PSBM3 (a genus in the Family Erysipelotrichaceae), were significant in numerous feature subsets, demonstrating their applicability in the identification of various diseases.American Gut Project applied supervised ML models like Random Forest (RF), Support Vector Machines (SVM), and Logistic regression on human gut data samples to identify the coronary artery disease (CAD) based on interactions between the gut microbiomes. Prediction algorithms were compared with and without interactions between food and gut microbiome and surprisingly, models performed better with diet and gut microbe included as primary parameters. Study Vilne et al. (2022) found that integrating DL/ML algorithms and taking the diet-gut microbiota into account could increase the precision of disease risk prediction, especially for complicated disorders like CAD. According to the case study, further study with larger amounts of data and more different population groups are needed to substantiate the findings.

Studies Liang et al. (2022) have identified that gut microbiome characteristics linked to cancer patients' responses to immunotherapy. To discover taxa and microbial interactions connected to response, investigated numerous 16S rRNA gene sequencing datasets, integrating a new cohort with the available data, and using a variety of analytical tools, including univariate analysis and a novel technique called selbal. To anticipate reaction based on taxonomic traits, they created statistical models utilizing machine learning algorithms such as LogitBoost and KNN. These models had good accuracy and were tested on various platforms for sequencing. The study Tabl et al. (2019) provides information on potential microbiome-based indicators for the effectiveness of immunotherapy. Machine learning algorithms are utilized to identify gene biomarkers for breast cancer survival and to enhance treatment choices and clinical outcomes. The gene activity of patients who had various therapies and either lived or died was evaluated using a model that employs a hierarchy of trees for classification. The model offers a high degree of accuracy in predicting the result of survival based on gene expression. The identified biomarkers may help physicians decide what treatments to be provided for the cancer patients to improve the disease condition.

A reliable and precise classifier for diagnosing individuals with constipation has been developed in a research project using machine learning to analyse gut microbiome sequence data samples of constipation patients from the American Gut Project. The model was created using cross-validation and a new cohort was used to validate the model to increasing its reliability. Feature-selection techniques were employed to increase prediction accuracy and streamline calculation. It was shown that gradient boosting regression trees (GBRT), chi-square, and logistic regression were all effective for data discretization and feature selection. Serratia, Dorea, Aeromonas, and Hungatella were discovered to be potential major contributors to the constipation-prediction model, which are among the high-ranking microbial indicator. This illustrates the importance of microbiome analysis and machine learning for identifying constipation-related symptoms and offers suggestions for future work on non-invasive tests and microbial treatments Chen et al. (2021).

Using machine learning algorithms Random Forest technique, which was used to examine by integrating many data layers, the relationship between the gut microbiota and Grave's disease (GD), a thyroid condition, was investigated. The model incorporated four taxonomic species-level relative abundance profiles, which showed high sensitivity in distinguishing GD from other metabolic diseases and may be helpful for clinical diagnosis Zhu et al. (2021). Three machine learning classifiers were developed to assess the prevalence of the brain disorder schizophrenia: Random Forest (RF), Support Vector Machine (SVM), and XGBoost (XGB). Two feature tables are used to these classifiers: one contains Amplicon Sequence Variants (ASVs), while the other has the table compressed at the genus level. These tables were adjusted per sample and converted into relative abundance tables to address the compositional nature of the microbiome data. The machine learning classification process is utilized to obtain the schizophrenia diagnosis utilizing the features from these tables as inputs. The trained model showed accurate predictions of the schizophrenia diagnosis based on the input characteristics of a specific sample Wang et al. (2023). Out of 10 machine learning classifiers used to categorize mild cognitive impairment and Alzheimer's disease, logistic regression yielded the best results. However, the study does have certain limitations, including the use of 16S rRNA data and the relatively small sample size, which could affect how precisely species and families are categorized. The study suggests that with increased access to metagenomics samples, classification algorithms may be improved and a deeper analysis of the function of the gut microbiome in various disorders may be made possible Hasic Telalovic et al. (2022). Six datasets were utilized to use three ML algorithms (RF, SVM, and NN) to identify between healthy controls (HC) and PD patients based on a variety of parameters. The

Random Forest (RF) approach discovered 22 bacterial families and delivered the greatest results. According to the results of this study, gut dysbiosis in PD is caused by a complex interaction between several bacterial families. During the examination, it was found that several of the observed bacterial families had not before been documented in the literature. These findings highlight the capability of ML systems to offer novel insights into the role played by certain bacterial families in PD status prediction. The study Pietrucci et al. (2020) discovered methodological differences in the collection, storage, and analysis of data across various laboratories, which may have affected the results' variability in PD case-control studies of the gut microbiota. The authors stressed the need for consistent processes in order to produce reliable assessments.

However, with every research that was conducted, gained more knowledge and understanding of the complex link between gut dysbiosis and PD prediction.Supervised Machine learning classifier algorithms have demonstrated to be great tools for analysing gut microbiome data in predicting Parkinson disorders at an early stage, and assisting researchers in creating innovative treatments.

3 Research Methodology

Knowledge discovery in databases (KDD) will be suitable for this project, as its main objective is to determine the changed gut microbiome composition and underlying causes of an imbalance in individuals with Parkinson's disease (PD).With the help of this methodology, it is possible to regularly identify significant, reliable, and clear patterns in data samples that are large and complex.Data selection,pre-processing or cleaning,Data Transformation,Taxonomy Analysis,Data Model, Evaluations of the model,these steps are carried out in this study to draw conclusions.Figure 1 gives the research methodology process workflow for this study.

3.1 Data Collection

From European Nucleotide Archive (ENA) Database, a data set with 17 randomly chosen fecal samples of parkinson and healthy subjects is chosen under the project number PRJEB27564[2] . Of these 17 samples, 9 correspond to Parkinson's patients (PD) and 8 to Healthy controls (HC). This bio project Aho et al. (2019) collected clinical information and stool samples twice throughout the course of, on average, 2.25 years from PD and HC subjects. On the Illumina MiSeq platform, amplicon sequencing of the 16S rRNA gene's V3-V4 variable regions was used to analyse microbial populations in these samples.

3.2 Data Quality Check

These downloaded 17 samples are in Pair end layout with forward and reverse read files in FastQ file format with biological sequence reads with respective quality scores. FastqC tool [3] was used to evaluate the quality of the these sequencing read files.

[2] https://www.ebi.ac.uk/ena/browser/view/PRJEB27564

[3] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

3.3 Data Transformation

The open-source bioinformatics application QIIME2 Bolyen et al. (2019) is mostly used for data transformation in this work. The pair end read sample files are combined using this program, and sequences were denoised Bokulich et al. (2013) using the Deblur method, which includes abundance filtering and internal chimera checking mechanisms. The sequences are organized into Amplicon Sequence Variants (ASVs) at the final stage of this data transformation. This results in the feature table, representative sequences, and statistics. These artifacts hold the key to the intended follow-up investigation on this study.

3.4 Data Analysis (Taxonomy Analysis)

At this phase taxonomy analysis is studied which gives the microbe abundance in each of the samples and its characteristics. A pre-trained Naive Bayes classifier, which is accessible on QIIME 2, is used for this analysis. Two significant databases , SILVA and Green Genes2, served as the training components for this Naive Bayes classifier. In this study, the SILVA (silva13899nbclassifier) and the most recent Green Genes2 (gg202210-backbonefulllength) classifiers are used for taxa analysis of the samples Resources (2023).

3.5 Model

In order to predict the patient's risk of having Parkinson's disease, the samples in this study were analysed using machine learning algorithms. A sample classifier plugin for QIIME 2 was created utilizing the Sci-Kit Learn package for supervised learning classifiers and feature selection, SciPy for statistics, and Sea born for data visualization classifier (2023). In this case, the model receives the feature table and labelled sample meta data files and divides the samples into train and test groups at random. It then divides the samples into two classes, Parkinson Disease (PD) and Healthy Control (HC), according to the meta data. The arguments to the Model include the number of K-fold cross validations, the size of the data samples, automatic feature selection and parameter tuning, and the various supervised learning classifier estimators are used .In this study two classifier algorithms Random Forest (RF) and Linear SupportVectorClassifier (LSVC) are used for prediction Model.

3.5.1 RF

This algorithm uses averaging to boost predicted accuracy and decrease over fitting after fitting numerous decision tree classifiers to various data set sub samples.These are some of the algorithm's parameters that on high level discussed, the size of the sub samples is decided by the argument max samples, and if bootstrap is set to default, each of tree is built using the whole data set. The number of decision trees is controlled by n estimators, which by default is 100. The max features parameter sets the optimal number of features to split; it takes every feature into account by default Pedregosa et al. (2011) RandomForest (2023).

3.5.2 LSVC

This is Support Vector Classifier but with the parameter kernel is set to 'linear', gives more freedom in the selection of penalties and loss functions and making it more likely to scale to more numbers of samples. This accepts both spare and dense data, and the multi class support is handled via one vs rest strategy Pedregosa et al. (2011) LinearSVC (2023).

3.6 Model Evaluation Metrics

Accuracy, Precision, Sensitivity, Specificity and F1 score are the metrics used to evaluate each supervised classification model. Heat Maps, ROC curves, and confusion matrix are used to compare and display the models. The performance of the machine learning models is majorly evaluated by the number of samples that are predicted into Parkinson's disease positive class or negative class. Confusion Matrix gives these numbers and summarize as TruePositive(TP),TrueNegative(TN),FalseNegative (FN), Falsepositive(FP).

True Positive (TP) model accurately identifies Parkinson's disease in the sample.True Negative (TN) model accurately identifies the sample as Healthy Control.False positive (FP) model identifies the sample as Parkison's Disease. False Negative (FN) model identifies the sample as Healthy Control.Evaluation Metrics are defined as below.

- Accuracy: Samples which are accurately identified /Total samples

  Formula: TP + TN/(TP+FP+TN+FN)

- Sensitivity (Probability of disease detection): How probable is it that the model will be correct if the patient has a Parkinson disease.

  Formula: TP/TP+FN

- Precision: What is the probability that a patient would develop a Parkinson's disease, if a model prediction is positive.

  Formula: TP/TP+FP

- F1 score: Represented as harmonic mean of Model Precision and Sensitivity.

  Formula : 2* Sensitivity * Precision/ (Sensitivity + Precision)

- Specificity: Probability of the model prediction as healthy control if the sample is healthy.

  Formula: TN/TN+ FP

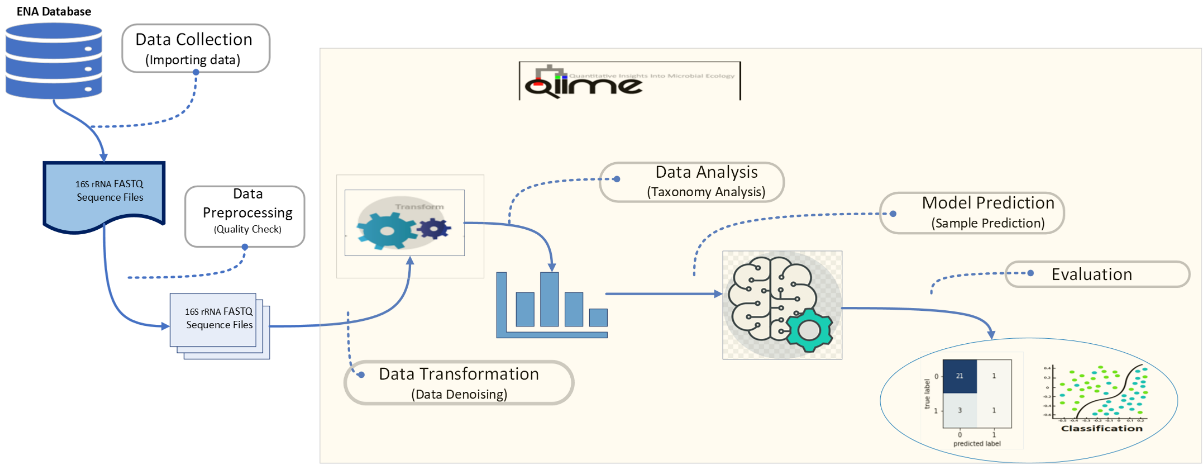

Figure 1: Research Methodology for 16S rRNA Sequence Sample Prediction

4 Design Specification & Implementation

In this research ,16S rRNA gene raw sequencing reads of fecal samples downloaded from the ENA database and the microbiome analysis of these raw pair end sequence files was performed using QIIME 2 (version: 2023.5.1) Core Concepts (2023). High level data analysis and visualization can be accomplished using this versatile and potent tool. This adaptable tool includes a standard methodology for analysing the raw sequence data.Figure 2 Core Concepts (2023) provides a detailed illustration of the fundamental overview of how QIIME 2 analyses the raw sequence and provides the amplicon sequence analysis and visualisations. The first stage of the workflow, however, may differ depending on the type and layout of the input sequences. Although there are many further steps for extra analysis in the QIIME 2 pipeline, the boxes highlighted in red in Figure 2 illustrates all the processes that were carried out for this study ,as the objective of this study is to analyse taxonomy classification and predict the samples.

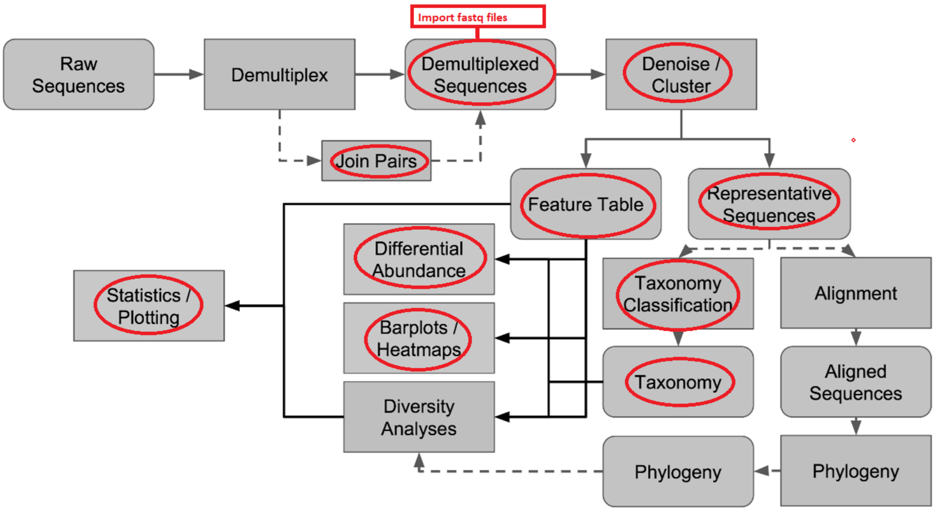

Figure 2: QIIME2 Design & Conceptional Overview

In QIIME 2 platform while we use any method or plugins it follows a simple structure of instructions which is illustrated by a flow chart in Figure 3 Core Concepts (2023). The output files (artifacts) produced at each stage are used as inputs during the subsequent phase of analysis. Artifacts are fed into the QIIME 2 pipeline, or the methods and a visualizer is used to display the artifacts.

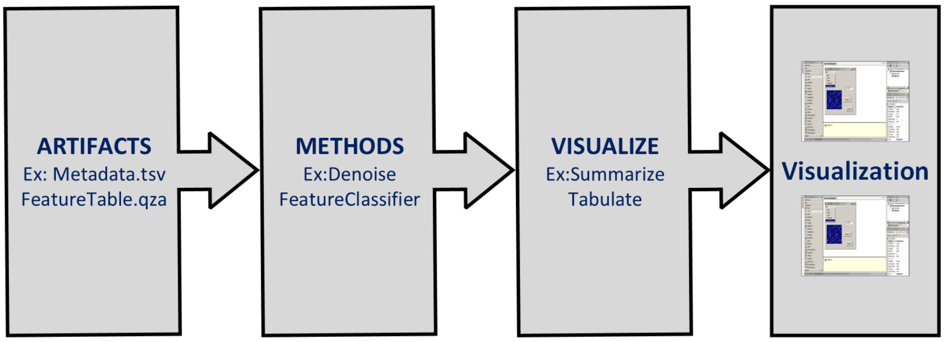

Figure 3: QIIME2 Process Flow Chart

For this study the selected raw sequence files are already de-multiplexed ,which means sequenced reads are divided into distinct files for each sample in a sequenced run. Below are the techniques listed to analyse these files and extract the features , ASV's or Taxonomy information.

4.1 Importing

The selected 17 samples are de-multiplexed paired end sequencing with biological sequences and quality scores FASTQ files. Therefore, importing the files in QIIME2 environment begins at De-multiplexed sequences step shown in Figure 2. The manifest file format, which is a flexible import technique, is used to import samples as paired-end sequences with quality scores with input format set to 'PairedEndFastqManifestPhred33V2' ImportFiles (2023).The samples are mapped with their absolute path on the local disk in this tab-separated manifest file.

4.2 Denoising

At this stage, the sequences are subjected to quality control. The imported de-multiplexed paired end Fastq files are used as input artifacts at this point. At this phase either the DADA2 or Deblur techniques can be used to conduct denoising in QIIME 2. Deblur can only manage single-end reads Callahan et al. (2016), in contrast to DADA2, which can perform paired-end reads natively. The sequences are denoised in unique ways by both techniques.

In this study, the sequences are denoised using the Deblur technique, a greedy de - convolution algorithm that uses the Illumina Miseq/Hiseq error profiles for amplicon sequencing. This is an alternative if memory or computing resources are limited or if need to work with smaller bp chunks (such as 150 bp or fewer), but it also appears to be effective for longer readings Nearing et al. (2018). As the input samples are paired end sequences, with the use of merge pair method in VSEARCH plugin Rognes et al. (2016), forward and reverse reads are joined to give single end reads. Deblur denoise-16S method is used, based on a decline in quality ratings, this p-trim length of 300 is chosen as it regulates the sequences length Amir et al. (2017). When the data has been sufficiently denoised, the findings can be verified by glancing at the summary of the feature table. Feature tables and representative sequences are the output artifacts at this stage, which are crucial for additional investigation. The frequency of each ”feature” (such as ASVs, OTUs, etc.) found in samples is represented by a feature table, which is effectively a matrix of samples vs. observations. Following the creation of this feature table, taxonomic analysis of the samples is carried out.Steps for denoising the samples in QIIME2 with deblur 16s method for this study are shown in Figure 4.

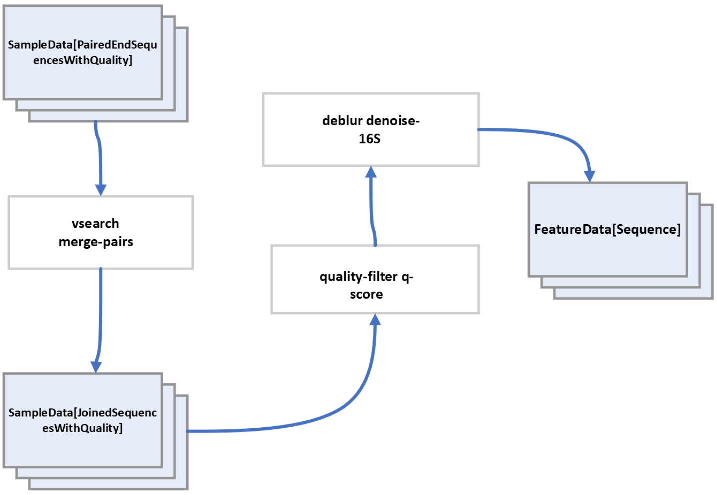

Figure 4: QIIME2 Deblur-DenoisePairendSequences

4.3 Taxonomy Classification

The goal of this step is to find which microbial species present in the samples, this is achieved using the q2-feature-classifier plugin available in QIIME2 . This classifier's primary input artifact is a feature table. Taxonomy classifiers identify the closest taxonomic association at this stage with some degree of confidence or consensus based on alignment, k-mer frequencies, and other variables. By evaluating the query sequences (i.e., features, whether ASVs or OTUs) with a database of reference of sequences with established taxonomic compositions, such as the Silva Robeson et al. (2021) or Green Genes data bases Bokulich, Kaehler, Rideout, Dillon, Bolyen, Knight, Huttley and Caporaso (2018). Here, python classify scikit learn is used to access a pre-trained Naive Bayes machine learning classifier Ziemski et al. (2021), which is suggested for usage with the 16srRNA gene analysis.Figure 5 illustrated the taxonomy pipeline.

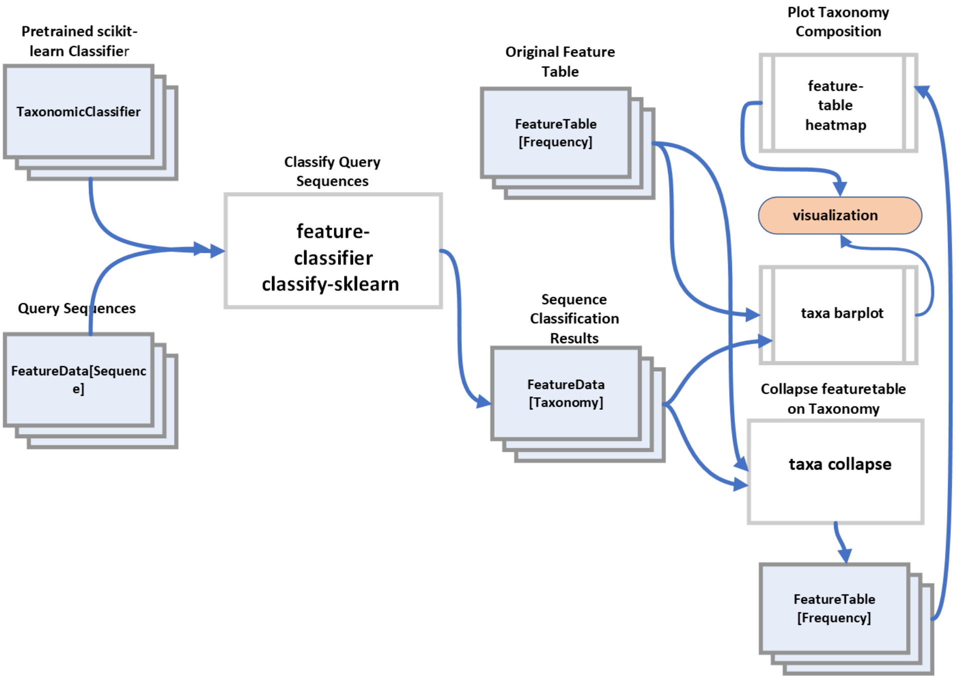

Figure 5: QIIME2 Pre-trained Taxonomy Classification Pipeline

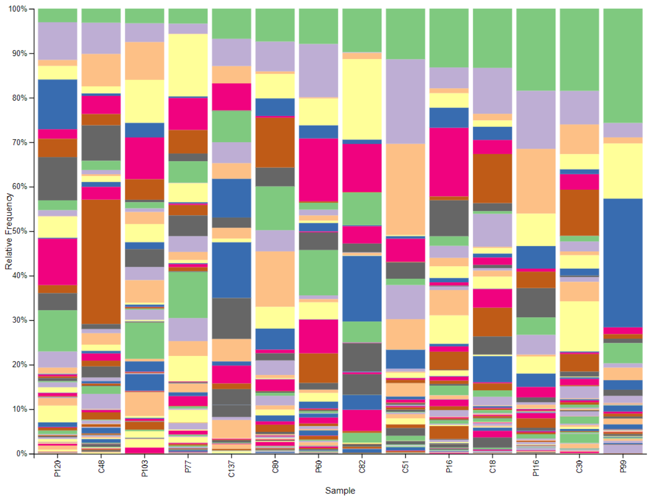

Figure 6: Taxonomy Classification per Sample with Silva Reference

4.4 Sample Prediction

The QIIME 2 sample classifier makes use of supervised learning algorithms. By studying the composition of labelled training samples given in sample metadata file , supervised learning classifiers make predictions about the categorical metadata classes of unlabelled samples. Depending on the composition of the stool microbiome, sample classifier is used to diagnose or predict Parkinson's disease susceptibility Bokulich, Dillon, Bolyen, Kaehler, Huttley and Caporaso (2018). This pipeline has set of actions to achieve the sample prediction which is listed below and Figure 7 describes the process flow.

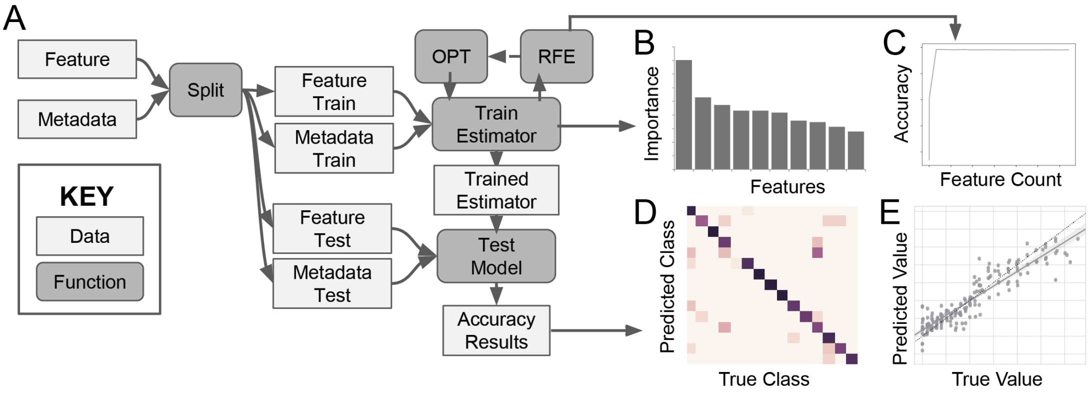

Figure 7: QIIME2 Sample Classifier Process Pipeline

- Training and Test sets are created by dividing samples at random. One the model training on train set is completed the test sample set is used for accessing the accuracy of the model at the end of the process. The –p-test-size argument is used to adjust the percentage of input samples to include in the test set.
- Based on the feature information corresponding to each sample (included in a metadata field), the model is trained to predict a certain target value. –p-estimator

argument allows to choose the algorithm and predict samples based on various scikit learn library Pedregosa et al. (2011). RandomForestClassifier, LinearSVC, SVC,GradientBoostingClassifier, AdaBoostClassifier,KNeighborsClassifier are some of the estimators currently available to choose.

- To fine-tune the model, automatic selection of features and parameter optimization processes are carried out. By default, five-fold cross-validation is used; the –p-cv argument can be used to change this value.

- Based on the feature data related to each test sample, the trained model is used to forecast the target values and likelihoods of classes for each sample.

- By contrasting the predicted value for each test sample with the actual value for that sample, the model's accuracy is determined.

5 Evaluation

5.1 Sample Classifier- Default Parameters

In this experiment, the QIIME2 default q2-sample classifier inputs are used ,samples are divided into two groups: 80% for training the model and 20% for testing and Jobs to run in parallel is set to 1.For sample prediction, the sample meta data column ”DiseaseStatus” is used as an argument. By default, the q2-sample classifier employs the Random Forest Classifier with 100 trees, 5-fold cross-validation (CV), and parameter tuning is carried out automatically by using cross validated randomized parameter grid search with Scikit Learn Randomized Search. Cross-validated recursive feature elimination from Scikit learn Recursive Feature Elimination (RFE) was used to select the features with the best predicted accuracy.

The accuracy for the classification model (classifying samples to PD and HC) is shown in confusion matrix , ROC and AUC is shown in Figure 8 and Figure 9 respectively.

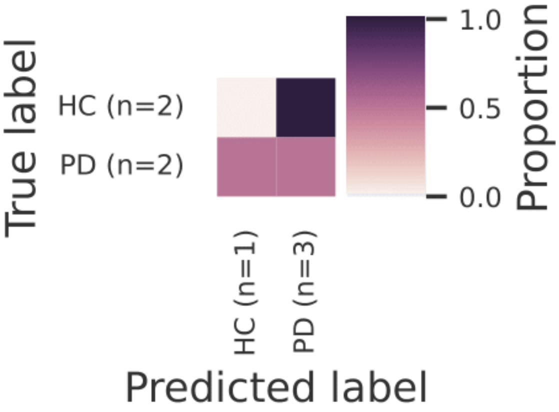

Figure 8: Confusion Matrix Heat Map-Default Inputs

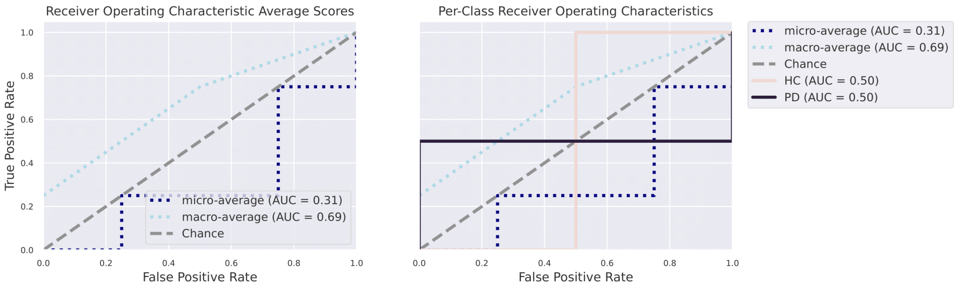

Figure 9: ROC- AUC Default Inputs

Accuracy of 25% , Precision 33%, Sensitivity 50% and F1 Score 0.397 achieved with experiment with default inputs.

5.2 Sample Classifier- RF Classifier

In this experiment, samples are divided into two groups: 70% for training the model and 30% for testing. For sample prediction, the sample meta column "DiseaseStatus" is used as an argument. q2-sample classifier employs the Random Forest Classifier with 50 decision trees, 5-fold cross-validation (CV), and hyper parameter tuning is carried out automatically using a cross-validated randomized parameter grid search with Scikit Learn Randomized Search. Cross-validated recursive feature elimination from Scikit learn Recursive Feature Elimination (RFE) was used to select the features with the best predicted accuracy.

The accuracy for the classification model (classifying samples to PD and HC) is shown in confusion matrix , ROC and AUC is shown in Figure 10 and Figure 11 respectively.

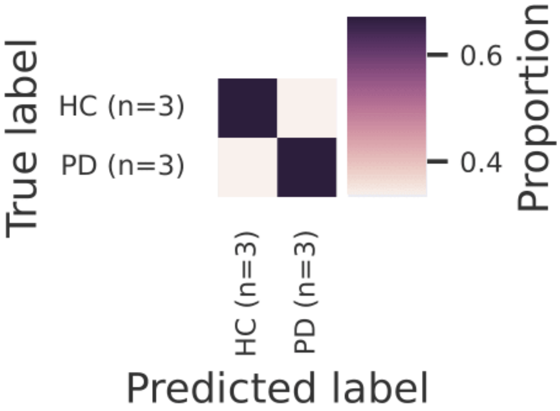

Figure 10: Confusion Matrix Heat Map- Random Forest

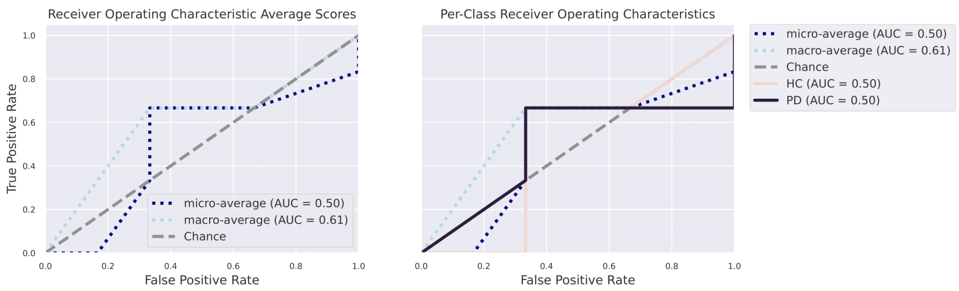

Figure 11: ROC- AUC Random Forest

Accuracy of 66% , Precision 66%, Sensitivity 66% and F1 Score 0.88 is achieved with this settings.Figure 12 shows the heat map with distinct features found within input samples with this experiment.Table 1 refers to the predictions and probabilities of the sample with this RF model.

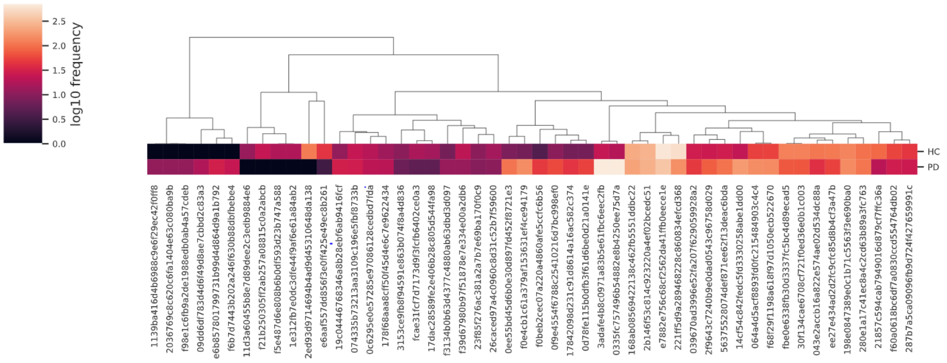

Figure 12: Heat Map Showing Sample Features

Table 1: q2 Sample Classifier RF estimator Probabilities & Predictions

Sample Id	HC	PD	prediction
C82	0.56	0.44	HC
P53	0.58	0.42	HC
P120	0.46	0.54	PD
C48	0.44	0.56	PD
C80	0.54	0.46	HC
P103	0.44	0.56	PD

5.3 Sample Classifier- LSVC Classifier

In this experiment, samples are divided into two groups: 70% for training the model and 30% for testing and random state of 123 is used. For sample prediction, the sample meta column ”DiseaseStatus” is used as an argument. The q2-sample classifier employs the Linear Support vector Classifier, 5-fold cross-validation (CV), and hyper parameter tuning is carried out automatically using a cross-validated randomized parameter grid search with Scikit Learn Randomized Search.Cross-validated recursive feature elimination from Scikit learn Recursive Feature Elimination (RFE) was used to select the features with the best predicted accuracy.The accuracy for this model is shown in confusion matrix , ROC and AUC is shown in Figure 13 and Figure 14 respectively.

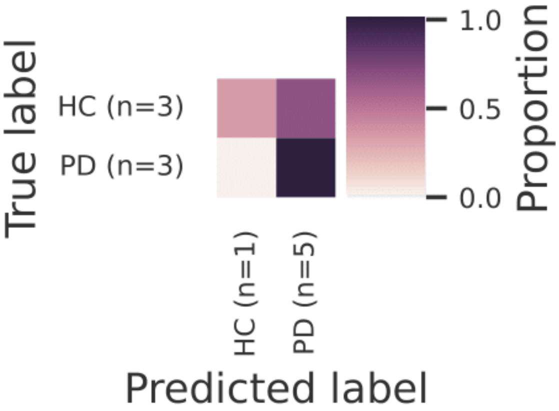

Figure 13: Confusion Matrix Heat Map- LSVC

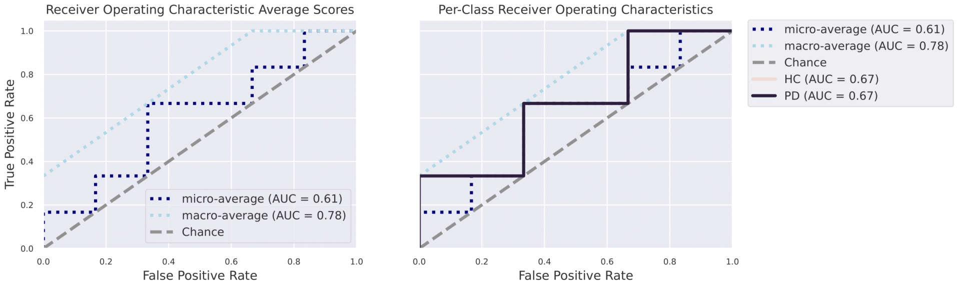

Figure 14: ROC- AUC Linear Support Vector Classifier

Finally,the results are evaluated in the Table2 with Accuracy, Recall, Precision,Specificity and F1 Score. From the findings RF Classifier with decision tress 50 were the performed best among other input settings, with an F1 score of 0.88 for a predicted class and an accuracy of 66.6%. LSVC model has high recall but low specificity means model is not performing good at predicting true negative samples.

Table 2: q2 Sample Classifier Model Results

Estimator	Accuracy	Precision	Recall	Specificity	F1 Score
RF-Default Inputs	0.25	0.33	0.5	0	0.397
RF	0.66	0.66	0.66	0.66	0.88
LSVC	0.66	0.66	1	0.33	0.82

6 Conclusions and Discussion

The Machine Learning data analysis of gut microbiome in PD patients and HC is presented in this study. The 16S rRNA sequence is gathered from the ENA database that were accessible and derived from studies that found a connection between microbiota and disease Parkinson's disease progression and specific bacterial taxa that differed between patients and controls. Using the most recent bioinformatic techniques, this study uniformly processed the downloaded data and re-analysed the data sets.The effectiveness of two machine learning (ML) algorithms (RF, Linear SVC) using QIIME 2 sample classifier in recognizing samples from HC or PD patients was assessed using a variety of measures (AUC, accuracy, precision, recall, and F-score). The relatively low accuracy could be attributed to the numerous reasons one is definitely very low samples used for training the model, to improve the model's accuracy and produce more reliable findings about the relationship between the gut microbiota and PD, a larger number of samples must be used to train the RF algorithm. The other reason could be bioinformatic techniques utilized for data processing and analysis.

This study was conducted with some limitations,one of them is very small sample size of only 17 samples. The major contributing reason to this restriction was computational resources, which had a direct impact on the data collection and its subsequent analysis. As a result, both the volume and diversity of the data set used to train the machine learning model was restricted.

In the future study the aim is to increase sample sizes in order to overcome this restriction. By increasing the sample size, machine learning models could be trained in a more reliable and consistent manner, improving their capacity to predict accurately. Additionally, the performance and prediction ability of the model would be improved by using sophisticated algorithms, feature selection approaches, and maybe even multi-omics data. The objective is to study about gut microbiome by using larger, diverse data sets together with advance machine learning techniques to better understand the complex relationships between the gut microbiota and various neurological disorders.

References

Aho, V. T., Pereira, P. A., Voutilainen, S., Paulin, L., Pekkonen, E., Auvinen, P. and Scheperjans, F. (2019). Gut microbiota in parkinson's disease: temporal stability and relations to disease progression, *EBioMedicine* **44**: 691–707.

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A. et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns, *MSystems* **2**(2): 10–1128.

Bang, S., Yoo, D., Kim, S.-J., Jhang, S., Cho, S. and Kim, H. (2019). Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data, *Scientific reports* **9**(1): 10189.

Boktor, J. C., Sharon, G., Verhagen Metman, L. A., Hall, D. A., Engen, P. A., Zreloff, Z., Hakim, D. J., Bostick, J. W., Ousey, J., Lange, D. et al. (2023). Integrated multi-cohort analysis of the parkinson's disease gut metagenome, *Movement Disorders* **38**(3): 399–409.

Bokulich, N. A., Dillon, M. R., Bolyen, E., Kaehler, B. D., Huttley, G. A. and Caporaso, J. G. (2018). q2-sample-classifier: machine-learning tools for microbiome classification and regression, *Journal of open research software* **3**(30).

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A. and Caporaso, J. G. (2018). Optimizing taxonomic classification of marker gene amplicon sequences, *PeerJ Preprints* **6**: e3208v2.

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., Mills, D. A. and Caporaso, J. G. (2013). Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing, *Nature methods* **10**(1): 57–59.

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. et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using qiime 2, *Nature biotechnology* **37**(8): 852–857.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. (2016). Dada2: High-resolution sample inference from illumina amplicon data, *Nature methods* **13**(7): 581–583.

Chen, Y., Wu, T., Lu, W., Yuan, W., Pan, M., Lee, Y.-K., Zhao, J., Zhang, H., Chen, W., Zhu, J. et al. (2021). Predicting the role of the human gut microbiome in constipation using machine-learning methods: a meta-analysis, *Microorganisms* **9**(10): 2149.

classifier, s. (2023). Sample-classifier.
URL: *https://docs.qiime2.org/2023.5/plugins/available/sample-classifier/*

Coelho, B. F. O., Massaranduba, A. B. R., dos Santos Souza, C. A., Viana, G. G., Brys, I. and Ramos, R. P. (2023). Parkinson's disease effective biomarkers based on hjorth features improved by machine learning, *Expert Systems with Applications* **212**: 118772.

Core Concepts, Q. . (2023). Qiime 2 core concepts.
URL: *https://docs.qiime2.org/2023.5/concepts/*

Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., Gaglio, S. and Urso, A. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data, *BMC bioinformatics* **19**: 61–76.

Hasic Telalovic, J., Cicak Bašić, D. and Osmanovic, A. (2022). Investigation of the role of the microbiome in the development of alzheimer's disease using machine learning techniques, *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies*, Springer, pp. 639–649.

Hey, G., Nair, N., Klann, E., Gurrala, A., Safarpour, D., Mai, V., Ramirez-Zamora, A. and Vedam-Mai, V. (2023). Therapies for parkinson's disease and the gut microbiome: evidence for bidirectional connection, *Frontiers in Aging Neuroscience* **15**: 1151850.

Hill-Burns, E. M., Debelius, J. W., Morton, J. T., Wissemann, W. T., Lewis, M. R., Wallen, Z. D., Peddada, S. D., Factor, S. A., Molho, E., Zabetian, C. P. et al. (2017). Parkinson's disease and parkinson's disease medications have distinct signatures of the gut microbiome, *Movement disorders* **32**(5): 739–749.

Huang, B., Chau, S. W., Liu, Y., Chan, J. W., Wang, J., Ma, S. L., Zhang, J., Chan, P. K., Yeoh, Y. K., Chen, Z. et al. (2023). Gut microbiome dysbiosis across early parkinson's disease, rem sleep behavior disorder and their first-degree relatives, *Nature Communications* **14**(1): 2501.

ImportFiles, Q. (2023). Importing data.
URL: *https://docs.qiime2.org/2023.5/tutorials/importing/*

Liang, H., Jo, J.-H., Zhang, Z., MacGibeny, M. A., Han, J., Proctor, D. M., Taylor, M. E., Che, Y., Juneau, P., Apolo, A. B. et al. (2022). Predicting cancer immunotherapy response from gut microbiomes using machine learning models, *Oncotarget* **13**: 876.

LinearSVC, L. (2023). Sklearn.svm.linearsvc.
URL: *https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC*

Nearing, J. T., Douglas, G. M., Comeau, A. M. and Langille, M. G. (2018). Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches, *PeerJ* **6**: e5364.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in python, *the Journal of machine Learning research* **12**: 2825–2830.

Pietrucci, D., Teofani, A., Unida, V., Cerroni, R., Biocca, S., Stefani, A. and Desideri, A. (2020). Can gut microbiota be a good predictor for parkinson's disease? a machine learning approach, *Brain Sciences* **10**(4): 242.

RandomForest, R. (2023). Sklearn.ensemble.randomforestclassifier.
URL: *https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier*

Resources, D. (2023). Qiime 2 data resources.
URL: *https://docs.qiime2.org/2023.5/data-resources/*

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T. and Bokulich, N. A. (2021). Rescript: Reproducible sequence taxonomy reference database management, *PLoS computational biology* **17**(11): e1009581.

Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. (2016). Vsearch: a versatile open source tool for metagenomics, *PeerJ* **4**: e2584.

Roos, D. S., Klein, M., Deeg, D. J., Doty, R. L. and Berendse, H. W. (2022). Prevalence of prodromal symptoms of parkinson's disease in the late middle-aged population, *Journal of Parkinson's Disease* **12**(3): 967–974.

Sampson, T. R., Debelius, J. W., Thron, T., Janssen, S., Shastri, G. G., Ilhan, Z. E., Challis, C., Schretter, C. E., Rocha, S., Gradinaru, V. et al. (2016). Gut microbiota regulate motor deficits and neuroinflammation in a model of parkinson's disease, *Cell* **167**(6): 1469–1480.

Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L. and Ngom, A. (2019). A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer, *Frontiers in genetics* **10**: 256.

Vilne, B., Ķibilds, J., Siksna, I., Lazda, I., Valciņa, O. and Krūmiņa, A. (2022). Could artificial intelligence/machine learning and inclusion of diet-gut microbiome interactions improve disease risk prediction? case study: coronary artery disease, *Frontiers in Microbiology* **13**: 627892.

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

Zhang, J. (2022). Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of parkinson's disease, *NPJ Parkinson's disease* **8**(1): 13.

Zhu, Q., Hou, Q., Huang, S., Ou, Q., Huo, D., Vázquez-Baeza, Y., Cen, C., Cantu, V., Estaki, M., Chang, H. et al. (2021). Compositional and genetic alterations in graves' disease gut microbiome reveal specific diagnostic biomarkers, *The ISME journal* **15**(11): 3399–3411.

Ziemski, M., Wisanwanichthan, T., Bokulich, N. A. and Kaehler, B. D. (2021). Beating naive bayes at taxonomic classification of 16s rrna gene sequences, *Frontiers in Microbiology* **12**: 644487.