# Cancer Prediction Based on Microbiome Profile

Shivam Sharma

Supervised by Dr Arief Gusnanto and Dr Henry M Wood

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Master of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

August 2024

# Abstract

This dissertation investigates the use of microbiome profiling combined with machine learning to enhance colorectal cancer (CRC) detection. By analyzing microbial data with models such as Random Forest, LightGBM, Logistic regression with L2 regularisation, and ensemble methods, the study identifies distinct microbial signatures associated with CRC, adenoma, and blood-negative cases. The ensemble model outperformed others, highlighting the potential of integrating machine learning in non-invasive CRC diagnostics. Key microbial families like Fusobacteriaceae were significant in model predictions, aligning with existing research. The findings suggest that microbiome-based approaches could revolutionize CRC screening, offering a promising path toward more accurate and personalised healthcare solutions.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Colorectal cancer (CRC) is the third most common cancer type in the world and, most tragically, the second leading cause of cancer mortality in women. In the past decade, the incidence of CRC has increased significantly in developing countries. Currently, CRC accounts for approximately 10% of all cancers in men and 9.2% in women. The reason for the high prevalence of CRC is due to the challenges in early diagnosis as patients often display nonspecific symptoms that may be associated with less serious conditions. Symptoms are mild and often delayed, including rectal bleeding, changes to bowel habits and abdominal discomfort. Unfortunately, due to the non-specific nature of the symptoms, they are often missed and signs are not identified. Ultimately, these factors lead to delayed diagnosis and poor prognosis.

Currently, CRC is most commonly detected with a colonoscopy or a faecal occult blood test (FOBT). A colonoscopy remains the best diagnostic test as it is highly sensitive and specific (more than 90 per cent), allowing for direct visualisation of the colon and the removal of polyps at the same time. Despite its benefits, colonoscopy is much more complicated and invasive, requiring the creation of a clear colon with laxatives, and may cause cramping and discomfort, which can lead to lower patient compliance with the procedure (Brenner et al., 2014). In contrast, FOBT is non-invasive, however, its sensitivity and specificity are significantly lower compared with a colonoscopy. FOBT can produce false-negative and false- positive results, suggesting further screening tests for true positive indications i.e., microscopic blood in the faeces (Levin et al., 2008). Therefore, additional studies such as the utilisation of a colonoscopy may be required. The combination of both a colonoscopy and FOBT testing is known as double negative screening and has higher accuracy than either method alone. Despite these findings, both tests are associated with patient discomfort and do not lead to higher compliance with the screening protocol. This makes it imperative to find a simpler, more effective, and more patient-friendly approach to early detection.

## 1.1 Importance of Early Detection

Colorectal cancer has a much better prognosis if diagnosed early. Early detection leads to fewer intrusive treatments and a higher survival percentage. Siegel et al. (2020) discovered that patients with localised CRC have a higher five-year survival rate (about 90%) than those with distant metastatic disease (14%). Early identification improves patient outcomes which reduces healthcare costs and boosts quality of life. Improving screening technologies for early detection is a critical public health priority.

## 1.2 Role of the Microbiome in Human Health

The human gut microbiome contains trillions of microorganisms, including bacteria, viruses, fungus, and archaea, that live in the digestive tract. Microorganisms are essential for maintaining health as they promote digestion, synthesise vitamins, regulate the immune system, and defend against harmful bacteria (Lynch & Pedersen, 2016). Diet, lifestyle, and genetics all influence the gut microbiome's makeup and functionality.

Recent literature has focused on the microbiome's role in the development of a variety of diseases which includes colorectal cancer. Dysbiosis, or microbial imbalance in the gut, has been linked to inflammation and the development of CRC. Certain bacteria like *Fusobacterium nucleatum* have been associated with tumorigenesis by creating a pro-inflammatory environment conducive to cancer development (Brennan & Garrett, 2019). As a result, analyzing the gut microbiome offers a promising avenue for non-invasive cancer detection, leveraging microbial signatures as potential biomarkers for early diagnosis.

## 1.3 Research Problem

### 1.3.1 Drawbacks and Challenges of Current Colonoscopy Methods

While colonoscopy remains the definitive method for CRC screening, its invasiveness, high cost, and associated risks pose significant challenges. The procedure requires patients to undergo rigorous bowel preparation, which many find unpleasant, and it carries risks of complications such as bleeding and perforation, particularly in older adults and those with comorbidities (Brenner et al., 2014). Additionally, the need for sedation during the procedure adds to the complexity and cost, often necessitating recovery time and affecting patient compliance.

The limited availability of endoscopic facilities and trained personnel further exacerbates the accessibility of colonoscopy, especially in resource-limited settings. These barriers underscore the need for alternative screening methods that are less invasive, more accessible, and equally effective.

### 1.3.2 Potential of Microbiome Profiles in Cancer Detection

The gut microbiome presents a promising alternative for CRC screening. The composition of the microbiome reflects the state of health and disease within the host, with specific microbial patterns associated with CRC (Zeller et al., 2014). Microbial dysbiosis, characterized by the presence of pathogenic bacteria and a decrease in beneficial microbial diversity, has been linked to colorectal carcinogenesis. For instance, Fusobacterium nucleatum has been found in higher abundance in CRC patients, suggesting its potential role as a biomarker for early detection (Brennan & Garrett, 2019).

The utilisation of microbiome profiles for cancer detection involves analyzing faecal samples, which are non-invasive and can be collected easily, enhancing patient compliance. Advances in high-throughput sequencing technologies have facilitated detailed characterisation of the microbiome, enabling the identification of microbial signatures associated with CRC. This approach could complement existing screening methods, offering a less invasive, cost-effective, and scalable solution for early cancer detection.

## 1.4 Significance of the Study

### 1.4.1 Contribution to Early Cancer Detection Methods

The thesis contributes to the field of cancer detection by exploring the use of microbiome profiles as non-invasive biomarkers for colorectal cancer. By leveraging the power of machine learning techniques, the research aims to develop predictive models that can accurately classify individuals based on their microbiome composition. This approach has the potential to enhance early detection which would lead to improved patient outcomes, and reduced burden on healthcare systems. The use of microbiome analysis could complement existing screening methods, providing a less invasive and more patient-friendly alternative.

### 1.4.2 Potential Impact on Clinical Practices and Patient Outcomes

Integrating microbiome analysis and machine learning models into routine screening programs could revolutionize colorectal cancer detection. It offers a non-invasive, cost-effective, and scalable alternative to traditional methods, potentially increasing participation rates in screening programs. Early detection through microbiome-based screening can lead to timely intervention, reducing mortality rates and improving the quality of life for patients. Furthermore, understanding the microbiome's role in cancer development could open new avenues for preventive and therapeutic strategies, advancing personalized medicine approaches in oncology.

## 1.5 Overview of the Project

1.6 Overview of the Dissertation This dissertation presents a complete investigation of cancer prediction using microbiome profiles through the extensive methodology and advanced machine learning approaches. The research begins with an introduction that outlines the research problem, objectives, and the significance of the work. It also establishes the context by describing colorectal cancer (CRC) prevalence and detection issues, the significance of early detection, and the role of the gut microbiome in health and disease.

The literature review delves into existing research on CRC detection meth- ods, and explores the role of microbiome alterations and health statuses. This section also reviews the application of machine learning techniques in microbiome research, providing a context for the current study and identifying gaps in the existing literature that this research aims to address.

The methodology section outlines the research design and methods used in the study, including the data collection and preprocessing methods. Detailed descriptions of preprocessing techniques, such as data integration and handling missing values, are also provided. It also covers exploratory data analysis (EDA), detailing numerical summaries and graphical visualizations that help understand the data's structure. It further explains the machine learning models employed, including Random Forest, Logistic Regression, LightGBM, and ensemble model, and discusses the specific techniques used for hyperparameter tuning and model evaluation.

The results section presents the findings from the EDA, diversity analysis, and the performance metrics of the ML models. Each model's results are discussed in detail, highlighting their effectiveness in predicting CRC based on microbiome profiles and examining the impact of hyperparameter tuning on model performance, including validation results obtained through cross-validation techniques.

The discussion interprets these results in light of existing literature, discussing their implications for early cancer detection and the potential impact on clinical practices and patient outcomes. This section also addresses suggestions for future research along with the study's contributions to the field, emphasizing how the integration of microbiome analysis can enhance current CRC screening methods.

Finally, the conclusion summarizes the key findings, and discusses the study's limitations. This section underscores the potential of microbiome-based screening as a non-invasive, cost-effective alternative for early CRC detection and highlights areas for further investigation to improve and validate these methods. This structured approach provides a comprehensive framework for advancing cancer prediction using microbiome profiles.

# Chapter 2

# Literature Review

## 2.1 Cancer Detection Methods

A lot of research needs to go into screening for colorectal cancer (CRC) to save more lives and make the disease less dangerous overall. The likelihood of a successful therapy and a long-life expectancy are both enhanced by early detection of several cancer types. Shaukat & Levin (2022) examined the current status of CRC detection methods, including their operation and potential hazards, in their study.

### 2.1.1 Pathogenesis and Risk Factors

Colorectal cancer can be caused by several things, including hormones, the environment, and genes. The usual way that colorectal cancer grows is through a tumour and then a carcinoma. Nevola et al. (2023) showed how the cancer starts in the epithelial of the intestines, then moves to adenomatous polyps, eventually to invasive carcinoma. Over time, there are more changes to key oncogenes (like KRAS) and cancer suppressor genes (like APC and TP53). Esophageal changes, such as DNA methylation and histone modification, are what make these changes happen.

Colorectal cancer is in its initial phases when the APC gene stops working. The Wnt/$\beta$-catenin signalling system gets messed up, and cells start to grow out of control. Genes like KRAS, SMAD4, and TP53 change more as the cancer gets worse. This helps the cancer grow faster and makes aggressive carcinoma (Ponziani et al., 2019) to emerge.

Colorectal cancer is more likely to happen in people who have certain types of bacteria in their gut according to new studies. Microbes in the gut, which number in the billions (Nevola et al., 2023), play a crucial role in keeping the gut healthy and regulating the immune system. When the gut microbiota is out of balance, which is called dysbiosis, it can lead to colorectal cancer in many ways.

First, chronic inflammation is a significant factor. *Escherichia coli* and *Fusobacterium nucleatum* are two types of bacteria that may cause long-term inflammation in the gastrointestinal

system. In this process, inflammatory pathways like NF-$\kappa$B are activated and pro-inflammatory cytokines like IL-6 and TNF-$\alpha$ are produced. Elinav et al. (2019) explained that a continuous inflammation reaction creates an atmosphere that is ideal for cancer growth and development.

Second, some bacteria can exhibit Genotoxicity, like *Bacteroides fragilis*, produce toxins that can damage the DNA of human epithelial cells and lead to changes and genetic instability. This is shown by the fact that *B. fragilis* toxin (BFT) activates the Wnt/$\beta$-catenin signalling pathway, which damages DNA and foster cancer growth.

Moreover, Pathogens can change the immune system of the host, which could promote or hinder the development of cancers. Some microbiome may not be discovered by the immune system, which can cause unchecked tumor growth. It has been found, though, that certain microbes in the gut could assist the body's immunity fight cancer by increasing the influx of immune cells into tumours (Poore et al., 2020).

Finally, the production of metabolites by bacteria in the gut can make chemicals that can lead to cancer development. For example, secondary bile acids have been shown in lab tests to promote growth and survival of intestinal epithelial cells. Microbes also make other chemicals, such as short-chain fatty acids (SCFAs), that help intestine epithelial cells develop and keep the gut barrier strong (Poore et al., 2020).

### 2.1.2 Current Detection Methods

If colorectal cancer is found early, it can be treated better and the death rate can be lowered. These days, most cancer screenings look for adenomatous polyps alongside additional growths that could turn into cancer. This way, they can be removed before the cancer gets worse. The two primary methods used for identifying colorectal cancer are a colonoscopy and FOBT.

#### Colonoscopy

A colonoscopy, which makes a clear picture of the colon, is one of the best ways to check for colorectal cancer while pregnant. Part of the process involves putting a bendable tube with a camera into the rectum and then gradually moving it into the stomach. By using these techniques doctors may clearly be able to check the digestive track walls that provides an option of complete examination of the system observed.

#### Faecal Occult Blood Testing (FOBT)

The technique is immensely helpful in assessing colorectal cancer effectively in patients, with the utilisation of microbiome research it can predict the chance of getting cancer by individual screening. The technique is quite helpful in identifying any hidden faecal sample of blood. This can be a fast and efficient way to check early stages of cancer in patients.

One form of FOBT, known as the Guaiac-based FOBT (gFOBT). This kind of techniques helps to examine any occurrence of blood present in the faeces by using a chemical reaction

which consists guaiac reagent, which changes color in the presence of blood. Using this kind of test, polyps and the early stages of cancer may not be found. Also, Schreuders et al. (2022) said that the data can be wrong because of what people eat and the drugs they take.

Another type of FOBT is the Faecal Immunochemical Test (FIT), in which, antibodies are created specifically to find out if there is blood in human faeces. The antibodies are used to detect hemoglobin, the protein in red blood cells. Another advantage of the FIT program is that it doesn't require the patients to change the way they eat. The FIT test is very effective at finding the first signs of colorectal cancer. An important reason for this is that the test is very good at finding blood that comes from the lower digestive system.

#### Emerging Detection Methods

With the resources that are available currently, colorectal cancer (CRC) is not likely to be discovered early by employing them. Because of this, researchers and scientists constantly search for exciting and simple signs that can help. Genetic tests using urine and blood as markers is one of the most hopeful ways. Biomarkers in the blood check for ctDNA and other cancer-related chemicals in the circulation. DNA tests in faeces samples look for genetic problems and epigenetic changes linked to colorectal cancer.

Cologuard® is a multi-target DNA test for faeces that may find blood and DNA signs of colorectal cancer that aren't obvious at first glance. Some of these signs are changes in the methylation of the NDRG4 and BMP3 genes and changes in the KRAS gene. These tests are better than FOBT at finding CRC and advanced adenomas, but they are more expensive and need more complicated laboratory operations (Ahlquist, 2019).

A fairly new method called "liquid biopsy" looks at ctDNA in blood samples. This DNA gets into the bloodstream when tumours grow. Colorectal cancer is linked to some genetic changes, including differences in copy number and methylation patterns that can be found with ctDNA research. Before liquid biopsy can be used to find CRC faster than other screening methods, its clinical usefulness is still being studied (Kartal et al., 2022).

As study into the gut microbiome goes on, it becomes evident that bacteria fingerprints can be used to find colorectal cancer. More and more evidence suggest that changes in the gut microbiome, which is a living, changing ecosystem, may play a part in the growth of colon cancer. Metagenomics as well as 16S rRNA gene sequencing are two advanced sequencing methods that can be used to get a more complete picture of the bacteria communities found in faeces samples (Poore et al., 2020).

## 2.2 Microbiome and Cancer

The intricate interplay between colorectal cancer (CRC) and gut flora alters the progression and development of the illness. The most significant objective is to learn more about this subject. This part of the research will talk about the bacteria that live in the gut in more depth, and

how this subject is connected to dysbiosis and the growth of colon cancer. Poore et al. (2020) explained that biomarkers made from bacterial samples could help find colorectal cancer early.

### 2.2.1 Overview of the Gut Microbiome

Various bacterial species thrive inside the gastrointestinal tract. Archaea, viruses, bacteria, and fungus belong to the same category. These bacteria do more than just live in people's gut systems; they make things worse for people. In addition, they play a part in or affect many important bodily processes (Poore et al., 2020).

The good bacteria in the gut system help break down food and absorb nutrients, among other things. Short-chain fatty acids (SCFAs) like butyrate, acetate, and propionate are made when bacteria in the gut break down complex carbohydrate chains and other food parts that can't be digested. These SCFAs not only give colonocytes energy, but they also help keep bowels healthy and keep immune responses in check (Sepich-Poore et al., 2021). It's also interesting to know that micriobiome in the gut can make some vitamins. Some of these vitamins, like vitamin K, are needed to make energy and keep the blood from clotting (Lai et al., 2022). Along with its part in metabolism, the gut microbiome is also very important for the defence mechanism. It guides and changes immunity cells to fight germs in a healthy way while preventing harmful inflammation. Each human gut microbiome is different because of changes in food, habits, genetics, and the surroundings (Lai et al., 2022). If gut microbiota is steady, varied, and well-balanced, that means the person is healthy. When this delicate balance is disrupted, it causes dysbiosis, which raises the risk of many negative health outcomes, such as obesity, colorectal cancer, inflammatory bowel disease (IBD), and other long-term illnesses.

### 2.2.2 Microbiome Dysbiosis and Colorectal Carcinogenesis

A lot of researchers are realising that dysbiosis, or an unbalance of gut microbiota, can cause colorectal cancer to develop and spread. Rahman et al. (2022) mentioned that this microbe mismatch might help cancer develop in a number of ways that are all linked.

Chronic inflammation is one of the main ways that dysbiosis leads to CRC. The inflammation in the digestive system may get worse when harmful bacteria like *Escherichia coli* and *Fusobacterium nucleatum* are present. *Fusobacterium nucleatum* can attack inner lining cells of the intestines. It can link to E-cadherin and start $\beta$-catenin signalling. This causes more cells to divide and fewer cells to die, which are both signs that the cancer is getting aggravated. According to Elinav et al. (2019), this long-term inflammation state helps cancer and DNA changes grow.

Genotoxins are made by some bacteria in the gut. They distress DNA directly and make inflammation worse. As an example, the bacteria *Bacteroides fragilis* makes a poison that cuts DNA strands and starts cancer pathways like Wnt/$\beta$-catenin, which are very important in the early stages of colorectal cancer (Doocey et al., 2022).

The gut microbiome also have an effect on the immune system. In cases of dysbiosis, the immune system can either aid or hinder the growth of a cancer. Some bacteria are able to get past the immune system, which means tumours may grow without being stopped. On the other hand, good bacteria like *Lactobacillus* and *Bifidobacterium* may make the immune system better at fighting cancer. Rising the activity of regulatory T cells (Tregs) and tumor-infiltrating lymphocytes (TILs) does this. These cells prevent cancer from spreading (Yoo et al., 2020).

The association between the gut microbiome and colon cancer is complicated by numerous factors. This category includes interactions between different types of microbes, human genes, and external factors. Diet, drug use, and way of life are factors that can change the gut bacteria. Because of this, it is harder to figure out what part it plays in CRC. Even so, more and more evidence show that dysbiosis is linked to the development of colon cancer. This shows how important it is to keep your microbiome healthy.

### 2.2.3 Microbial Signatures as Biomarkers for CRC Detection

Many researchers are interested in using microbial fingerprints as medical biomarkers to find colorectal cancer (CRC) early because the gut microbiome has a big effect on how the disease develops. Studying the gut microbiome with new high-throughput sequencing methods like 16S rRNA gene sequencing and metagenomics, have helped the researchers find patterns of bacteria that are related to CRC (Olovo et al., 2021).

Microbes in patients with colorectal cancer (CRC) are different from those in healthy people. Olovo et al. (2021) discovered a link between colorectal cancer and harmful bacteria like *Fusobacterium nucleatum*, *Porphyromonas*, and *Peptostreptococcus*. These bacteria cause inflammation and damage to the colon. It is also linked to some good bacteria, like *Faecalibacterium* and *Roseburia*, but not as much. Microbes may have changed in ways that make it easier to find people who are more likely to get colorectal cancer promptly.

By looking at the whole genomes of populations of microbes, metagenomic methods shed light on the gut microbiome. These tests might find genes and processes that are linked to CRC. For example, CRC might be linked to genes that change the immune system or make chemicals that cause cancer. Freitas et al. (2023) did metabolomic studies that list the small chemicals that gut bacteria make, are another example. These studies shed light on the microbiome's biological processes and how they affect the health of the host.

A new way to find CRC is with the faecal microbiome test. Samples of faeces are examined to see what kinds of microbes are present. Scientists can check people for CRC, now that they know how to find risk factors in microbiome. For example, Yang et al. (2021) used the microbiome to create a model that linked clinical risk factors to the amount of certain bacterial types present. The CRC screening model in this study was more accurate and specific than regular faecal occult blood tests (FOBT).

## 2.3 Machine Learning in Biomedical Research

### 2.3.1 Overview of Machine Learning in Biomedical Research

Researchers like machine learning (ML) because it lets them look at big, complicated datasets like those from microbiome studies. In contrast to standard statistical methods, algorithms that are taught on data may find connections and patterns that would not have been observed otherwise (Hermida et al., 2022). ML is utilised widely in microbiome study for tasks like identifying diseases, find biomarkers, and classification.

Any kind of microbe can be used to make samples that machine learning systems can utilise. It is possible for these algorithms to look for small trends in very large datasets. After observing these patterns, they might be able to predict how different human and bacterial factors will interact. Furthermore, autonomous systems have the capability to continuously enhance their performance via self-learning. As their knowledge expands, they will become more adept at processing new information and making more accurate forecasts.

### 2.3.2 Machine Learning Models for CRC Prediction

Recent progress in predicting colorectal cancer has relied more on microbiome data mixed with machine learning methods. Hermida et al. (2022) showed that microbiome data has the potential to be used to train different machine learning models that can identify the risk of colorectal cancer. Two types of models that demonstrate a lot of promise as useful tools in the area are neural networks as well as ensemble methods. Some techniques for machine learning that are frequently employed to find cancer were talked about by Oh & Zhang (2020). These are Random Forest (RF), Support Vector Machines (SVM), and Logistic Regression (LR).

By mixing data on microbiome alongside additional clinical factors, machine learning methods could make the process simpler to detect CRC. It is possible that ML models may benefit from a person's medical history, lifestyle choices, DNA, and the way bacteria proliferate (Oh & Zhang, 2020). Put another way, one can acquire better, more actionable risk data. It's intriguing to learn that colon or colorectal cancer (CRC) might progress due to the intricate web of relationships between bacteria. Ensemble learning makes use of many of these ML methods to improve the prediction process. Several new approaches to learning these relationships have recently emerged.

### 2.3.3 Challenges and Limitations of ML in Microbiome Analysis

ML needs to be used to study the microbiome, but there are some issues that need to be fixed first. These other kinds of data are not like microbiome data because it is uncommon, high-dimensional, and might have noise. So that it doesn't over-fit, it also needs to be validated very well (Zhang et al., 2020). In healthcare, it's very important to know how bodily processes work from the inside out. It's even more important that ML models are easy to understand because

of this. System biology and machine learning are both tools that experts use to help them create interpretable models, such as decision trees and rule-based systems.

It might be difficult to use machine learning on microbiome data because the groups of bacteria are so complicated and varied. The microbiome can be changed by the environment, eating habits, and medical treatments. Due to this, the data might not be as consistent, which would make it harder to find trends that appear across different studies. The number of different species of bacteria in the microbiome datasets is directly related to how complicated the datasets are. Overfitting happens when a model works well on the training data but poorly on unseen data. If this occurs, the likelihood of overfitting increases. To solve these problems, researchers are looking into ways to make ML models more accurate and useful. Some of the methods in this group are feature selection, normalisation, and dimensionality reduction. The goal of making machine learning models interpretable is to help doctors make better choices about colorectal cancer patients (Malla et al., 2019).

## 2.4 Gaps in Existing Research

Testing for CRC is getting better, and more researchers are interested in learning more about the microbiome, but there are still big gaps in the studies that have been already conducted. These gaps need to be filled in order to make microbiome-based tests more useful for public health.

Colorectal cancer (CRC) can be detected with colonoscopy, but some people may not be able to go through with the process because it is painful. It is very important to find new, accurate measures that don't require surgery, since FOBT and other non-invasive methods aren't very dependable. More research is needed to find out how well microbiome-based screening works in different groups of people and to add these markers to current screening methods (Kim et al., 2019).

Many factors can cause the microbiome to change or evolve at any time. To put it another way, the results from the different study groups are very different. Due to this difference, it is difficult to find broad bacterial biomarkers that may help diagnose CRC. It's hard to compare studies because people don't agree on the best way to get, store, and look at study samples (Saus et al., 2019).

Even though ML models have shown potential in CRC detection, there are still a lot of challenges that hinder their widely usage. Some of these problems are the need for large well-annotated datasets, the fact that models are hard to understand, and the fact that microbiome data doesn't work with other health records. Building and testing machine learning models can use a lot of computational resources, especially for big projects (Gupta et al., 2020).

For many reasons, including having to get government approval, being very expensive, and teaching doctors how to use microbiome-based screening, it is hard to make it work in real life. This can be risky to use the model for clinical purpose as the results can be manipulated which can lead to moral consequences.

# Chapter 3

# Methodology

## 3.1 Introduction

This chapter takes a deep dive into the approach taken for developing the predictive model for CRC using microbiome data. The chapter starts by detailing the data collection and preprocessing steps that were taken to transform the raw microbiome data into a suitable format for the analysis. The discussion then shifts from preprocessing to the data analysis (EDA and Diversity Analysis) and the core machine learning models implemented: Random Forest, LightGBM, and Logistic Regression. It talks about how each model functions and how it can potentially aid in the prediction of CRC. The chapter further introduces an innovative ensemble model that utilises these algorithms by capitalising on their individual strengths to achieve good predictive accuracy and generalisability. The later part of the chapter covers the Hyperparameter tuning of the models using Optuna along with the matrices used for the evaluation.

## 3.2 Data Collection

The microbiome dataset used in this study was provided by the supervisors and is based on the data collected through the NHS Bowel Cancer Screening Programme (NHSBCSP) which used the gFOBT test to collect the feacal samples from the population who were part of the routine colorectal cancer screening process. The population consisted of people who were tested blood-negative (where the blood was not found in the faeces), diagnosed with colorectal cancer, adenoma (low, intermediate, high; based on the growth of benign tumor), normal colonoscopy results (where the blood was found but it was not related to any disease), and non-neoplastic conditions. This ensured that the dataset consists of a wider range of microbiome compositions and clinical outcomes which is important for developing robust predictive models (Young et al., 2021).

In order to preserve sample integrity and relevance, samples were kept at room temperature and stored under similar settings until DNA extraction to mirror the situation in the real world where immediate freezing/refrigeration is not feasible. Every step of the data collection was

validated properly, to ensure that the results were appropriate for the national screening setting, through experiments that confirmed that storing the samples at room temperature had minimal effect on the microbiome structure (Young et al., 2021).

To effectively capture the diversity and abundance of microbial communities in the faecal samples, the DNA extraction process was carried out meticulously using QIAmp DNA Mini Kit protocol along with 16S rRNA gene V4 amplicon sequencing for taxonomic profiling (Young et al., 2021).

The utilisation of NHSBCSP samples ensured a realistic reflection of the conditions under which faecal microbiome analysis could be integrated into screening programmes as it makes sure that the findings are robust and applicable. Also, the use of the gFOBT and its specific handling circumstances is also consistent with real-world public health practices, increasing the study results' generalisability to larger populations and contexts.

## 3.3 Data Preprocessing

After the data collection, the next step was to prepare the data for analysis and modelling. This step was necessary as the dataset contained a high number of zero values, which could either result from stochastic variations, like the error caused during the data collection, or the presence of the specific microbiome being too low to be captured by the data collection technique used, or it could represent the genuine absence of microbial species. The dataset initially contained features categorised across several taxonomic levels: Domain (D_0), Phylum (D_1), Class (D_2), Order (D_3), Family (D_4), and Genus (D_5).

Domain, the D_0 level, is the broadest classification, typically representing groups like "Bacteria" or "Archaea" in the microbiome data—microbiomes belonging to phylum (D_1) according to their primary structural and functional characteristics. The Class (D_2) further refines this classification based on more specific characteristics. Order (D_3) includes organisms with shared features which are then divided into Families at the D_4 level. The last level of the dataset, which is the D_5 level, represents the Genus, the most specific grouping of closely related species.

Initially, the dataset contained 648 features, leading to a high number of zero values. To refine the data and reduce the dimensionality for better modelling, microbiomes were aggregated which shared the same D_4 level. The aggregation was based on the observation that the microbiomes at the genus level sharing the same family, generally exhibited the same behaviour. It is similar to the animal kingdom, for example, the *Cervidae* family (the family of deer) contains around 16 genera and around 51 species such as white-tailed deer (*Odocoileus virginianus*) and mule deer (*Odocoileus hemionus*) that share behavioural traits like quick response to stimuli and similar foraging habits. Further refinement involved applying a 10% threshold to filter out features present in less than 10% of the samples in the dataset, thereby decreasing the noise as well as the dimension of the data, enhancing the overall data quality, and facilitating more effective

analysis and predictive model development. Another significant preprocessing that took place was that initially, the adenoma group was divided into three groups as: Low, Intermediate, and High, so for better analysis, these three groups were combined under one single group called 'adenoma'.

The final step of the data preprocessing was transforming the abundance count data into relative abundance ratio. It was a necessary as microbiome data is compositional, meaning the abundance of each microbiome is relative to the total count in a sample. Absolute counts are less informative as they are influenced by total sequencing depth or other sample-specific factors, therefore, converting to relative abundance ensures that the data reflects the proportional representation of each microbe which allows for meaningful comparisons across samples. This provided a more accurate depiction of the microbial community structure.

## 3.4 Exploratory Data Analysis (EDA)

The preprocessing of the data set the foundation for the EDA. It was conducted to gain a detailed understanding of the microbiome dataset to ensure its readiness for further analysis. It is an important step before developing the models as it uncovers the underlying patters in the dataset and identify any anomalies (if present) that could impact the model performance.

The EDA was conducted in two parts, being, numerical and graphical summaries. Numerical summaries including measures like mean, median, standard deviation, variance, and IQR range, provide an overview of the central tendency, variability, and distribution of the data. They are quite a significant part of the EDA as they help in identifying skewness, outliers, and understanding the shape of the data distribution, which helps in making informative decisions about potential transformations that might be required.

Graphical summaries, was also an integral part of the data understanding process. Plots like bar plot and violin plot were utilised to gain further information about the data. Bar graphs were used to examine categorical data and understand the frequency distribution of different microbial categories that helped in identifying class imbalances, signalling potential biases. Also, violin plots were generated to gain insight into the spread and quartiles of the data.

A special focus was placed on looking for data inconsistencies during the EDA, like missing values, incorrect data entries, or inconsistent data points, which are crucial to identify and address, as they can lead to misleading conclusions. Overall, the combination of numerical and graphical summaries provided extensive insight into the dataset's quality and structure aiding in identifying any potential anomalies in the microbiome dataset and setting up the foundation for building reliable predictive models.

## 3.5 Diversity Analysis

Diversity Analysis is essential in microbiome research to understand the complexity and variation within the microbiome colonies across different environments or health conditions. The Diversity Analysis can be broken down into three sub-analysis: Alpha Diversity Analysis, Beta Diversity Analysis, and Differential Abundance Analysis. The `skbio` Python package was used to conduct the diversity analysis.

### 3.5.1 Alpha Diversity Analysis

Alpha diversity plays an important role in the analysis as it measures the diversity within a single sample, focusing on both the richness (number of species) and evenness (distribution of abundances) of microbial colonies. Alpha diversity was found using the well-known Shannon Diversity Index, which is calculated using the equation 3.1:

$$H' = -\sum (p_i \times \ln(p_i)) \tag{3.1}$$

where $p_i$ represents the proportion of each species in the sample (Shannon, 1948). The reason behind choosing the Shannon Index was due to its assumption that all species in the community contribute to its diversity which makes the index sensitive to both abundant and rare species. This is a particularly important assumption as in microbiome studies, rare species can have significant impacts or associations with health conditions.

### 3.5.2 Beta Diversity Analysis

Beta diversity's role in this analysis was to examine the differences in microbial community composition between the samples, providing insights into the methods by which microbial populations adapt to various environments and conditions. The Bray-Curtis dissimilarity matrix aided in the analysis of beta diversity, which can be calculated using the equation 3.2:

$$BC_{ij} = \frac{\sum |x_{ik} - x_{jk}|}{\sum (x_{ik} + x_{jk})} \tag{3.2}$$

where $x_{ik}$ and $x_{jk}$ are the counts of species $k$ in samples $i$ and $j$, respectively (Bray & Curtis, 1957). The primary assumption of this matric is that it assumes that dissimilarity between the communities is proportional to both species presence and abundance differences which makes it specifically effective for the comparison of microbial communities with varying species distributions and abundances. To further evaluate it, Principal Coordinates Analysis (PCoA) was employed to visualise the differences in 2D space. Finally, PERMANOVA (Permutational Multivariate Analysis of Variance) was conducted using `skbio.stats.distance` package to test the differences statistically, assuming that the group differences are detectable in the multivariate space defined by the dissimilarity matrix (Anderson, 2001).

### 3.5.3 Differential Abundance Analysis

Differential abundance analysis was the last step of the diversity analysis which was conducted to identify specific microbial taxa that differ significantly in abundance between different health groups. The Wilcoxon rank-sum test, also known as the Mann-Whitney U test, was conducted using `mannwhitneyu` from `scipy.stats` Python package. The reason behind choosing the U test was, that it is non-parametric and does not require data to be normally distributed (Mann & Whitney, 1947), which is a frequent case in the microbiome data, caused because of the rare taxa. The test conducts a comparative analysis of the distribution of two independent groups to determine if there is a significant difference in their median values.

Along with this test, the Bonferroni correction was applied to the p-values which takes into account multiple comparisons, reducing the chances of false positives by making the significant threshold more strict. By applying these adjustments, the findings tend to remain statistically robust and preserve the reliability of the results despite the numerous tests that are conducted (Benjamini & Hochberg, 1995), which leads to the identification of the most statistically significant differences.

These sub-analysis allows the diversity analysis to provide robust insights into the microbiome's role in various health status, capturing both the diversity within individual samples and the differences between groups.

## 3.6 Machine Learning Models

### 3.6.1 Random Forest

Random Forest has become a standard in the toolbox of data scientists and statisticians due to its versatility and robustness. It is based on the ensemble learning method and is primarily used for classification and regression problems. Random Forest works by building a large number of decision trees during the training phase and using their results to produce a combined result to improve the overall predicted performance. Breiman (2001) invented this method, which is highly praised for its ability to handle complex datasets with numerous features, just like the microbiome dataset.

#### Assumptions

Just like any other model, Random Forest has its key assumptions too, which are critical to its performance and utility.

One of the primary assumptions of Random Forest is that the individual trees created in the process of training are relatively independent of each other. This independence is achieved through the process of bootstrap sampling and random feature selection which makes sure that each tree of the forest is trained on a different subset of the dataset and takes into account different features at each node. The primary idea behind this is that each individual tree may

be a weak learner, the ensemble of trees, because of their independence, can become a strong learner when united (Breiman, 2001).

Along with Independence, Random Forest assumes that the input features contribute at different degrees in determining the output. This assumption is evaluated by determining how much the prediction error increases when the values of one feature are permuted while the values of other features remain constant. The features that increase the prediction error significantly, are deemed important (Strobl et al., 2007).

Another key assumption of Random Forest is that by aggregating predictions from several decision trees, the final model will be less prone to overfitting than a single decision tree. Random Forest can better generalise to new data by averaging the outcomes of numerous uncorrelated trees, reducing the model's variance while not significantly increasing the bias (Breiman, 2001).

In the real world, most of the relationships are non-linear which makes Random Forest an ideal choice as it does not assume a linear relationship between the input features and target variable, which is particularly crucial for complex datasets like microbiome data, where the interaction between the microbial species can be non-linear and intricate (Cutler et al., 2007).

**Explanation of the Random Forest Algorithm:**

To delve deeper into the application of Random Forest to the microbiome data, the algorithm can be divided into three major steps, which are as follows:

1. **Bootstrap Sampling:** Support there are $n$ samples (2252 in this case), each with $p$ features (49) representing microbiome ratios, age, and gender. Bootstrap sampling allows the creation of different subsets by randomly selecting n samples with replacements from the original dataset. This ensures that each tree is trained on a different subset, introducing variation and reducing the overall correlation among the trees.

2. **Decision Tree Construction:** In the Random Forest, a decision tree is created for each of the bootstrap samples, but unlike a typical decision tree, only a random subset of features is evaluated. This randomness reduces the correlation between the trees in the ensemble, boosting the overall performance. Generally, for classification tasks, Gini impurity is utilised to determine the appropriate split as it assesses how pure or homogeneous the classes are within a node (Breiman, 2001).

3. **Prediction Consolidation:** After building all the trees, Random Forest combines their individual predictions to arrive at the final outcome. For classification problems, like in the case of CRC prediction, a majority vote is used to select the class predicted by the most trees. On the other hand, for regression, the final result is calculated by taking the arithmetic mean of all the trees' predictions (Breiman, 2001).

### Mathematical Framework for Random Forest Application

The Random Forest can be mathematically formalized in the context of the microbiome dataset, which has been transformed from count data to ratio data to maintain its compositional nature. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ represent the dataset, where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the feature vector of microbiome ratios for the $i$-th sample, and $y_i$ is the class label (CRC, adenoma, or negative).

1. Create $B$ bootstrapped samples $\mathcal{D}_b$ for $b = 1, 2, \ldots, B$, each containing $n$ samples drawn with replacement from $\mathcal{D}$.

2. For each bootstrapped sample $\mathcal{D}_b$, a decision tree $T_b$ is grown by recursively splitting the nodes. At each node, a random subset of $m$ microbiome features is selected from the $p$ available features. The best split among these $m$ features is determined using a criterion such as Gini impurity. The Gini impurity $G(t)$ at a node $t$ is defined as:

$$G(t) = \sum_{i=1}^{c} p_i(1 - p_i) \tag{3.3}$$

   where $p_i$ is the proportion of samples belonging to class $i$ at node $t$, and $c$ is the total number of classes. This criterion is extensively discussed by Breiman (2001).

3. For a new sample $\mathbf{x}$, obtain the prediction from each tree $T_b$ as follows:

$$\hat{y}_b = T_b(\mathbf{x}) \tag{3.4}$$

   The final prediction $\hat{y}$ is then determined by majority voting:

$$\hat{y} = \text{mode}\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_B\} \tag{3.5}$$

#### Accuracy, Convergence, and Generalization Error

The accuracy and generalization of Random Forest are characterized by the margin function and generalization error. The margin function for a Random Forest, as defined by Breiman (2001), can be expressed as:

$$mg(X, Y) = \text{av}_k \, I(h_k(X) = Y) - \max_{j \neq Y} \text{av}_k \, I(h_k(X) = j) \tag{3.6}$$

where $h_k(X)$ denotes the prediction of the $k$-th tree, $I(\cdot)$ is the indicator function, and $Y$ is the true class label. The margin function measures the difference between the average number of votes for the correct class and the maximum average vote for any other class. A larger margin signifies higher confidence in the classification (Breiman, 2001).

The generalization error $PE^*$, also discussed by Breiman (2001), is defined as the probability that the margin is negative:

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{3.7}$$

As the number of trees increases, the generalization error converges to a limiting value. This convergence can be explained by the Strong Law of Large Numbers, as shown by Breiman (2001):

$$PE^* \to P_{X,Y}\left(P_\Theta\left(h(X,\Theta) = Y\right) - \max_{j \neq Y} P_\Theta\left(h(X,\Theta) = j\right) < 0\right) \tag{3.8}$$

where $\Theta$ represents the randomness in the tree construction, including the selection of bootstrap samples and features at each node.

**Relevance to the Dataset**

In the context of microbiome data with numerous features, the Random Forest model will certainly provide different advantages. The creation of decision trees with different configurations using bootstrap sampling and feature selection allows it to capture the various facets of the data, which leads the model to yield more robust and reliable predictions.

The margin function serves as a measure of classification confidence. For example, when discriminating between CRC and Adenoma, a wide positive margin indicates high confidence in the correct classification of a sample, whereas a negative margin may indicate a probable misclassification, further contributing to the generalisation error. The generalisation error tends to converge as the number of trees grows. Due to this convergence, stability and reliability are maintained even in the face of the inherent unpredictability of microbiome data, preventing the model's performance from being significantly altered by the addition of additional trees.

### 3.6.2 Logistic Regression with L2 Regularisation

Logistic regression is a powerful yet, one of the simplest statistical methods widely used among statisticians, primarily for binary classification tasks. The basic functionality of it is that it models the probability that a given input belongs to a specific class using the logistic function, also known as the sigmoid function. Logistic regression is intended to handle categorical outcomes, usually binary, and is particularly effective in situations where the relationship between the dependent and independent variables can be modeled as a linear relationship in the log-odds (logit) space. One of the reasons for its wide usage is its interpretability and effectiveness in various domains, including health data analysis.

### Assumptions

Just like any other model to be accurate and useful, logistic regression also works on several assumptions. The primary assumption of logistic regression is that the logit is linear, which means that the natural logarithm of the odds of the dependent variable is a linear function of the independent variables. While the relationship between the predictors and outcome probability may be non-linear in nature, the relationship between predictors and log-odds should be linear (Hosmer et al., 2013). Additionally, the observations in the dataset must be independent of one another as any dependence between the data can result in biased parameter estimations and inaccurate predictions (Menard, 2002).

Logistic regression assumes that the predictors are not perfectly correlated as this might make it difficult to estimate the model's coefficients effectively which further leads to inflated standard errors and reduced reliability of the results (Menard, 2002). While logistic regression may be used for multinomial regression with more than two categories, its primary use is binary classification problems (Hosmer et al., 2013).

Logistic regression often requires a large sample size to produce accurate results, especially when the model includes infrequent occurrences or a large number of predictors. A large sample size allows the model's estimates to be stable and generalisable (Hosmer et al., 2013).

### Algorithm and Mathematical Application to Microbiome Data

For this research, Logistic Regression is used with L2 Regularisation which is also known as Ridge Regression, an extension or modified version of the basic logistic regression model that includes a penalty term, which helps the model prevent overfitting. In high-dimensional datasets, like microbiome data with a high number of features, this regularisation is very helpful. The algorithm can be broken down into three parts, which are as follows:

1. **Logistic Function and Probability Estimation:** Logistic regression utilises logistic function that models the probability $P(y = 1 \mid x)$ that a given sample $x$ belongs to a particular class (eg: CRC, Adenoma or Negative). The logistic function is defined as:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w \cdot x + b)}} \tag{3.9}$$

   where $w$ is the weight vector, $x$ is the feature vector (e.g., microbial species ratios), and $b$ is the bias term. The logistic function aims to transform the output into a value between 0 and 1, making it suitable for modelling probabilities (Hosmer et al., 2013).

2. **Log-Likelihood and Cost Function:** The parameters $w$ and $b$ are estimated by maximising the log-likelihood function which quantifies the model's fit to the data. For a dataset with $n$ observations, the log-likelihood can be calculated as:

$$\text{Log-Likelihood} = \sum_{i=1}^{n} [y_i \log P(y_i = 1 \mid x_i) + (1 - y_i) \log(1 - P(y_i = 1 \mid x_i))] \tag{3.10}$$

When applying L2 regularisation, the model does not directly maximises this likelihood, instead it reduces the cost function which contains a regularisation term as explained by Ng (2004):

$$\text{C.F} = -\frac{1}{n} \sum_{i=1}^{n} [y_i \log P(y_i = 1 \mid x_i) + (1 - y_i) \log(1 - P(y_i = 1 \mid x_i))] + \frac{\lambda}{2} \sum_{j=1}^{p} w_j^2 \tag{3.11}$$

Where the first term is the negative log-likelihood and the second term is the L2 penalty term, where $\lambda$ is the regularisation parameter which controls the extent of penalisation, and $p$ is the number of features. By decreasing big coefficients towards the zero, the regularisation term improves the model's capacity to generalise and help prevent overfitting (Ng, 2004).

3. **Optimization and Gradient Descent:** Gradient descent is used to optimise the cost function by iteratively updating the model parameters $w$ and $b$ to minimise the cost function. The following formula is used to calculate the cost function's gradients with respect to $w$ and $b$:

$$\frac{\partial \text{Cost Function}}{\partial w_j} = -\frac{1}{n} \sum_{i=1}^{n} [y_i - P(y_i = 1 \mid x_i)] \, x_{ij} + \lambda w_j \tag{3.12}$$

$$\frac{\partial \text{Cost Function}}{\partial b} = -\frac{1}{n} \sum_{i=1}^{n} [y_i - P(y_i = 1 \mid x_i)] \tag{3.13}$$

The gradients help in parameter updation during each iteration to make sure that the model converges to a solution that minimises the cost function to balance between the regularisation and model fit (Ng, 2004).

**Relevance to Microbiome Data**

Logistic regression with L2 regularisation is quite effective for microbiome data as the number of features is very high. Due to regularisation, the model's chances to overfit the data reduce as it penalises large coefficients, making it more generalisable. For microbiome data, logistic regression is capable of predicting the likelihood of health status, such as CRC, based on the microbiome profile. Along with this, the coefficients of the model highlight the feature's importance by revealing how changes in a specific microbial species affect the likelihood of different

health outcomes, which can help in identifying potential biomarkers for early CRC detection.

### 3.6.3 LightGBM

LightGBM stands for Light Gradient Boosting Machine, it is a gradient boosting algorithm that utilises decision tree algorithms (just like Random Forest) to perform machine learning tasks efficiently. It is particularly optimised for speed and accuracy when dealing with high-dimensional data with numerous features. This makes it especially useful for complex datasets like the microbiome dataset, which contains the relative abundance of different microbial species (Ke et al., 2017).

#### Assumptions

LightGBM is designed around a set of foundational assumptions that help it boost its performance and efficiency in machine learning applications. It is assumed by LightGBM that the decision trees within the ensemble are initially independent and sequentially constructed where each tree is constructed to correct the errors of the previous tree. This refinement allows the model to continuously improve by focussing on residual error, maximising the efficiency of the framework (Friedman, 2001).

LightGBM also relies on additive and sequential model training, where each new tree improves the ensemble by focusing on the errors made by the previous trees. This approach is particularly adept at handling complex/ non-linear relationships and high-dimensional data, resulting in an overall improvement in the model's accuracy as the iterations progress (Ke et al., 2017).

The algorithm is also optimised for sparse features as it assumes that, there are features in the dataset that will have many zeroes or missing values, which is the case with the microbiome data in which some microbial species may be absent from many samples. Additionally, it is also optimised to handle categorical features without the need for any extensive preprocessing. Thanks to these core assumptions, LightGBM effectively manages diverse data types, ensuring that it delivers robust and accurate models even in complex situations.

#### Algorithm and Mathematical Application to Microbiome Data

LightGBM works by constructing a series of decision trees where each tree aims to correct the errors of the previous ones. This is particularly advantageous for microbiome data as it can handle high number of features with complex relationships. LightGBM works on three main concepts which are as follows:

1. **Gradient Boosting Framework:** LightGBM takes a bold approach with its gradient boosting framework, where each tree is not just another layer but a purposeful addition. Trees are built sequentially, each one learning from the mistakes of the previous ones.

The focus is clear:minimise the logistic loss function, used for this research, for binary classification:

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{3.14}$$

As described by Ke et al. (2017), this loss function is minimized using gradient descent, where each new tree is trained to predict the negative gradient of the loss function with respect to the current model's predictions.

2. **Leaf-wise Growth Strategy:** The traditional gradient boosting methods work on level-wise trees expansion, but this is not the case with LightGBM. It focuses on growing the trees by expanding the leaf which leads to the most significant reduction in the loss function, which is represented by:

$$\Delta L = L(T_{\text{old}}) - L(T_{\text{new}}) \tag{3.15}$$

where $T_{\text{old}}$ represents the tree before adding the new leaf, and $T_{\text{new}}$ represents the tree after adding it. As explained by Ke et al. (2017) it allows LightGBM to achieve higher accuracy by focusing on the most significant splits.

3. **Regularization Techniques:** For this project, both L1 (Lasso) and L2 (Ridge) regularization are utilised for LightGBM. The objective function including these regularization terms can be expressed as:

$$\text{Objective Function} = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \lambda_{\text{L1}} \sum_{j=1}^{p} |w_j| + \frac{\lambda_{\text{L2}}}{2} \sum_{j=1}^{p} w_j^2 \tag{3.16}$$

where $L(y_i, \hat{y}_i)$ represents the loss function (e.g., logistic loss for binary classification), $\lambda_{\text{L1}}$ and $\lambda_{\text{L2}}$ are the L1 and L2 regularization parameters, $w_j$ are the leaf weights, and $p$ is the number of leaves.

**Relevance to Microbiome Data**

LightGBM would be an ideal model, just like the random forest, for the microbiome data as it has the ability to handle high-dimensional and sparse data efficiently. As the microbiome dataset is of high dimensionality and consists of complex relationships, LightGBM's approach to leaf-wise growth would allow it to focus on the most informative splits, which would result in the improvement of predictive accuracy. Also, LightGBM can efficiently handle categorical features and use regularisation techniques making it an ideal choice for this study, where the data variability and diverse measurements are common. The regularisation would also ensure that the model doesn't overfit.

### 3.6.4 Ensemble Model

The ensemble model, utilised for this study combines Random Forest, LightGBM, and Logistic Regression to improve predictive power and robustness. Ensemble models are often used in the domain of machine learning as they integrate multiple models to exploit their strengths and mitigate their weaknesses.

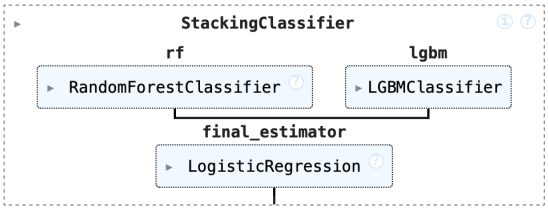



*Figure 3.1*: Ensemble model architecture.

From Fig. 3.1 it can be seen that in this research's ensemble model, Random Forest and LightGBM will serve the purpose of base models for capturing the various patterns in the data, while Logistic Regression functions as the meta-model for the ensemble which integrates the predictions of the base models into a final output.

#### Method and Application to Microbiome Data

The ensemble model starts by training Random Forest and LightGBM individually on the same microbiome dataset. Where Random Forest uses multiple decision trees, trained on different subsets of the data to reduce overfitting and enhance gernalisability, at the same time, LightGBM with its gradient boosting mechanism builds trees sequentially to efficiently manage the dataset, and identify the most crucial splits.

After the base models generate their predictions, these outputs are fed forward to the Logistic Regression model, which uses these predictions as additional input features and learns to combine them into a final prediction. This method offers a clear explanation of how both the base models contribute to the final prediction, balancing the advantages of Random Forest and LightGBM.

#### Relevance to Microbiome Data

This ensemble model has a very high potential for analysing microbiome data due to its ability to handle complexity and variability. The utilisation of Random Forest and LightGBM ensures reduction in variance, noise data management, and improved accuracy through efficient learning. On the other hand, Linear Regression ensures balanced and interpretable final output by using the insights from both the base models. Overall, this makes the ensemble model a suitable option for predicting colorectal cancer and understanding microbial patterns related to CRC.

## 3.7 Hyperparameter Tuning and Evaluation Matrices

### Overview of Hyperparameter Tuning

Optimising a model is an important step in machine learning, to make the model more robust and generalisable. This can be achieved using hyperparameter tuning. For this particular study, hyperparameter tuning is conducted using a Python package called `Optuna`, which is considered one of the best hyperparameter optimisation framework. The reason behind choosing `Optuna` is its efficiency and flexibility. It operates by employing an advanced search algorithm, called Tree-structured Parzen Estimator, that adaptively explores the hyperparameter space to select the best values for the parameters. The advanced search algorithm allows `Optuna` to perform better than other traditional methods like grid or random search (Akiba et al., 2019), making it a better option for hyperparameter tuning of the models.

### Baseline Evaluation Metrics

Before applying hyperparameter tuning, the base models are evaluated using multiple matrices to determine a baseline performance. These matrices include, Confusion Matrix, Accuracy, Precision, Recall, F1-Score, and AUC-ROC. These matrices provide a strong evaluation of the model altogether. The confusion matrix gives a detailed breakdown of true positives, false positives, true negatives, and false negatives, giving a thorough overview of the model's performance (Powers, 2011). To assess the model's sensitivity in capturing all the relevant instances and its ability to accurately identify positive cases, respectively, precision and recall are essential matrices.

The F1-score which is the harmonic mean of precision and recall, offers a balanced evaluation, which is particularly useful for datasets with class imbalances, a clear case in the microbiome data. On the other hand, AUC-ROC scores and curves provide a robust judgment about the model's ability to distinguish between classes across various thresholds, making it effective for uneven class distributions just like F1-score (Bradley, 1997).

### Post-Tuning Evaluation and Cross-Validation

After fine-tuning the models using Optuna, the models are reassessed using the same matrices to compare against the baseline performance and prevent overfitting. A 10-fold cross-validation is also applied to the refined models to ensure the models' consistency and reliability. Cross-validation provides a robust measure of the model's generalisation performance by dividing the data into various subsets and repeating the whole training and testing process (Kohavi, 1995). This approach makes sure that the models are robust and capable of generalising effectively to the unseen data.

# Chapter 4

# Results

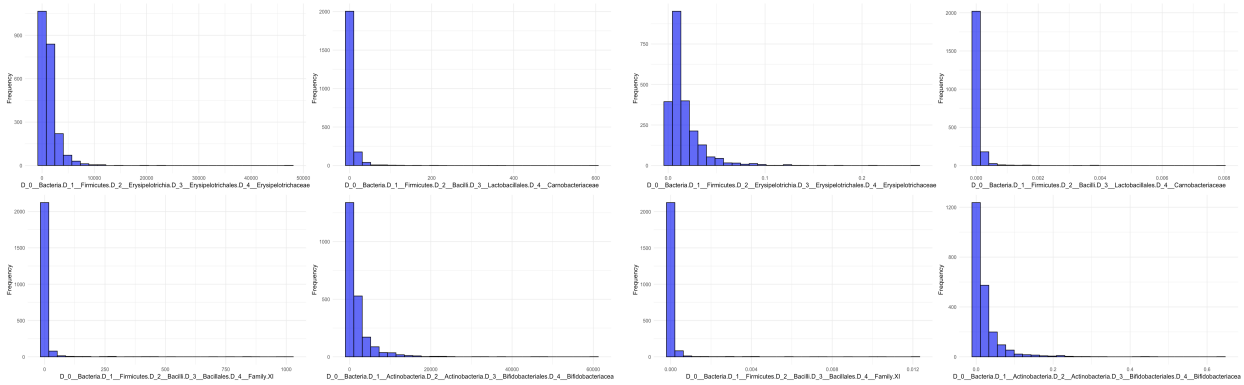
## 4.1 Data Preprocessing Results

The data quality was significantly improved because of the methods employed for data preprocessing, which were crucial in laying the groundwork for accurate modelling and analysis. The raw microbiome dataset was primarily made up of zero values, totalling 1,239,393, around 85% of the data. It was crucial to address these zeroes as a part of preprocessing to ensure the suitability of the data for further analysis.

The first major result of the preprocessing was observed after aggregating the data at the D_4 level, or Family level. The aggregation resulted in the reduction of the dataset's dimensionality from 648 to 195 features which also reduced the zeroes to 362,832. This step effectively reduced the sparsity within the dataset and the decrease in zeroes assisted in addressing the sparsity issue, which frequently complicates statistical analysis and can produce biased conclusions. By focusing on broader taxonomic categories, the microbiome dataset became more manageable, thus enhancing the data quality.

Further, the zeroes were handled by applying a 10% threshold which removed the microbes present in less than 10% of the samples. The threshold helped in further bringing down the number of features from 195 to 51 and reducing the zeroes drastically to 40,812, filtering out the noise and less informative features that could distort the analysis, ensuring that only the most relevant features were included in the analysis, therefore minimising the risk of overfitting.

The transformation of count data to relative abundance ratios produced another important outcome in the process.



(a) Before transformation

(b) After transformation

*Figure 4.1*: Figures (a) and (b) represents the distribution of random four features.

The histograms from Fig. 4.1 represent the overall distribution of the features, before and after the transformation. It can be observed from Fig. 4.1 that the overall distribution of the features remained largely unchanged, indicating that the conversion to ratio data preserved the original variability while adjusting for the compositional nature of the data. It was an important transformation as it aided in maintaining the data's integrity and making sure that the comparisons between the samples were meaningful.

The preprocessing steps effectively refined the dataset, reducing noise and ensuring the suitability of the data for robust analysis.

## 4.2 Exploratory Data Analysis Results

### 4.2.1 Numerical Summaries

The exploratory data analysis began with numerical summaries to understand the distribution and central tendency of the participants' age across different groups. The overall age distribution for the dataset showed a mean age of 66.89 years with a standard deviation of 4.70 years, suggesting a relatively homogeneous age distribution within the studied population. The median age is slightly lower at 67 years, indicating a slight left skew in the data.

From Table 4.1, distinct patterns can be observed. the mean ages vary slightly, with "cancer" groups showing the highest mean age (68.3 for females and 68.0 for males) and the "high" risk groups showing lower means (66.2 for females and 65.5 for males). Variances are generally higher in "cancer" groups (30.9 for females, 22.5 for males), reflecting greater age spread in these categories. The gender distribution reveals a higher proportion of males in the "cancer" (67.2%) and "high" (76.4%) groups, while females predominate in the "negative" (58.2%) and "normal_colonoscopy" (48.3%) groups, indicating potential gender-based demographic differences.

| Group | Mean | Var. | SD | Q1 | Q2 | Q3 | IQR |
|---|---|---|---|---|---|---|---|
| **cancer_F** | 68.3 | 30.9 | 5.56 | 64 | 69 | 72 | 8 |
| **cancer_M** | 68.0 | 22.5 | 4.74 | 64 | 68 | 72 | 8 |
| **high_F** | 66.2 | 17.9 | 4.23 | 62 | 66 | 69 | 7 |
| **high_M** | 65.5 | 20.7 | 4.55 | 62 | 65 | 69 | 7 |
| **intermediate_F** | 65.3 | 20.5 | 4.53 | 60 | 66 | 68 | 8 |
| **intermediate_M** | 66.0 | 23.9 | 4.89 | 62 | 66 | 69.8 | 7.75 |
| **low_F** | 67.3 | 21.1 | 4.59 | 64 | 68 | 71 | 7 |
| **low_M** | 67.2 | 20.5 | 4.53 | 64 | 68 | 71 | 7 |
| **negative_F** | 66.9 | 19.0 | 4.35 | 64 | 67 | 70 | 6 |
| **negative_M** | 67.2 | 21.2 | 4.60 | 64 | 68 | 71 | 7 |
| **normal_colonoscopy_F** | 66.6 | 19.8 | 4.45 | 62 | 67 | 70 | 8 |
| **normal_colonoscopy_M** | 66.7 | 17.7 | 4.20 | 64 | 66 | 70 | 6 |
| **other_F** | 66.3 | 21.5 | 4.64 | 62 | 66 | 70 | 8 |
| **other_M** | 67.0 | 23.4 | 4.83 | 62 | 67 | 71 | 9 |

*Table 4.1*: Summary Statistics of Age by Group and Gender

These findings (From Table 4.1) suggest potential age and gender-specific patterns in disease prevalence and risk, emphasizing the importance of demographic factors in microbiome-based studies. The variance and standard deviation across groups underscore the importance of accounting for age variability when interpreting microbiome profiles in these populations.

### 4.2.2 Graphical Summaries

The graphical summaries provided further insights into the age and gender distributions across various groups, enhancing the understanding of demographic variations.

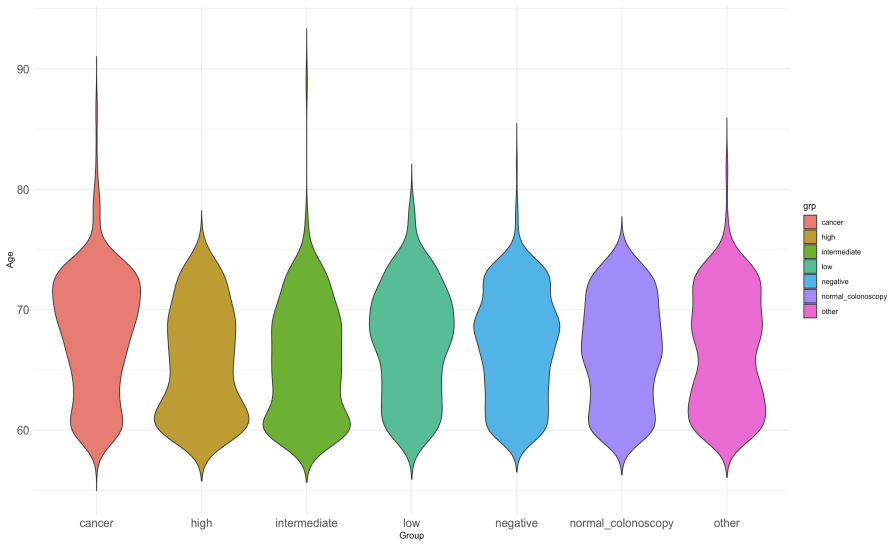



*Figure 4.2*: Violin Plot of Age Distribution by Group

The violin plot in Fig. 4.2 reveals that the ”cancer” group has a wider distribution with a peak around the upper age range which means a higher concentration of older individuals. On the other hand, the distributions of the ”high” and ”normal_colonoscopy” are narrower, reflecting more homogeneous age profiles. This highlights the age-related variation in cancer prevalence

and further supports the numerical summaries.

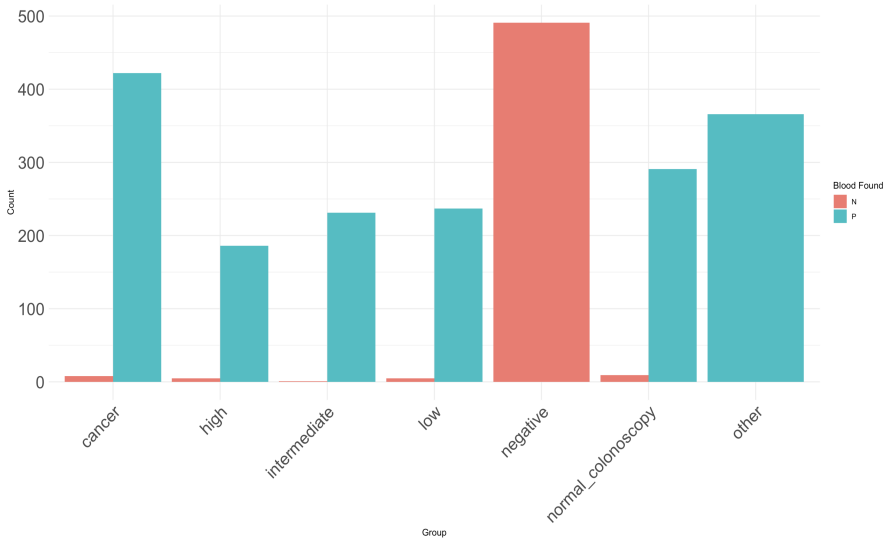



*Figure 4.3*: Bar Plot of Group by Blood Found in Faeces (P/N)

Fig. 4.3 reveals a data integrity issue in the dataset. It can be observed from the plot that ”cancer”, ”low”, ”intermediate,” and ”high” groups which inherently require the presence of blood in faeces for their classification through colonoscopy, showed entries in the ”Negative” bars, indicating that the blood was not found in those samples. This disparity shows that there may have been a mistake made in the data entry process as these health statuses should not contain any ”Negative” labels. It is important to change these entries to ”Positive” before developing predictive models. This would result in preserving the validity of the dataset and the accuracy of the conclusions derived from it.

## 4.3 Diversity Analysis Results

### 4.3.1 Alpha Diversity Results

The analysis of alpha diversity was conducted using the Shannon Diversity Index (SDI) which measures both the richness as well as evenness of microbial species within a sample. The Shannon Diversity Index provided valuable insights into the microbiome diversity across different health statuses.

From Table 4.2, it can be observed that the mean SDI values are relatively similar across groups, indicating only slight variations in diversity. Specifically, the cancer group had the highest mean diversity (1.88), and ’other’ (1.79) being the lowest. This suggests that, on average, the microbiome in CRC patients might be slightly more diverse when compared to other groups. The median values support this trend, with the CRC group again showing the highest median diversity of 1.88, and again ’other’ being the lowest (1.79). The consistency between the mean and median indicates that the diversity distribution within each group is relatively symmetrical,

| Group | Mean | Median | Mode |
|---|---|---|---|
| **Adenoma** | 1.794173 | 1.811657 | 0.390115 |
| **Cancer** | 1.884239 | 1.880966 | 0.863512 |
| **Blood-Negative** | 1.838860 | 1.840855 | 1.106315 |
| **Normal Colonoscopy** | 1.805913 | 1.857718 | 0.214767 |
| **Other** | 1.792339 | 1.796881 | 0.392459 |

*Table 4.2*: Mean, Median, and Mode Shannon Index for different groups.

although the CRC group tends to have a marginally high central tendency. The mode values provide additional insight into the distribution of diversity within the groups. The cancer group has a higher mode of 0.86 compared to adenoma, 'other', and normal colonoscopy which shows that although the cancer group exhibits a generally higher diversity, there is still a significant proportion of samples with lower diversity values.

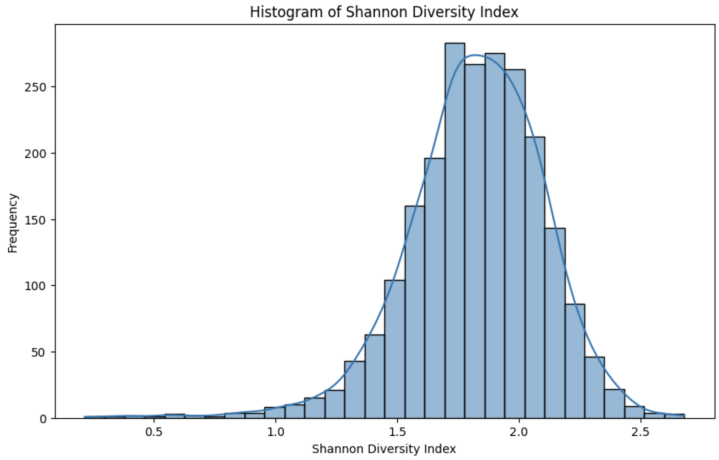



*Figure 4.4*: Histogram plot for Shannon Diversity Index.

Fig. 4.4 supports this insight, as the histogram shows a right-skewed distribution and a long tail extending toward lower diversity values, suggesting that there are notable outliers, particularly in the cancer group.

To check the significance of the differences in SDI across the health statuses, the Kruskal-Wallis test was conducted, which showed that the results were statistically significant ($H = 30.11$, $p < 0.00001$), which means that the differences in microbiome diversity between the groups are unlikely to be due to chance. The very low $p$-value of $4.65 \times 10^{-6}$ reinforces the finding. This confirmed that microbiome diversity had differences in the different health statuses.

However, the effect size measured by Eta squared came out to be 0.0116, which was small, implying that while the differences are statistically significant, their actual impact on microbiome diversity is modest, suggesting that other factors like diet, medication, and genetic predispositions may play substantial roles in determining microbiome diversity.

Overall the alpha diversity analysis revealed that while there are statistically significant dif-

ferences in microbiome diversity across various health groups., the magnitude of these differences is relatively small, which underscores the importance of considering additional variables related to diet, medication, or other lifestyle habits.

### 4.3.2 Beta Diversity Results

For the purpose of beta diversity analysis, Bray-Curtis dissimilarity and Principal Coordinates Analysis (PCoA) were employed which provided key insights into how the microbial community composition changes are associated with various health statuses. It highlights how shifts in the microbiome can result in different health conditions like colorectal cancer. The results of Bray-Curtis dissimilarity were used to construct the PCoA plot, to increase the interpretability of the result.

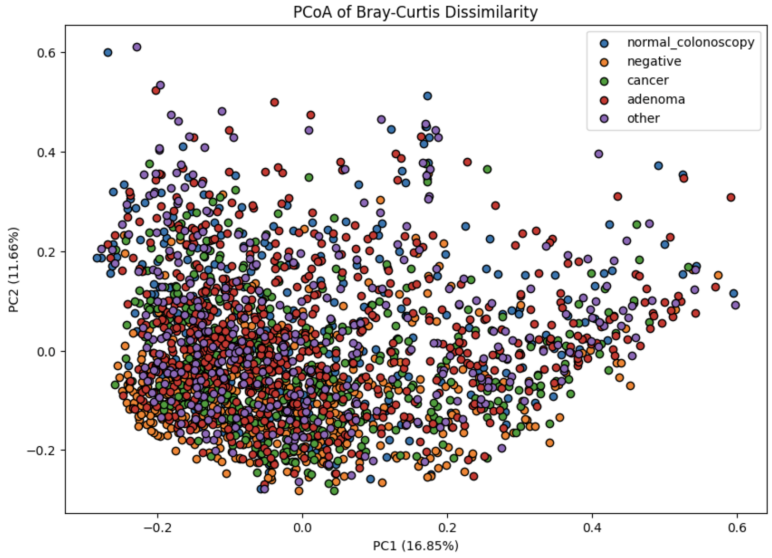



*Figure 4.5*: PCoA plot for Bray-Curtis dissimilarity.

Fig 4.5 illustrates these results, with the two principal coordinates (PC1 and PC2) explaining 28.51% of the variance (16.85% and 11.66%, respectively). The plot indicates that these axes are able to capture a substantial amount of the variance in the microbiome composition. It also reveals some overlapping among the groups, particularly between blood-negative, cancer, and adenoma which suggests that although the microbial compositions differ, there are shared characteristics across these health statuses. It also revealed that the wider spread of points for adenoma and cancer groups indicates greater variability, potentially reflecting a higher degree of dysbiosis or microbial imbalance, which is often associated with health statuses like colorectal cancer and adenoma.

| Comparison | Test Statistic (pseudo-F) | p-value |
|---|---|---|
| Blood-Negative vs Cancer | 16.947935 | 0.001 |
| Blood-Negative vs Adenoma | 39.366034 | 0.001 |
| Cancer vs Adenoma | 6.500141 | 0.001 |

*Table 4.3*: PERMANOVA results for different comparisons.

PERMANOVA results, from Table 4.3, further confirmed the differences in microbial composition across different groups, with an overall pseudo-F statistic of 15.84 and a p-value of 0.001. This low p-value proved the statistical significance of the observed results. Further, pairwise PERMANOVA comparisons elucidated these differences. The comparison between the blood-negative and cancer groups yielded a pseudo-F value of 16.95 with a p-value of 0.001, indicating distinct microbial community compositions between these groups. The blood-negative vs. adenoma comparison showed an even higher pseudo-F value of 39.37, suggesting substantial differences between these groups, likely reflecting a distinct microbial environment associated with adenoma formation. The cancer vs. adenoma comparison, with a pseudo-F value of 6.50, revealed that while there are statistically significant differences, they are less pronounced, indicating some shared microbial characteristics between these groups.

### 4.3.3 Differential Abundance Analysis Results

#### Cancer vs Blood-Negative

The differential abundance analysis of the 'cancer' and 'blood-negative' groups revealed significant variations in the abundance of many microbial families, revealing significant modifications in the microbiome associated with colorectal cancer. The Mann-Whitney U test, with Bonferroni correction, identified top 10 microbial families with most statistically significant differences.

| Feature | Statistic | adjusted p-value |
|---|---|---|
| Fusobacteriaceae | 144312.0 | 4.863085e-26 |
| Bacillales (Family.X) | 131599.5 | 3.749689e-24 |
| Bacillales (Family.XI) | 131599.5 | 3.749689e-24 |
| Ruminococcaceae | 63601.5 | 9.567024e-24 |
| Carnobacteriaceae | 118004.0 | 1.525834e-06 |
| Peptococcaceae | 87198.0 | 1.235398e-04 |
| Christensenellaceae | 87195.0 | 2.394251e-04 |
| Clostridiales.vadinBB60.group | 88851.5 | 1.492233e-03 |
| Corynebacteriaceae | 113377.0 | 2.775109e-03 |
| Verrucomicrobiaceae | 90470.0 | 7.982410e-03 |

*Table 4.4*: Results of Mann-Whitney U test for Cancer vs Blood-Negative.

From Table 4.4, it can be observed that *Fusobacteriaceae* exhibited the largest significant difference, with a test statistic of 144,312 and an adjusted p-value of $4.86 \times 10^{-26}$, indicating a

strong link to colorectal cancer via inflammation promotion and possible carcinogenesis. *Bacillales* Families X and XI also showed significant changes, each with a test statistic of 131,599.5 and adjusted p-values of $3.75 \times 10^{-24}$, implying that their altered abundance may disturb normal gut microbial metabolism, thereby contributing to cancer progression. The differential abundance of *Carnobacteriaceae*, marked by an adjusted p-value of $1.53 \times 10^{-6}$, further underscores the microbiome changes associated with colorectal cancer.

*Ruminococcaceae*, known to be involved in short-chain fatty acid production, which has anti-inflammatory properties; achieved a test statistic of 63,601.5 with an adjusted p-value of $9.57 \times 10^{-24}$. These changes in its abundance may reduce the protective effects, facilitating a more carcinogenic environment. Additionally, significant results were found for *Peptococcaceae* (adjusted p-value = $1.24 \times 10^{-4}$) and *Christensenellaceae* (adjusted p-value = $2.39 \times 10^{-4}$), which are crucial for maintaining mucosal integrity and modulating inflammation. The shift in their abundance could mean a disruption in these protective mechanisms. *Corynebacteriaceae* is associated with lipid metabolism, indicating metabolic alterations associated with cancer, with a test statistic of 113,377.0 and an adjusted p-value of $2.78 \times 10^{-03}$. Changes in the *Verrucomicrobiaceae* family may result in a decrease in mucosal protection and an increase in inflammation.

### Cancer vs Adenoma

In the comparison between 'cancer' and 'adenoma' groups, significant microbial shifts were also observed, indicating distinct microbiome changes as the disease progresses from precancerous lesions to malignant tumors.

| Feature | Statistic | Adjusted p-value |
|---|---|---|
| Fusobacteriaceae | 181,756.5 | $1.28 \times 10^{-15}$ |
| Clostridiales (Family XI) | 180,865.0 | $5.68 \times 10^{-12}$ |
| Bacillales (Family XI) | 167,667.0 | $2.03 \times 10^{-11}$ |
| Clostridiales (Family XIII) | 179,953.5 | $2.10 \times 10^{-11}$ |
| Bacillales (Family X) | 167,510.0 | $3.18 \times 10^{-11}$ |
| Rikenellaceae | 175,003.0 | $1.73 \times 10^{-08}$ |
| Clostridiales.vadinBB60.group | 171,267.5 | $9.59 \times 10^{-07}$ |
| Mollicutes.RF9 (uncultured.bacterium) | 169,146.5 | $2.55 \times 10^{-06}$ |
| Verrucomicrobiaceae | 167,522.0 | $5.03 \times 10^{-05}$ |
| Christensenellaceae | 167,750.5 | $5.86 \times 10^{-05}$ |

*Table 4.5*: Results of Mann-Whitney U test for Cancer vs Adenoma.

From Table 4.5 it can again be observed that *Fusobacteriaceae* continued to show a marked difference, with a test statistic of 181,756.5 and an adjusted p-value of $1.28 \times 10^{-15}$, reinforcing its role in cancer progression and its potential as a biomarker for distinguishing cancer from adenoma. Along with this, other differences were also observed in *Clostridia* families, including Family XI, XIII, and *vadinBB60* group, suggesting that alterations in microbial fermentation

could influence colonic health and increase the cancer risk.

Other microbes like *Rikenellaceae* (adjusted p-value = $1.73 \times 10^{-8}$), *Mollicutes RF9* group (adjusted p-value = $2.55 \times 10^{-6}$), and *Verrucomicrobiaceae* (adjusted p-value = $5.03 \times 10^{-5}$) also pointed towards microbial shifts that could impair gut barrier functions, promoting CRC progression through increased inflammation.

Differential abundance analysis results showed different microbial patterns in the 'cancer', 'adenoma', and 'blood-negative' health groups. These findings imply that some microbial families may function as biomarkers for the progression of colorectal cancer, facilitating early detection and focused treatment approaches.

## 4.4 Model Performance

The Model Performance section unveils a detailed evaluation of the predictive power of various machine learning models trained on two critical cases: CRC vs Blood-Negative and CRC vs Adenoma. Each model's performance is rigorously assessed, shedding light on their effectiveness and limitations in distinguishing between colorectal cancer and other health conditions. For the development of various machine learning models for colorectal cancer (CRC) prediction, an 80-20 train-test split was consistently employed across all models and cases. This split ensures a robust assessment of each model's ability to generalize and accurately predict CRC status, whether distinguishing between CRC and Blood-Negative or CRC and Adenoma cases.

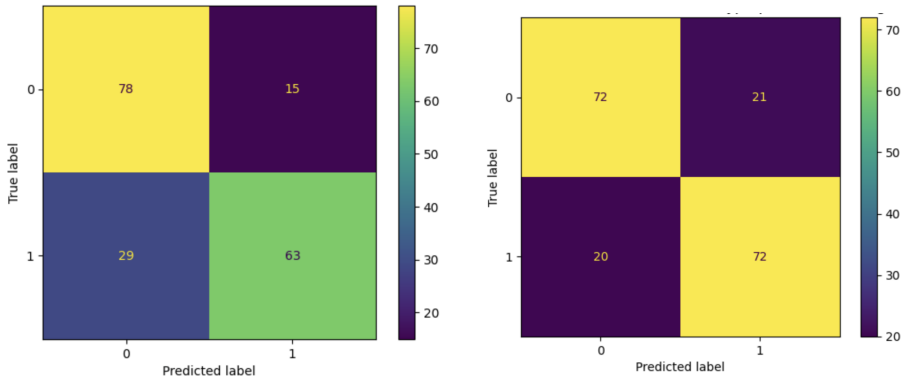
### 4.4.1 Cancer vs Negative

**Random Forest**

| Model | AUC-ROC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| **Random Forest** | 0.87 | 0.76 | 0.80 | 0.68 | 0.74 |
| **Tuned Random Forest** | 0.87 | 0.78 | 0.77 | 0.78 | 0.78 |

*Table 4.6*: Performance Metrics for Random Forest and Tuned Random Forest Models (CRC vs Blood-Negative)

The Random Forest model was initially trained with 400 estimators, exhibited a balanced performance in distinguishing CRC from Blood-Negative cases, as evidenced by Fig. 4.6 with an accuracy of 76% and an F1-score of 0.74.




(a) Base Random Forest (b) Tuned Random Forest

*Figure 4.6*: Figures (a) and (b) represents the confusion matrices of Base RF and Tuned RF

The confusion matrix in Fig. 4.6a showed the model's ability to correctly identify 78 out of 93 blood-negative cases and 63 out of 92 CRC cases, though 29 CRC cases were incorrectly classified as negative. This resulted in a slightly higher recall for blood-negative (84%) compared to CRC (68%), reflecting a moderate imbalance in sensitivity between the two groups. The AUC-ROC score of 0.87 suggests a strong overall model performance, indicating that the Random Forest can effectively differentiate between these health statuses with a reasonably high true positive rate across various thresholds.

After hyperparameter tuning using Optuna, the optimized Random Forest model, with 281 estimators and a maximum depth of 8, demonstrated improved performance metrics. The accuracy increased to 78%, and the F1-score rose to 0.78 (from Fig. 4.6), reflecting better-balanced sensitivity between the blood-negative and CRC groups. The tuned model's confusion matrix (4.6b) showed 72 correct classifications for both groups, highlighting enhanced precision and recall rates of 78%, effectively reducing errors.

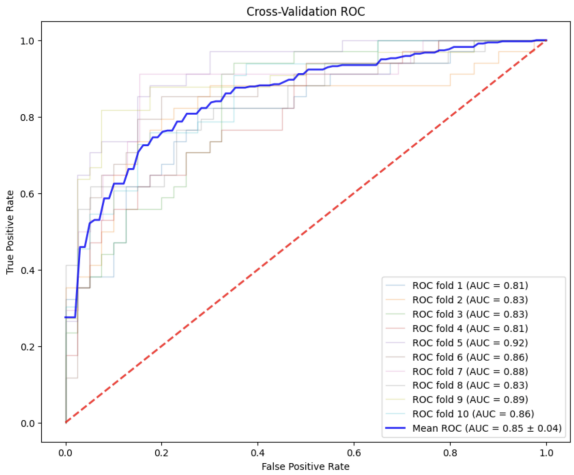



*Figure 4.7*: Mean AUC-ROC plot after Cross-Validation for Tuned Random Forest Model

Cross-validation further validated the model's robustness, yielding a mean AUC of $0.85 \pm 0.04$

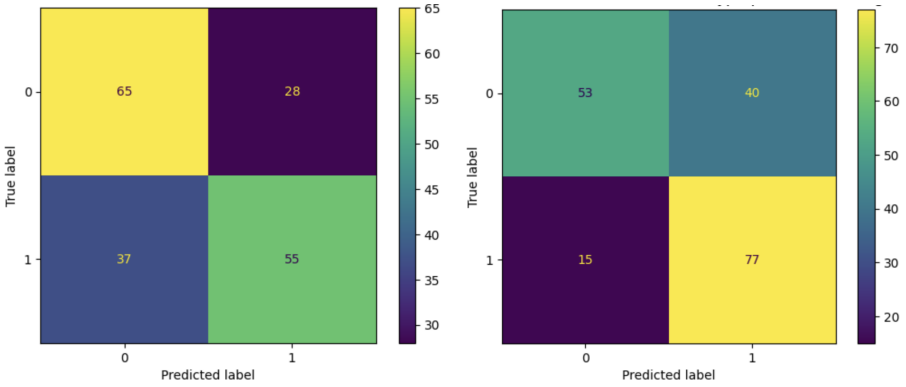
(observed from Fig. 4.7), confirming consistent performance across different folds. The feature importance analysis identified several critical microbiome families, including *Fusobacteriaceae, Ruminococcaceae*, Family X and XI within the *Bacillales*, *Erysipelotrichaceae*, and *Porphyromonadaceae*, as significant contributors to CRC prediction, indicating that the Random Forest model effectively handles complex interactions within high-dimensional microbiome data, making it particularly effective for CRC prediction.

### Logistic Regression with L2 Regularisation

| Model | AUC-ROC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| **Logistic Regression (L2)** | 0.70 | 0.65 | 0.66 | 0.60 | 0.63 |
| **Tuned Logistic Regression (L2)** | 0.73 | 0.70 | 0.66 | 0.83 | 0.74 |

*Table 4.7*: Performance Metrics for base Logisitic Regression (L2) and Tuned Logistic Regression (L2) (CRC vs Blood-Negative)

The Logistic Regression model with L2 regularization initially achieved an accuracy of 65%, a precision of 66% for CRC cases (label 1), and a recall of 60% (Table 4.7). The F1-score for the model is 0.63.




(a) Base Logistic Regression (b) Tuned Logistic Regression

*Figure 4.8*: Figures (a) and (b) represents the confusion matrices of Base Logisitic Regression and Tuned Logistic Regression

The confusion matrix (Fig. 4.8a) reflects these metrics with a notable number of false positives (28) and false negatives (37), indicating challenges in correctly identifying CRC cases. The AUC-ROC curve, with an area of 0.70 (Table 4.7), shows a moderate ability to distinguish between CRC and blood-negative cases.

After tuning the model with ($C = 42.08$), the model's performance improved across several metrics. The tuned model reached an accuracy of 70%, with a precision of 66% for CRC detection and a recall increase to 84% (Table 4.7), indicating a substantial reduction in false negatives (now 15, from Fig. 4.8b). The F1-score improved to 0.74, suggesting a better balance

between precision and recall. The AUC-ROC curve also improved to 0.73, indicating a stronger ability to differentiate between CRC and non-CRC cases after tuning.

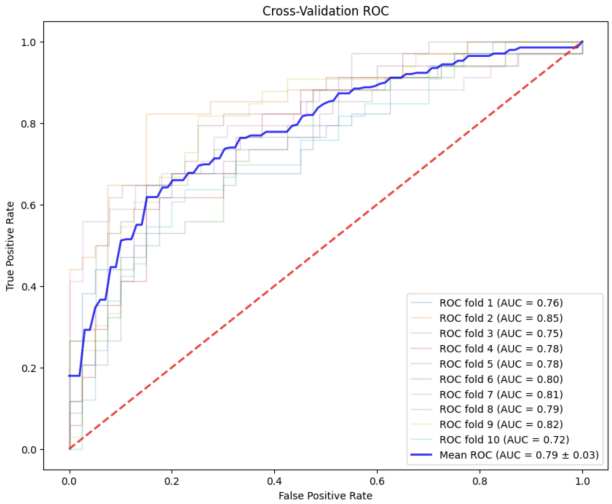



*Figure 4.9*: Mean AUC-ROC plot after Cross-Validation for Tuned Logistic Regression Model

Cross-validation further assessed the robustness of the tuned logistic regression model. Using 10-fold cross-validation, the model achieved a mean AUC of 0.79±0.03 (from Fig. 4.9). This consistency across folds indicates the model's improved generalizability and reduced overfitting, underscoring the model's stability when applied to different subsets of the data.

Feature importance analysis reveals several key predictors, including microbial families like *Ruminococcaceae, Porphyromonadaceae, Peptostreptococcaceae, Erysipelotrichaceae*, and *Rikenellaceae*. Additionally, *Rhodospirillaceae* and *Veillonellaceae* were identified as significant features, suggesting distinct microbial compositions in CRC patients that are critical for model prediction.

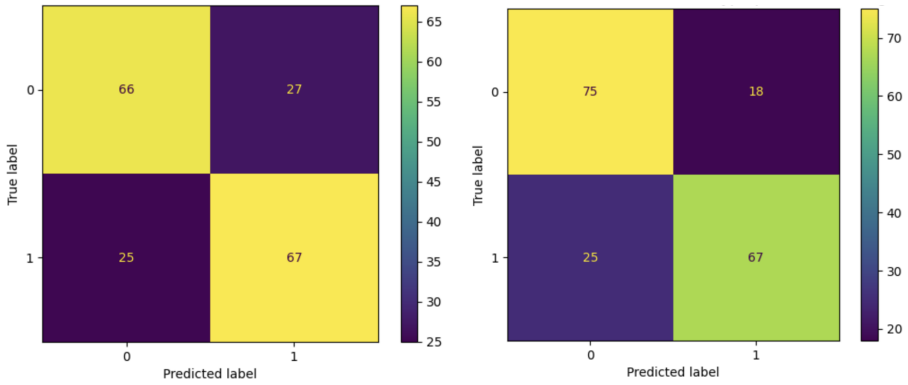
### LightGBM

| Model | AUC-ROC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| **LightGBM** | 0.82 | 0.72 | 0.71 | 0.73 | 0.72 |
| **Tuned LightGBM** | 0.82 | 0.77 | 0.79 | 0.73 | 0.76 |

*Table 4.8*: Performance Metrics for base LightGBM and Tuned LightGBM (CRC vs Blood-Negative)

The LightGBM model, showed moderate performance both before and after hyperparameter tuning. From Table 4.8 and Fig. 4.10a it can be observed that, initially, the model achieved an accuracy of 72% on the test data, with the confusion matrix showing 66 true negatives, 27 false positives, 25 false negatives, and 67 true positives. The precision and recall for blood-negative samples (class 0) were 73% and 71%, respectively, while for CRC (class 1), both metrics were

reversed, demonstrating balanced detection capabilities across both classes. The F1 score for the base model was 0.72, and the AUC-ROC was 0.82, reflecting good discriminatory power between the two groups.




(a) Base LightGBM (b) Tuned LightGBM

*Figure 4.10*: Figures (a) and (b) represents the confusion matrices of Base LightGBM and Tuned LightGBM

The hyperparameter tuning, optimized parameters such as `learning_rate`, `num_leaves`, `max_depth`, `min_data_in_leaf` and the regularization parameters `lambda_l1` (1.36e-05) and `lambda_l2` (0.0023). The accuracy increased to 77% (Table 4.8), and the confusion matrix showed 75 true negatives, 18 false positives, 25 false negatives, and 67 true positives (Fig. 4.10b). The precision for blood-negative samples improved to 75%, while for CRC, it increased to 79%. The recall for blood-negative samples rose to 81% and for CRC to 73%. The F1 score also saw an enhancement to 0.76, with the AUC-ROC remaining consistent at 0.82, indicating reliable model performance post-tuning.

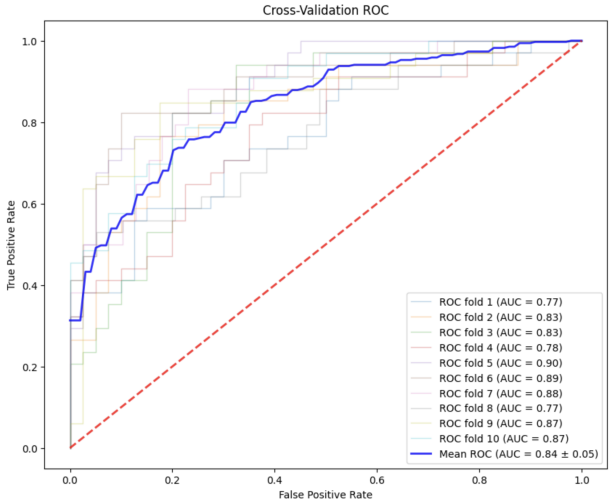



*Figure 4.11*: Mean AUC-ROC plot after Cross-Validation for Tuned LightGBM Model

Cross-validation further supported these results, producing a mean AUC-ROC of 0.84±0.05,

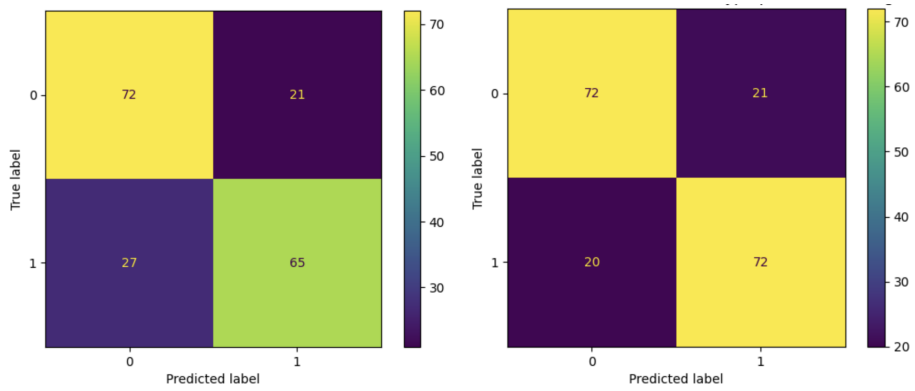
suggesting consistent predictive ability across various subsets of the data. The feature importance analysis further identified significant microbes, including *Ruminococcaceae, Acidaminococcaceae, Erysipelotrichaceae, Desulfovibrionaceae,* and *Porphyromonadaceae*, underscoring their potential roles in CRC prediction. These families' prominence in the model aligns with their established associations with colorectal cancer, reinforcing their relevance in this predictive context.

**Ensemble Model**

| Model | AUC-ROC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| **Ensemble Model** | 0.85 | 0.74 | 0.75 | 0.70 | 0.73 |
| **Tuned Ensemble Model** | 0.86 | 0.78 | 0.77 | 0.78 | 0.78 |

*Table 4.9*: Performance Metrics for base Ensemble Model and Tuned Ensemble Model (CRC vs Blood-Negative)

The ensemble model, comprising tuned Random Forest and LightGBM as base models and Logistic Regression as the meta-model, demonstrated significant performance in distinguishing between CRC and blood-negative samples. The initial ensemble model (as observed from Table 4.9) exhibited a balanced performance with an accuracy of 74%, precision of 73% for the blood-negative class, and 76% for the CRC class, with an F1 score of 0.73. The AUC-ROC score of 0.85 further validates its capability to distinguish effectively between classes.




(a) Base Ensemble Model (b) Tuned Ensemble Model

*Figure 4.12*: Figures (a) and (b) represents the confusion matrices of Base Ensemble model and Tuned Ensemble model

The confusion matrix (Fig. 4.12a) reveals that the ensemble model correctly identified 72 instances of blood-negative samples and 65 instances of CRC out of 185 samples, though it misclassified 21 blood-negative and 27 CRC samples, indicating areas for improvement.

After performing hyperparameter tuning, the meta-model parameters were optimized to `C`: 1.005, `solver:saga`, `penalty:l1`, and `max_iter`: 571. These optimised parameters enhanced the model's performance, achieving an accuracy of 78%, precision of 77%, and an F1

score of 0.78. The AUC-ROC also increased to 0.86, reflecting improved discrimination ability. The confusion matrix post-tuning indicated better balance (Fig. 4.12b), with 72 correctly classified CRC samples and blood-negative samples both, decreasing the overall misclassifications.

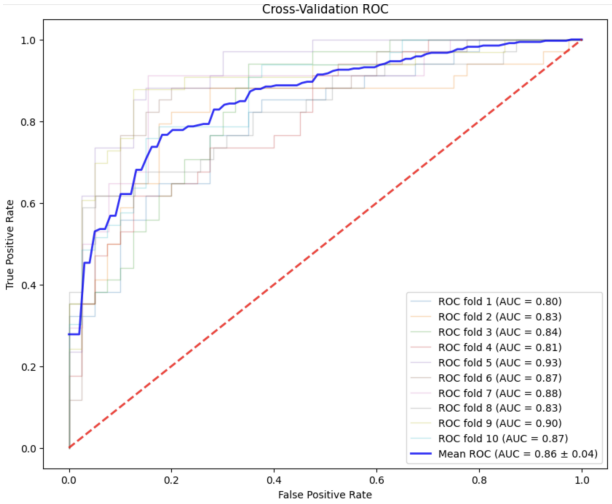



*Figure 4.13*: Mean AUC-ROC plot after Cross-Validation for Tuned Ensemble Model

Cross-validation of the tuned ensemble model produced a mean AUC-ROC of 0.86±0.04, highlighting the model's robustness and consistent performance across different folds. The ensemble's strong cross-validation results emphasize its reliability and potential utility in clinical settings for early CRC detection.

The top features that were contributing to the ensemble model's predictive power, with the most important being, *Fusobacteriaceae, Ruminococcaceae, Clostridiales.vadinBB60.group, Bacillales*' Family X and XI, *Lachnospiraceae,* and *Streptococcaceae*. These features (microbial families), suggest that shifts in specific microbial communities play a critical role in differentiating CRC from non-CRC cases.

### 4.4.2 Cancer vs Adenoma

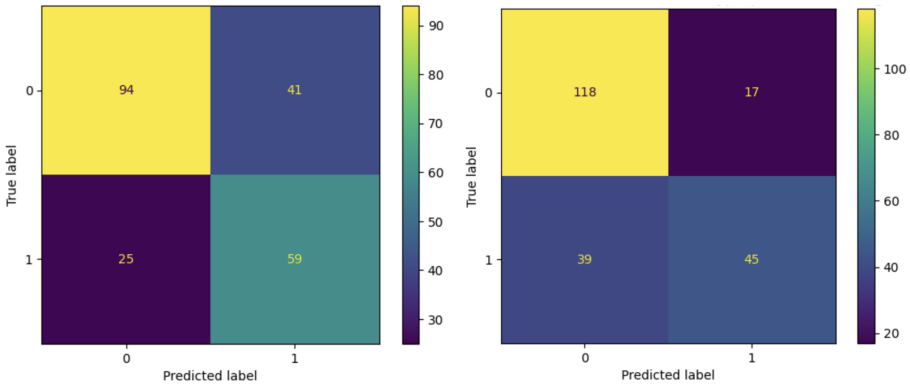
#### Random Forest

| Model | AUC-ROC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| **Random Forest** | 0.76 | 0.70 | 0.59 | 0.70 | 0.64 |
| **Tuned Random Forest** | 0.77 | 0.74 | 0.73 | 0.54 | 0.62 |

*Table 4.10*: Comparison of Performance Metrics for Random Forest and Tuned Random Forest (Cancer vs Adenoma)

The Random Forest model, initialized with 400 estimators, was used to classify colorectal cancer (CRC) versus adenoma. The base model achieved an overall accuracy of 70% and an F1-score of 0.64, indicating moderate differentiation between the two classes (Table 4.10). The

model's precision for adenoma (class 0) was 79%, with a recall of 70%, while for CRC (class 1), the precision was 59%, and recall was 70%.




(a) Base Random Forest Model (b) Tuned Random Forest Model

*Figure 4.14*: Figures (a) and (b) represents the confusion matrices of Base RF Model and Tuned RF Model

The confusion matrix showed that 41 adenoma samples were incorrectly classified as CRC, and 25 CRC cases were missed (Fig. 4.14a). The ROC AUC score of 0.76 (Table 4.10) demonstrates a fair capability of the model in distinguishing between CRC and adenoma based on microbial profiles.

Post hyperparameter tuning, the Random Forest model was optimized with parameters such as 567 `n_estimators`, `maximum_depth` of 2, and `min_samples_split` of 34. The tuned model displayed an improved overall accuracy of 74% and an F1-score of 0.62 (Table 4.10). The confusion matrix for the tuned model shows reduced false negatives for adenoma but an increase in missed CRC cases, with 39 samples misclassified (Fig. 4.14b), suggesting some trade-off in class prediction. The ROC AUC score improved slightly to 0.77 (Table 4.10).

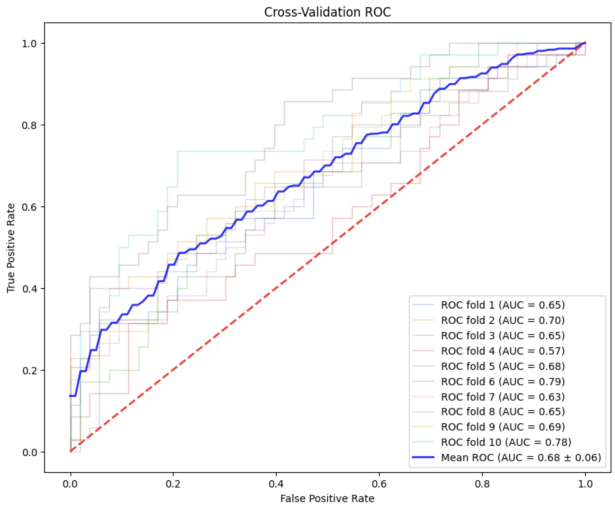



*Figure 4.15*: Mean AUC-ROC plot after Cross-Validation for Tuned RF Model

Cross-validation of the tuned model revealed a mean ROC AUC of 0.68±0.06 (Fig. 4.15), highlighting variability in the model's performance across different data subsets, suggesting the model's sensitivity to underlying data patterns.

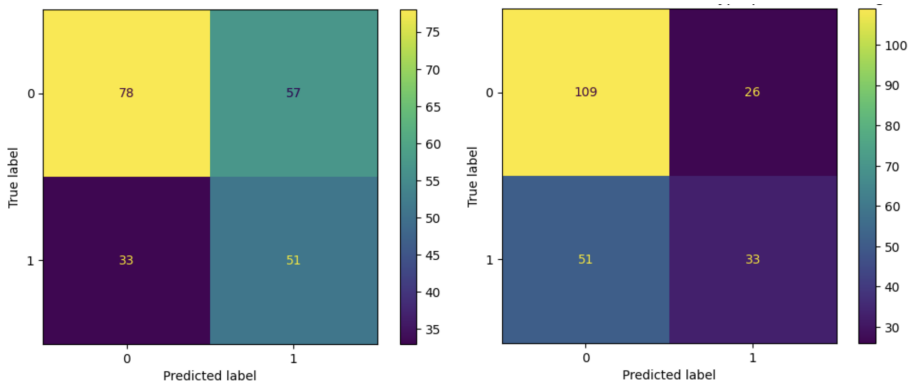
Feature importance analysis highlighted key microbial taxa as significant contributors to the model's predictions. Notably, *Fusobacteriaceae, Rikenellaceae, Clostridiales* Family XI & XIII, and *Ruminococcaceae* were among the top features, emphasizing their potential roles in distinguishing CRC from adenoma, and aligning with existing literature, linking these microbiomes to CRC.

### Logistic Regression with L2 Regularisation

| Model | AUC-ROC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| **Logistic Regression (L2)** | 0.61 | 0.59 | 0.47 | 0.61 | 0.53 |
| **Tuned Logistic Regression (L2)** | 0.62 | 0.65 | 0.56 | 0.40 | 0.46 |

*Table 4.11*: Comparison of Performance Metrics for Logistic Regression with L2 and Tuned Logistic Regression with L2 (Cancer vs Adenoma)

The logistic regression model with L2 regularization demonstrated moderate performance. The initial model achieved an accuracy of 59%, with an overall F1 score of 0.53, indicating moderate ability to differentiate between cancer and adenoma (Table 4.11).




(a) Base Logistic Regression (b) Tuned Logistic Regression

*Figure 4.16*: Figures (a) and (b) represents the confusion matrices of Base Logisitic Regression and Tuned Logistic Regression

The confusion matrix in Fig. 4.16a reveals that the model correctly identified 78 adenoma cases (true negatives) and 51 cancer cases (true positives), while misclassifying 33 cancer cases and 57 adenoma cases. The ROC-AUC score of 0.61 suggests limited discrimination capability, implying the model's performance is not robust for this dataset.

Upon hyperparameter tuning, with the best parameter for $C$ was approx. 5.42, the model showed a slight improvement in accuracy to 65%, but the overall F1 score decreased to 0.46

(Table 4.11). The confusion matrix (Fig. 4.16b) for the tuned model shows better precision for adenoma (68%) but lower recall for CRC (39%), reflecting a conservative bias, potentially favoring adenoma classification due to class imbalance. The ROC-AUC post-tuning was slightly improved to 0.62 (Table 4.11).

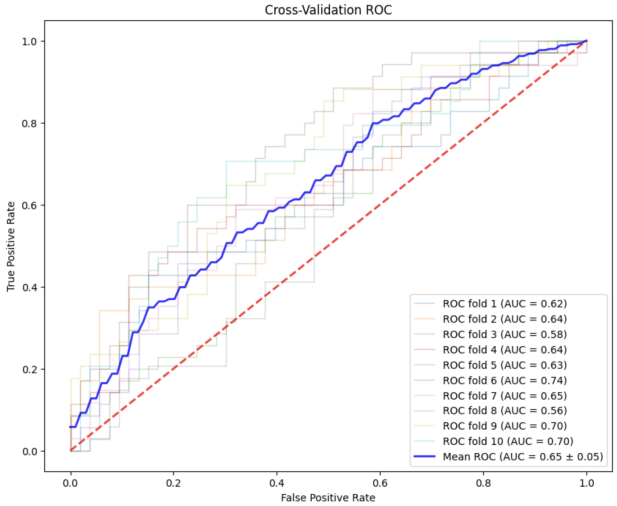



*Figure 4.17*: Mean AUC-ROC plot after Cross-Validation for Tuned Logistic Regression (L2) Model

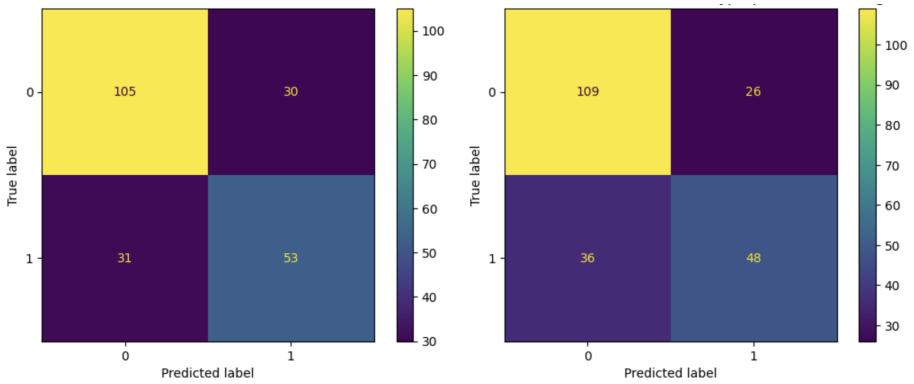
The cross-validation results further solidify the model's modest predictive power, with variability in ROC-AUC scores across different folds, averaging around 0.65±0.05 (Fig. 4.17), highlighting the model's sensitivity to data partitioning, which is affecting generalisability. Feature importance analysis shows that features such as *Enterobacteriaceae, Lachnospiraceae, Bacteroidaceae, Rikenellaceae,* and *Prevotellaceae* were among the most influential in determining model's predictions.

### LightGBM

| Model | AUC-ROC | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| **LightGBM** | 0.75 | 0.72 | 0.64 | 0.63 | 0.63 |
| **Tuned LightGBM** | 0.77 | 0.72 | 0.65 | 0.60 | 0.61 |

*Table 4.12*: Performance metrics for LightGBM and Tuned LightGBM models (Cancer vs Adenoma)

The LightGBM model initially demonstrated modest performance in distinguishing between adenoma and colorectal cancer (CRC) cases, achieving an accuracy of 72%, an AUC-ROC of 0.75, and an overall F1 score of 0.63 (Table 4.12). The confusion matrix (Fig. 4.18a) shows that the model correctly identified 105 adenoma cases (true negatives) and 53 CRC cases (true positives). However, it misclassified 30 adenoma (false positives) and 31 CRC cases (false negatives), highlighting a nearly balanced sensitivity and specificity in detecting both classes.




(a) Base LightGBM (b) Tuned LightGBM

*Figure 4.18*: Figures (a) and (b) represents the confusion matrices of Base LightGBM and Tuned LightGBM

Hyperparameter tuning showed slight changes in the LightGBM model's performance. The tuned model maintained an accuracy of 72% and an improved AUC-ROC of 0.77 but reported a slightly lower overall F1 score of 0.61 (Table 4.12). The confusion matrix (Fig. 4.18b) for the tuned model reveals 109 true negatives, 48 true positives, 26 false positives, and 36 false negatives. The parameter tuning primarily enhanced the model's ability to correctly classify adenoma cases, while the decline in the F1 score suggests a trade-off in sensitivity towards detecting CRC cases.

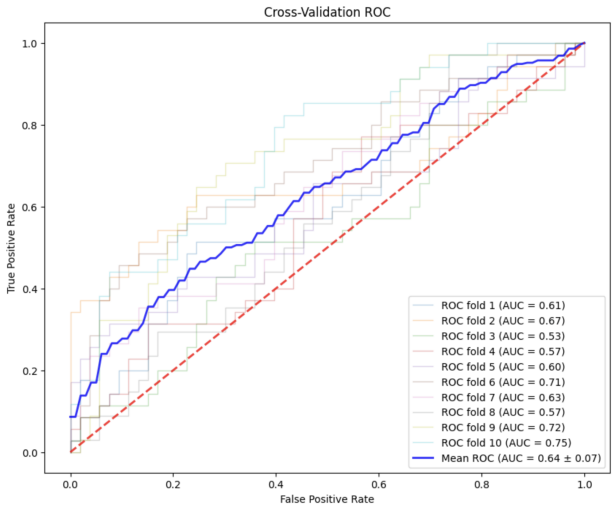



*Figure 4.19*: Mean AUC-ROC plot after Cross-Validation for Tuned LightGBM Model

Cross-validation results for the tuned model showed consistent performance, with an average AUC-ROC of 0.64±0.07 across folds (Fig. 4.19), indicating stable performance across different subsets of the data. Regarding feature importance, the most influential features included *Peptostreptococcaceae, Rikenellaceae, Clostridiaceae.1, Veillonellaceae,* and *Bacteroidaceae*, with *Fusobacterium* also being highlighted due to its known association with CRC.

**Ensemble Model**

*Table 4.13*: Performance Metrics of the Ensemble Model

| Model | AUC-ROC | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Ensemble Model | 0.78 | 0.74 | 0.68 | 0.64 | 0.66 |

The ensemble model achieved an overall accuracy of 74%. Precision was recorded at 78% for adenoma (class 0) and 68% for CRC (class 1), while recall was 81% for adenoma and 64% for CRC. This balance resulted in an overall F1 score of 0.66, and an AUC-ROC of 0.78 (Table 4.13), suggesting good discriminative ability between the two classes.

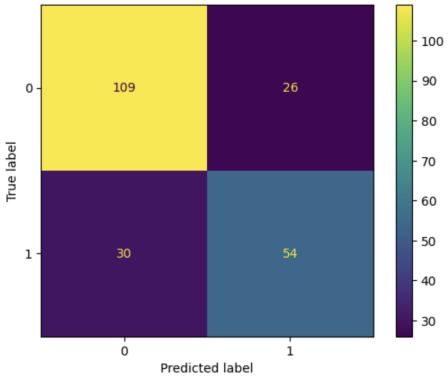



*Figure 4.20*: Confusion matrics for the Ensemble Model

The confusion matrix (Fig. 4.20) for the ensemble model shows that it correctly identified 109 adenoma cases (true negatives) and 54 CRC cases (true positives). However, there were also 30 CRC cases and 26 adenoma cases misclassified.

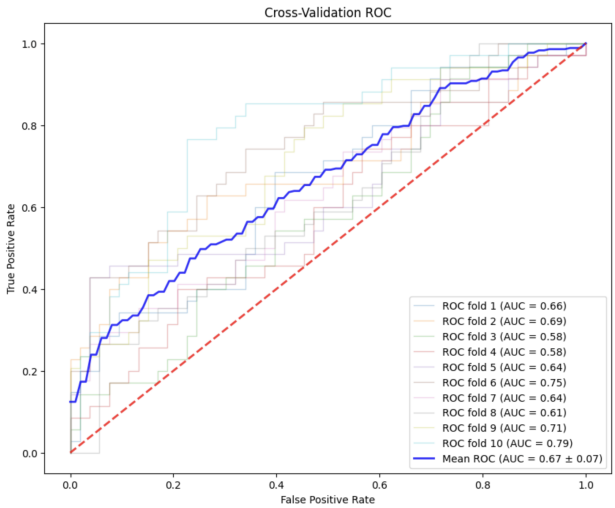



*Figure 4.21*: Mean AUC-ROC plot after Cross-Validation for Ensemble Model

The cross-validation AUC-ROC for the ensemble model averaged 0.67±0.07 (Fig. 4.21),

which confirms the model's robustness across different data subsets. This consistency is crucial for ensuring the model's applicability in various clinical settings and patient populations.

The feature importance analysis for the ensemble model reveals several critical microbial taxa associated with distinguishing between CRC and adenoma. The most significant features included: *Fusobacteriaceae, Peptostreptococcaceae, Pasteurellaceae, Ruminococcaceae, Bacteroidales.S24.7.group, Lachnospiraceae,* and *Bifidobacteriaceae*. These features provide insight into the complex microbial landscape associated with colorectal cancer, which is crucial for both diagnostic and therapeutic strategies.

# Chapter 5

# Discussion

## 5.1 Model Performance in CRC vs Blood-Negative:

In this classification, the models were evaluated using various matrices, particularly the mean AUC-ROC from cross-validation. The ensemble model, with the strengths of Random Forest and LightGBM with a logistic regression meta-model, outperformed other models with the highest mean AUC-ROC of 0.86. Using Random Forest's ability to capture complex, non-linear relationships and LightGBM's ability to handle high-dimensional data and class imbalances enabled the ensemble model to achieve greater accuracy and generalisability. The Random Forest model also performed well, with a mean AUC-ROC of 0.85, slightly high than LightGBM's 0.84. However, LightGBM demonstrated highest precision (79%) almost all the models, which suggests its robust ability to classify true positives among CRC samples correctly. On the other hand, ensemble model and Random Forest had the highest recall of 78%, strong performance in identifying all positive cases.

## 5.2 Model Performance in CRC vs Adenoma

The task of distinguishing CRC form adenoma was challenging due to the bilogical similarities and overlapping microbial profiles. For this classificication, Random Forest model outperformed other models with the highest mean AUC-ROC of 0.68. The model also exhibited the highest precision (73%), highlighting the ability to identify adenoma samples. Whereas, the ensemble model achieved the highest recall of 64%, indicating better performance in capturing all positive cases.

The logistic regression performed poorly, especially in CRC vs adenoma, which could be due to the violation of its underlying assumptions. Logistic regression assumes linearity and independence of features, which may not hold in the context of microbiome data. The class imbalance and overlapping microbial features between CRC and adenoma further might be the reason for the logistic regression's low performance. The choice to keep the base logistic regression without tuning was made because of the overlap and imbalance that were obvious in

the ensemble model's decreased efficacy of hyperparameter-tuned logistic regression.

## 5.3 Key Microbial Features in CRC Detection

The prominent microbial features identified across these models, such as *Fusobacteriaceae, Ruminococcaceae, Veillonellaceae*, *Bacillales' Family X and XI*, and *Rikenellaceae*, are consistent with the literature. *Fusobacteriaceae*, in particular, is frequently linked to CRC due to its inflammatory and pro-carcinogenic potential, emphasizing its importance as a biomarker in distinguishing cancerous conditions from non-cancerous states. Although, *Bacillales Family X and XI* emerged as correlated yet significant features across both machine learning models and differential abundance analyses, suggesting their shared role in CRC pathogenesis which is also supported by the existing literature. Despite their correlation, their consistent presence highlights the potential importance of these microbial families in influencing CRC development and progression, warranting further investigation. The presence of these microbial families across models aligns with the diversity analysis findings, which indicated distinct microbial patterns in CRC samples compared to blood-negative cases, reinforcing their validity as biomarkers for CRC detection.

## 5.4 Validation with Diversity Analysis and Existing Research

The results of diversity analysis validate the microbial features identified in the machine learning models. Alpha diversity analysis revealed relatively similar Shannon Diversity Index values across all the groups, with CRC having the highest mean and median SDI. The higher diversity in CRC samples perfectly aligns with the existing literature that suggest a complex microbial ecosystem in CRC patients, which is probably due to dysbiosis and the presence of microbes like *Fusobacterium nucleatum*.

In Beta diversity analysis, PERMANOVA further validated these findings by highlighting statistically significant differences in the microbial compositions between different health statuses. The distinct microbial composition identified in differential abundance analysis, specifically the prominence of *Fusobacteriaceae* and *Ruminococcaceae*, align with these diversity findings and existing literature, which associates these taxa with inflammation and cancer progression.

Microbial families like *Bacillales' Family X and XI* and *Rikenellaceae* were identified as important features in the models, are also supported by literature and diversity analysis. While *Rikenellaceae* has been associated with inflammation and changes in gut barrier function, *Bacillales' Families X and XI* have been connected to metabolic pathways that could influence the immune system and accelerate the development of cancer. The difference in abundance between these families in CRC and adenoma samples highlights their potential relevance in colorectal cancer diagnosis and distinction.

## 5.5 Implications for Healthcare and Future Research

The results of this research successfully highlight the potential of using microbiome profiles to develop machine learning models for non-invasive CRC screening and differentiating from other health statuses. By utilisation of rich microbiome diversity present in the gut, these machine learning models offer a less invasive alternative to conventional screening methods such as colonoscopy which could cause discomfort and is expensive. The research also identified specific microbial signatures which are associated with colorectal cancer which not only make early detection easier, but also create opportunities for tailored treatment plans based on a person's microbiome profile. This could lead to enhancement of early detection rates, improving patient outcomes and survival rates.

Nevertheless, there are few factors which need careful consideration to translate these results into clinical practice. It is important to develop standardisation in the collection, processing and analysis of microbiome samples to make sure the reproducibility and reliability of the results across diverse population of patients and healthcare environment. The gut microbiome can be significantly impacted by changes in diet/eating habits, medicine, genetic background and other lifestyle habits. Due to this reason, thorough guidelines must be devised to lower the impact of these factors in clinical applications.

As it was observed in this research that class imbalances and overlapping of microbiological features (especially between CRC and adenoma) should be the focus of future research. To resolve this issue, future research should also focus on the advanced sampling methods like oversampling or synthetic data generation to balance the classes, further improving model performance. Along with this, more type of data should be captured/integrated along with microbiome profiles like lifestyle habits, proteomic (information regarding structure of proteins), and metabolomic profiles, etc. this could provide a more broader understanding of the host-microbiome interactions and their role in colorectal neoplasia. Integration of this information could significantly enhance the predictive power of the machine learning models which would lead them to capture more non-linear relationships.

Furthermore, the future research should also focus on integrating and developing explainable AI models to increase transparency and confidence in these technologies. To obtain acceptance and integration of the model's predictions into routine clinical care, both patients and doctors must understand the reasoning behind the forecasts.

# Chapter 6

# Conclusion

In order to improve the colorectal cancer screening procedures, this study has investigated the integration of advance machine learning models with microbiome profiling. It presents an extensive approach that combines data-driven insights with biological relevance. The outperforming models of this study, Random Forest and ensemble model, showed their capability to accurately distinguish between CRC and other health statuses like adenoma and blood-negative. The findings establish an opportunity for non-invasive microbiome-based diagnostic tools in clinical settings by demonstrating that these models can detect different microbial signatures that discriminate colorectal cancer.

Even with the encouraging results, a number of challenges still remain. The models' performance differed depending on the classification tasks; models like logistic regression performed less effectively since its underlying assumptions did not align well with the characteristics of the microbiome data, emphasising the need for continous refinement of model selection and tuning strategies, especially when taking class imbalances and high dimensionality of the data into account.

The modest effect size observed in the diversity analysis further implies that, although significant, microbiome differences alone may not fully account for CRC development, underscoring the importance of taking a multifaceted approach that takes lifestyle, genetics, and environmental factors into account.

The findings of this research go beyond CRC screening, highlighting how microbiome data and machine learning can be utilised to study microbiome-disease relationships in various conditions. In order to improve predictive accuracy, the future research should use more omics data and validate these results in larger population. These tools could be further improved by utilising advanced technologies like deep learning. This research pave the way for improvements in microbiome based diagnostics as well as tailored, non-invasive CRC screening.

# Bibliography


Brenner, H., Kloor, M., & Pox, C. P. (2014). Colorectal cancer. *The Lancet*, 383(9927), 1490-1502.

Brennan, C. A., & Garrett, W. S. (2019). Fusobacterium nucleatum—symbiont, opportunist, and oncobacterium. *Nature Reviews Microbiology*, 17(3), 156-166.

Levin, B., Lieberman, D. A., McFarland, B., Andrews, K. S., Brooks, D., Bond, J., et al. (2008). Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA: A Cancer Journal for Clinicians*, 58(3), 130-160.

Lynch, S. V., & Pedersen, O. (2016). The human intestinal microbiome in health and disease. *The New England Journal of Medicine*, 375(24), 2369-2379.

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1), 7-30.

Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., et al. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11), 766.

Ahlquist, D.A. (2019). Stool-based tests vs screening colonoscopy for the detection of colorectal cancer. *Gastroenterology & Hepatology*, 15(8), p.437. Available from: `https://link.springer.com/content/pdf/10.1007/s11894-020-00770-6.pdf`.

Doocey, C.M., Finn, K., Murphy, C., & Guinane, C.M. (2022). The impact of the human microbiome in tumorigenesis, cancer progression, and biotherapeutic development. *BMC Microbiology*, 22(1), p.53. Available from: `https://link.springer.com/content/pdf/10.1186/s12866-022-02465-6.pdf`.

Elinav, E., Garrett, W.S., Trinchieri, G., & Wargo, J. (2019). The cancer microbiome. *Nature Reviews Cancer*, 19(7), 371-376. Available from: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6700740/`.

Freitas, P., Silva, F., Sousa, J.V., Ferreira, R.M., Figueiredo, C., Pereira, T., & Oliveira, H.P. (2023). Machine learning-based approaches for cancer prediction using microbiome data. *Scientific Reports*, 13(1), p.11821. Available from: https://www.nature.com/articles/s41598-023-38670-0.pdf.

Gupta, V.K., Kim, M., Bakshi, U., Cunningham, K.Y., Davis III, J.M., Lazaridis, K.N., Nelson, H., Chia, N., & Sung, J. (2020). A predictive index for health status using species-level gut microbiome profiling. *Nature Communications*, 11(1), p.4635. Available from: https://www.nature.com/articles/s41467-020-18476-8.pdf.

Half, E., Keren, N., Reshef, L., Dorfman, T., Lachter, I., Kluger, Y., Reshef, N., Knobler, H., Maor, Y., Stein, A., & Konikoff, F.M. (2019). Fecal microbiome signatures of pancreatic cancer patients. *Scientific Reports*, 9(1), p.16801. Available from: https://www.nature.com/articles/s41598-019-53041-4.pdf.

Hermida, L.C., Gertz, E.M., & Ruppin, E. (2022). Predicting cancer prognosis and drug response from the tumor microbiome. *Nature Communications*, 13(1), p.2896. Available from: https://www.nature.com/articles/s41467-022-30512-3.pdf.

Huang, K., Gao, X., Wu, L., Yan, B., Wang, Z., Zhang, X., Peng, L., Yu, J., Sun, G., & Yang, Y. (2021). Salivary microbiota for gastric cancer prediction: an exploratory study. *Frontiers in Cellular and Infection Microbiology*, 11, p.640309. Available from: https://www.frontiersin.org/articles/10.3389/fcimb.2021.640309/pdf.

Kartal, E., Schmidt, T.S., Molina-Montes, E., Rodríguez-Perales, S., Wirbel, J., Maistrenko, O.M., Akanni, W.A., Alhamwe, B.A., Alves, R.J., Carrato, A., & Erasmus, H.P. (2022). A faecal microbiota signature with high specificity for pancreatic cancer. *Gut*, 71(7), 1359-1372. Available from: https://gut.bmj.com/content/71/7/1359.

Kim, S.Y., Kim, H.S., & Park, H.J. (2019). Adverse events related to colonoscopy: Global trends and future challenges. *World Journal of Gastroenterology*, 25(2), 190. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6337013/.

Lai, Y., Masatoshi, H., Ma, Y., Guo, Y., & Zhang, B. (2022). Role of vitamin K in intestinal health. *Frontiers in Immunology*, 12, p.791565. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.791565/pdf.

Malla, M.A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., & Abd_Allah, E.F. (2019). Exploring the human microbiome: the potential future role of next-generation sequencing in disease diagnosis and treatment. *Frontiers in Immunology*, 9, p.2868. Available from: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2018.02868/pdf.

Marcos-Zambrano, L.J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., & Klammsteiner, T. (2021). Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction, and treatment. *Frontiers in Microbiology*, 12, p.634511. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.634511/pdf.

Nevola, R., Tortorella, G., Rosato, V., Rinaldi, L., Imbriani, S., Perillo, P., Mastrocinque, D., La Montagna, M., Russo, A., Di Lorenzo, G., & Alfano, M. (2023). Gender differences in the pathogenesis and risk factors of hepatocellular carcinoma. *Biology*, 12(7), p.984. Available from: https://www.mdpi.com/2079-7737/12/7/984/pdf.

Oh, M., & Zhang, L. (2020). DeepMicro: deep representation learning for disease prediction based on microbiome data. *Scientific Reports*, 10(1), p.6026. Available from: https://www.nature.com/articles/s41598-020-63159-5.pdf.

Olovo, C.V., Huang, X., Zheng, X., & Xu, M. (2021). Faecal microbial biomarkers in early diagnosis of colorectal cancer. *Journal of Cellular and Molecular Medicine*, 25(23), 10783-10797. Available from: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcmm.17010.

Ponziani, F.R., Nicoletti, A., Gasbarrini, A., & Pompili, M. (2019). Diagnostic and therapeutic potential of the gut microbiota in patients with early hepatocellular carcinoma. *Therapeutic Advances in Medical Oncology*, 11, p.1758835919848184. Available from: https://journals.sagepub.com/doi/pdf/10.1177/1758835919848184.

Poore, G.D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S.J., & Kanbar, J. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, 579(7800), p.567. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7500457/.

Rahman, M.M., Islam, M.R., Shohag, S., Ahasan, M.T., Sarkar, N., Khan, H., Hasan, A.M., Cavalu, S., & Rauf, A. (2022). Microbiome in cancer: role in carcinogenesis and impact in therapeutic strategies. *Biomedicine & Pharmacotherapy*, 149, p.112898. Available from: https://www.sciencedirect.com/science/article/pii/S0753332222002876.

Saus, E., Iraola-Guzmán, S., Willis, J.R., Brunet-Vega, A., & Gabaldón, T. (2019). Microbiome and colorectal cancer: roles in carcinogenesis and clinical potential. *Molecular Aspects of Medicine*, 69, 93-106. Available from: https://www.sciencedirect.com/science/article/pii/S0098299719300329.

Schreuders, E.H., van Roon, A., van Dam, L., Zauber, A.G., Lansdorp-Vogelaar, I., Bramer, W., Berhane, S., Deeks, J.J., Steyerberg, E.W., van Leerdam, M.E., & Spaander, M.C.

(2022). Guaiac-based faecal occult blood tests versus faecal immunochemical tests for colorectal cancer screening in average-risk individuals. *Cochrane Database of Systematic Reviews*, (6). Available from: https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD009276.pub2/pdf/full.

Sepich-Poore, G.D., Zitvogel, L., Straussman, R., Hasty, J., Wargo, J.A., & Knight, R. (2021). The microbiome and human cancer. *Science*, 371(6536), p.eabc4552. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8767999/.

Shaukat, A., & Levin, T.R. (2022). Current and future colorectal cancer screening strategies. *Nature Reviews Gastroenterology & Hepatology*, 19(8), 521-531. Available from: https://www.nature.com/articles/s41575-022-00612-y.pdf.

Tzeng, A., Sangwan, N., Jia, M., Liu, C.C., Keslar, K.S., Downs-Kelly, E., Fairchild, R.L., Al-Hilli, Z., Grobmyer, S.R., & Eng, C. (2021). Human breast microbiome correlates with prognostic features and immunological signatures in breast cancer. *Genome Medicine*, 13, p.1-17. Available from: https://link.springer.com/content/pdf/10.1186/s13073-021-00874-2.pdf.

Yang, J., McDowell, A., Kim, E.K., Seo, H., Lee, W.H., Moon, C.M., Kym, S.M., Lee, D.H., Park, Y.S., Jee, Y.K., & Kim, Y.K. (2019). Development of a colorectal cancer diagnostic model and dietary risk assessment through gut microbiome analysis. *Experimental & Molecular Medicine*, 51(10), p.1-15. Available from: https://www.nature.com/articles/s12276-019-0313-4.pdf.

Yang, Y., Du, L., Shi, D., Kong, C., Liu, J., Liu, G., Li, X., & Ma, Y. (2021). Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nature Communications*, 12(1), p.6757. Available from: https://www.nature.com/articles/s41467-021-27112-y.pdf.

Yoo, J.Y., Groer, M., Dutra, S.V.O., Sarkar, A., & McSkimming, D.I. (2020). Gut microbiota and immune system interactions. *Microorganisms*, 8(10), p.1587. Available from: https://www.mdpi.com/2076-2607/8/10/1587/pdf.

Zhang, S., Kong, C., Yang, Y., Cai, S., Li, X., Cai, G., & Ma, Y. (2020). Human oral microbiome dysbiosis as a novel non-invasive biomarker in detection of colorectal cancer. *Theranostics*, 10(25), p.11595. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7545992/.

Zheng, Y., Fang, Z., Xue, Y., Zhang, J., Zhu, J., Gao, R., Yao, S., Ye, Y., Wang, S., Lin, C., & Chen, S. (2020). Specific gut microbiome signature predicts the early-stage lung cancer. *Gut Microbes*, 11(4), 1030-1042. Available from: https://www.tandfonline.com/doi/pdf/10.1080/19490976.2020.1737487.

Young, C., Wood, H.M., Balaguer, A.F., Bottomley, D., Gallop, N., Wilkinson, L., Benton, S.C., Brealey, M., John, C., Burtonwood, C., Thompson, K.N., Yan, Y., Barrett, J.H., Morris, E.J.A., Huttenhower, C., & Quirke, P. (2021). Microbiome analysis of more than 2,000 NHS bowel cancer screening programme samples shows the potential to improve screening accuracy. *Clinical Cancer Research*, 27(8), 2246-2254. Available from: https://clincancerres.aacrjournals.org/content/27/8/2246.

Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1), 32-46. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1442-9993.2001.01070.pp.x.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289-300. Available from: https://www.jstor.org/stable/2346101.

Bray, J.R., & Curtis, J.T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, 27(4), 325-349. Available from: https://www.jstor.org/stable/1942268.

Mann, H.B., & Whitney, D.R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1), 50-60. Available from: https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-18/issue-1/On-a-Test-of-Whether-one-of-Two-Random-Variables/10.1214/aoms/1177730491.full.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-423. Available from: https://ieeexplore.ieee.org/document/6773024.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), 25.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11), 2783-2792.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Menard, S., 2002. *Applied Logistic Regression Analysis*. 2nd ed. Thousand Oaks, CA: SAGE Publications.

Ng, A.Y., 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, Banff, Canada, 4-8 July 2004. New York: ACM, p. 78.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), pp.1189-1232.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems*, vol. 30, pp.3146-3154.

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp.2623-2631.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), pp.1145-1159.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, vol. 2, pp.1137-1143.

Powers, D.M., 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), pp.37-63.

# Appendix

This section contains all the important functions and codes.

## R Code: Function to extract D4 level from the column name

```r
extract_d4 <- function(colname) {

  parts <- str_extract_all(colname, "D_[0-5]__[A-Za-z0-9.]+(?=\\.D|$)")[[1]]
  print(colname)  #debug
  print(parts)    # debug

  if (length(parts) >= 5) {
    d4_level <- paste(parts[1:5], collapse = ".")
    print(d4_level)
    return(d4_level)
  } else {
    print(paste("Invalid:", colname))
    return(NA)
  }
}
```

## R Code: Aggregation of features with same D4 levels

```r
for (d4 in unique_d4_levels) {
  print(d4)
  matching_cols <- grep(d4, colnames(data_csv), value = TRUE)
  print(matching_cols)
  # Initialising a vector to store the combined values for the current D4
      level
  combined_values <- numeric(nrow(data_csv))

  # Looping through each row and calculate the sum for the current D4 level
  for (i in 1:nrow(data_csv)) {
    row_values <- data_csv[i, matching_cols]
    combined_values[i] <- sum(row_values, na.rm = TRUE)
  }

  combined_data[[d4]] <- combined_values
}
}
```

### R Code: Converting count data to ratio data

```r
microbiome_data <- data[, microbiome_cols]
relative_abundance_data <- microbiome_data

for (i in 1:nrow(microbiome_data)) {
  row_sum <- sum(microbiome_data[i, ])
  if (row_sum > 0) {
    relative_abundance_data[i, ] <- microbiome_data[i, ] / row_sum
  } else {
    relative_abundance_data[i, ] <- 0
  }
}
```

### Python Code: Alpha Diversity Analysis

```python
alpha_diversity_shannon = microbiome_data.apply(skbio.diversity.alpha.shannon
    , axis=1)

# Summary statistics for each group
mean_shannon = test_data.groupby('grp')['Shannon Diversity Index'].mean()
median_shannon = test_data.groupby('grp')['Shannon Diversity Index'].median()
mode_shannon = test_data.groupby('grp')['Shannon Diversity Index'].apply(
    lambda x: x.mode().iloc[0] if not x.mode().empty else float('nan'))

statistics_df = pd.DataFrame({
    'Mean Shannon Index': mean_shannon,
    'Median Shannon Index': median_shannon,
    'Mode Shannon Index': mode_shannon
})

# Perform the Kruskal-Wallis test
kruskal_stat, kruskal_p_value = kruskal(*data_by_group)

# Calculate eta squared (    ) effect size for Kruskal-Wallis
n = len(test_data)
eta_squared = (kruskal_stat - len(groups) + 1) / (n - len(groups))
```

### Python Code: Beta Diversity Analysis

```
# Calculating Bray-Curtis dissimilarity matrix
bray_curtis_dm = beta_diversity('braycurtis', microbiome_data, data.index)

# PCoA
pcoa_results = pcoa(bray_curtis_dm)

# Perform PERMANOVA
distance_matrix = beta_diversity('braycurtis', microbiome_data, ids=
    microbiome_data.index)

results = permanova(distance_matrix, test_data['grp'], permutations=999)
print(results)

groups = test_data['grp'].unique()
pairwise_results = {}

for group1, group2 in combinations(groups, 2):
    subset_data = test_data[(test_data['grp'] == group1) | (test_data['grp']
        == group2)]
    subset_microbiome = microbiome_data.loc[subset_data.index]
    distance_matrix = beta_diversity('braycurtis', subset_microbiome, ids=
        subset_microbiome.index)
    result = permanova(distance_matrix, subset_data['grp'], permutations=999)
    pairwise_results[(group1, group2)] = result
```

### Python Code: Differential Abundance Analysis (for one comparison)

```
# Function to perform Wilcoxon rank-sum test
def perform_wilcoxon(group1, group2):
    results = []
    for col in microbiome_data.columns:
        stat, p_value = mannwhitneyu(group1[col], group2[col])
        results.append((col, stat, p_value))
    return pd.DataFrame(results, columns=['Feature', 'Statistic', 'p-value'])

diff_abundance_results = perform_wilcoxon(group1, group2)

# Adjusting p-values for multiple testing (Bonferroni correction)
diff_abundance_results['adjusted_p-value'] = diff_abundance_results['p-value'
    ] * len(diff_abundance_results)

significant_results = diff_abundance_results[diff_abundance_results['
    adjusted_p-value'] < 0.05]

significant_results = significant_results.sort_values('adjusted_p-value')
```

### Python Code: Model Evaluation Function

```
def evaluate_model(y_true, y_pred, y_pred_prob, title='Model Evaluation'):
    print(classification_report(y_true, y_pred))

    cm = confusion_matrix(y_true, y_pred)
    disp = ConfusionMatrixDisplay(confusion_matrix=cm)
    disp.plot()
    plt.title(f'Confusion Matrix: {title}')
    plt.show()

    fpr, tpr, _ = roc_curve(y_true, y_pred_prob)
    roc_auc = auc(fpr, tpr)
    plt.figure()
    plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area = {
        roc_auc:0.2f})')
    plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title(f'Receiver Operating Characteristic (ROC): {title}')
    plt.legend(loc="lower right")
    plt.show()

    precision, recall, _ = precision_recall_curve(y_true, y_pred_prob)
    plt.figure()
    plt.plot(recall, precision, color='blue', lw=2)
    plt.xlabel('Recall')
    plt.ylabel('Precision')
    plt.title(f'Precision-Recall Curve: {title}')
    plt.show()

    f1 = f1_score(y_true, y_pred)
    print(f'F1-Score: {f1}')
```

## Python Code: Function for Cross-Validation and mean AUC-ROC Curve

```
# Function for cross-validation and CV AUC plot.
def CV_ROC(model, X, y, n_splits=10):
    cv = StratifiedKFold(n_splits=n_splits, shuffle=True, random_state=42)
    tprs = []
    aucs = []
    mean_fpr = np.linspace(0, 1, 100)

    plt.figure(figsize=(10, 8))
    for i, (train_idx, val_idx) in enumerate(cv.split(X, y)):
        model.fit(X.iloc[train_idx], y.iloc[train_idx])
        y_val_pred_prob = model.predict_proba(X.iloc[val_idx])[:, 1]
        fpr, tpr, _ = roc_curve(y.iloc[val_idx], y_val_pred_prob)
        tprs.append(np.interp(mean_fpr, fpr, tpr))
        roc_auc = auc(fpr, tpr)
        aucs.append(roc_auc)
        plt.plot(fpr, tpr, lw=1, alpha=0.3, label=f'ROC fold {i+1} (AUC = {
            roc_auc:.2f})')

    mean_tpr = np.mean(tprs, axis=0)
    mean_tpr[-1] = 1.0
    mean_auc = auc(mean_fpr, mean_tpr)
    std_auc = np.std(aucs)
    plt.plot(mean_fpr, mean_tpr, color='b', lw=2, alpha=0.8, label=f'Mean ROC
        (AUC = {mean_auc:.2f}    {std_auc:.2f})')

    plt.plot([0, 1], [0, 1], linestyle='--', lw=2, color='r', alpha=0.8)
    plt.xlim([-0.05, 1.05])
    plt.ylim([-0.05, 1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('Cross-Validation ROC')
    plt.legend(loc='lower right')
    plt.show()

    print(f"Cross-validation AUC Scores: {aucs}")
    print(f"Mean Cross-validation AUC: {mean_auc:.4f}")
```

### Python Code: Hyperparameter Tuning using OPTUNA (for a single model)

```
def objective_rf(trial):
    n_estimators = trial.suggest_int('n_estimators', 50, 1000)
    max_depth = trial.suggest_int('max_depth', 2, 50)
    min_samples_split = trial.suggest_int('min_samples_split', 2, 100)
    min_samples_leaf = trial.suggest_int('min_samples_leaf', 1, 50)
    max_features = trial.suggest_categorical('max_features', [ 'sqrt', 'log2'
        , None])

    rf = RandomForestClassifier(
        n_estimators=n_estimators,
        max_depth=max_depth,
        min_samples_split=min_samples_split,
        min_samples_leaf=min_samples_leaf,
        max_features=max_features,
        random_state=42
    )

    rf.fit(train_x, train_y)

    y_pred_prob = rf.predict_proba(test_x)[:, 1]
    y_pred = (y_pred_prob > 0.5).astype(int)
    accuracy = accuracy_score(test_y, y_pred)
    return accuracy

# Optimize hyperparameters using Optuna
study_rf = optuna.create_study(direction='maximize')
study_rf.optimize(objective_rf, n_trials=70)
```