UPPSALA
UNIVERSITET

UPTEC X 22031
Examensarbete 30 hp
Juni 2022

Metagenomic analysis of Crohn's Disease

Gustav Ahlström

Civilingenjörsprogrammet i molekylär bioteknik

UPPSALA
UNIVERSITET

Metagenomic analysis of Crohn's Disease

Gustav Ahlström

Abstract

Inflammatory Bowel Disease (IBD) is a chronic and incurable condition that is increasing in prevalence across the globe. This illness consist of two forms: Crohn's Disease (CD) and Ulcerative Colitis (UC). CD is characterised by a patch inflammation pattern across the gut and a multitude of different factors, such as diet. Contemporary research has found a link between gut dysbiosis and the development of IBD, suggesting that the microbial flora colonising the gut have a vital part to play in the development of CD.

This paper aims to identify taxa associated with CD. This is done through the application of machine learning algorithms as standard univariate statistical methods fail to apply in the highly interdependent domain of the gut microbiome. The compositionally of the data and external factors influencing variance in the data will be taken into account.

After applying a Center Log ratio transformation (CLR) to a MetaPhlAn3 taxonomic profile and using a random forest classifier the following five taxa were identified as the most important in the association to CD: *Ruminococcaceae bacterium*, *Akkermansia muciniphila, Streptococcus parasanguinis*, *Flavonifractor plautii* and *Bifidobacterium bifidum*.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Stefanie Prast-Nielsen Ämnesgranskare: Sergi Sayols

Examinator: Siv Andersson

1 Populärvetenskaplig sammanfattning

Den mänskliga kroppen innehåller mer än ett miljardtal mikroorganismer i tarmen som utövar väsentliga biologiska processer. Tarmikrobiomets komposition beror mycket på vad man äter, hur man lever och vilka gener man har. Crohns sjukdom är en tarminflammation som beror delvis på en dysbios, en förändring i tarmflorans komposition som påverkar hälsan negativt. Crohns sjukdom är kronisk och kan inte botas. Sjukdomen är inte dödlig men har en negativ påverkan på livskvalitén och kan leda till tarmcancer. Symptomen av Crohns sjukdom inkluderar diarré, magknip och blod i avföringen. För att utveckla botemedel mot denna sjukdom är det väsentligt att känna till vilka mikrober som orsakar denna dysbios. Men mikrobiom-data har egenskaper som man behöver känna till innan man kan dra några slutsatser.

Vanliga statistiska metoder kan inte användas när man utforskar vilka mikrober påverkar Crohns sjukdom. Detta beror både på vilka instrument som används för att sekvensera fekala prover, externa faktorer och de komplexa relationerna inom mikrobiomet. Kompositionell data kallas den data som skapas av sekvenseringsmaskiner och kan beskrivas som att man studerar en del av en helhet. Problemet med kompositionella data är att endast prediktioner inom vad som har observerats går att tolka. För att kunna tolka prediktioner utanför detta så måste den kompositionella datan transformeras. Den mänskliga kroppen är komplex. Det finns många olika faktorer som kan påverka mikrobiomet data som är oberoende av Crohns sjukdom, till exempel om någon röker kan detta påverka tarmmikrobiomet. Dessa faktorer behöver identifieras och faktorer med störst påverkan behöver behandlas. Detta görs genom att jämföra observationer som håller dessa faktorer konstant, t.ex. att man jämför endast icke-rökande patienter mot varandra.

Vanliga statistiska metoder jämför observationer från experiment med kända sannolikhetsfördelningar. Dessvärre är det inte lika lätt att jämföra observationer från mikrober. Flera olika mikrober kan påverka varandra och detta fenomen modelleras inte av kända sannolikhetsfördelningar. I så fall finns det mer komplexa och anvancerade metoder som kan ta hänsyn till detta. I detta projekt användes maskininlärningsmetoder. Flera olika maskininlärningsmetoder finns och det är inte alltid uppenbart vilken metod som ska användas i vilken kontext, utan metoder behöver jämföras med varandra. Efter att ha tagit hänsyn till externa faktorer och funnit en optimal maskininlärningsmetod kan det visas att mikrober som associeras främst med en hälsosam tarm är de mikrober som är avgörande i en prognos av Crohns sjukdom.

Table of contents

2 Abbreviations

Abbreviations	
Additive Log-Ratio	ALR
Area Under the Curve PR	AUCPR
Area Under the Curve ROC	AUCROC
Canadian Dollars	CDN
Centered Log-Ratio	CLR
Crohn's Disease	CD
Center for Translational Microbiome Research	CTMR
Cyclooxygenase	COX
False Negative	FN
False Positive	FP
False Positive Rate	FPR
Inflammatory Bowel Disease	IBD
Isometric Log-Ratio	ILR
K-Nearest Neighbor	KNN
Last Common Ancestor	LCA
Leave-one-out cross-validation	LOOCV
Logistic Regression	LR
Machine Learning	ML
Muramyl Dipeptide	MDP
Neural Networks	NN
Nucleotide Binding Oligomerization Domain containing 2	NOD2
Nonsteroidal Anti-inflammatory Drugs	NSAIDs
Pairwise Log-Ratio	PWLR
Polymerase Chain Reaction	PCR
Precision-Recall	PR
Random Forest	RF
Receiver Operator Characteristic	ROC
Ribosomal RNA	rRNA
Short Chain Fatty Acids	SCFAs
Support Vector Machines	SVM
True Negative	TN
True Positive	TP
True Positive Rate	TPR
Ulcerative Colitis	UC
United Sates of America	USA
United Sates of America Dollars	USD

3 Introduction

This project aims to investigate differences between the microbiomes of individuals suffering from Crohn's Disease (CD), which is a form of inflammation in the gut, and healthy individuals. Symptoms of CD can be varied but a few common symptoms include the following: fatigue, weight loss, loose stool and rectal bleeding. CD is a chronic and incurable disease, though remissions can occur in some cases (Kaplan 2015; Zhang & Li 2014). CD is characterized by damaged areas in the gastrointestinal tract, dispersed in a "patchy" fashion, where the inflammation may reach through multiple layers of the gastrointestinal tract (Kaplan 2015; Zhang & Li 2014). Since the 1950s, Inflammatory Bowel Disease, which CD is a form of, has increased in prevalence across the western world. Currently, healthcare costs to treat the 2-3 million European patients every year, are around €4-6 million (Kaplan 2015). These costs are covered by healthcare services but omit the costs of lowered productivity and life quality for patients suffering from IBD. By characterising IBD, new diagnostic techniques and biomarkers could be developed for easier diagnosis as well as "personalised medicine" for patients (Lacroix *et al.* 2021). Previous work on this topic has been performed, identifying changes in microbiota composition as well as metabolites but the aetiology of IBD still needs further investigation (Lacroix *et al.* 2021; Zhang & Li 2014). CD has a plethora of inter playing factors which could cause it to manifest, increasing the complexity of CD pathogenesis. Three distinct sets of factors play a role in development of CD: genetics, environmental factors and the patients microbiota, which will be further elaborated on. Children whose parents suffer from CD have a 2-14% chance of developing CD and there are specific loci, 163, linked to IBD development, that have been discovered thus far (Zhang & Li 2014; Ananthakrishnan 2015). A less diverse microbiome has been associated with increased risk of developing IBD as well. Presence of certain microbes, such as *Escherichia coli* (AIEC), is also associated with an increased risk of developing CD. Finally environmental factors play a role in the development of IBD; smoking, usage of antibiotics in early childhood and a diet with low fiber intake are such factors (Zhang & Li 2014; Ananthakrishnan 2015).

4 Background

4.1 The human microbiome

Being the most densely populated microbe communities on earth the microbes in the human gut live in a symbiotic relationship with their human hosts (Loyd-Price *et al.* 2016). Microbes form dynamic ecosystems on or in the human body with a variety of microbes which fluctuate in compositions depending on external factors such as diet, lifestyle and antibiotic usage as well as internal dynamics within said microbiome community (A Gilbert *et al.* 2018; Loyd-Price *et al.* 2016). Colonization of the gut begins right after birth, with ca 10^{13} to 10^{15} microbes colonizing the human gut within a year of birth while becoming more firmly established as a human ages. Once established these communities are resilient to change and alterations, relative to when the human body is first colonized (G. Albenberg *et al.* 2012; A Gilbert *et al.* 2018).

What can be said about a humans gut microbiome, is that it can have a drastic effect on the human hosts health (Loyd-Price *et al.* 2016; A Gilbert *et al.* 2018). Dysbiosis can be considered a disorder in the microbial community that can prolong, exacerbate or induce detrimental effects on someones health. Dysbiosis in the human gut has been observed, though no causality has been proven for some illnesses, to be associated with the development of:

- Autism
- Obesity
- Clostridium difficile infection
- IBD and many other diseases

Given the intricacy of the human gut, it is hard to define what a ”healthy human gut microbiome” would look like and, in turn, what is considered ”dysbiosis” as development of a humans microbiome is highly personalized and depends on multiple factors (Loyd-Price *et al.* 2016). Loyd-Price et al. in 2016 present an alternative hypothesis that there is a ”functional core” that can define a healthy microbiome (Loyd-Price *et al.* 2016). Central to this alternative hypothesis is that the microbiome present in an environment have to full fill certain metabolic functions in order for someone to have a ”healthy” gut.

While the functional human genome can be regarded as immense, with ca 22,000 genes present, it is dwarfed by the 3.3 million genes present in the human gut microbiome (Ursell *et al.* 2012). Through the gut microbiome humans are able to perform metabolic functions which they can not perform naturally, such as fermentation of indigestible carbohydrates into short chain fatty acids, synthesis of certain vitamins and bio-transformation of conjugated bile acids (G. Albenberg *et al.* 2012). Another service which the gut microbiome provides is the ability to moderate an individuals immune response, reducing the risk of the host developing allergies. In return the human gut provides an optimal environment for the microbes to flourish by producing a physical barrier between the world and the microbes while granting regular sustenance to the microbes (G. Albenberg *et al.* 2012). Another service which is provided by the gut microbiome is granting resistance towards pathogenic colonisation; commensal gut bacteria occupy niches within the gut ecology, limiting the opportunities which pathogens can exploit. Commensal gut bacteria also compete for sustenance against pathogens making colonisation even more difficult (Martín *et al.* 2013).

Furthermore, there are multiple gut-organ axes, with the ability of changing organ dynamics. Such axes include: the gut-brain axis, the gut-skin axis and the gut-heart axis, to name a few. The gut-brain axis is bifacial, meaning that the gut can influence the brain while the brain has the same ability to influence the gut (Ahlawat *et al.* 2020). This can occur through a direct link from the central nervous system to the gut, called the vagus nerve, as well as humoral and endocrine pathways. Dysbiosis of the gut has been associated with different neurological illnesses, such as Alzheimer's disease and Autism (Ahlawat *et al.* 2020). The gut-skin axis is not well understood to this day but is hypothesized to involve communication though metabolites, neurons and endocrines between the skin and microbiome. Gut disorders have been associated with cutaneous manifestations (Ahlawat *et al.* 2020). A gut-heart axis exists as well. Patients who experience heart related problems are at higher risk of having gut microbiome related problems, such as dysbiosis and lower levels of microbiome diversity. These problems can include, but are not limited to, coronary heart disease and heart failure (Ahlawat *et al.* 2020).

4.2 IBD

IBD is a chronic and incurable disease which is composed of two major forms: CD and UC. Both illnesses share symptoms but are different; CD is defined has having a ”patchy” inflammation pattern across the gut while UC has a more ”consistent” inflammation usually limited to the colon specifically. CD is also associated with more complications, such as the development of abscess and strictures of the colon, unlike

UC, which in turn increase the chances of requiring medical attention or care (Kaplan 2015). Symptoms of each illness include but are not limited to; fatigue, diarrhea, blood in stool and abdominal pain. While not a deadly disease, IBD has a detrimental effect on a patients well being and has been associated as a precursor to more server diseases, such as colon cancer (Kaplan 2015).

IBD has been increasing in prevalence across the globe since the 1950s as regions across the globe industrialise, such as Asia, the Middle East and South America. Currently more than 3.5 million individuals world wide are suffering from this illness which incurs a cost to the society as a whole (Kaplan 2015). In 2004 ca more than 1 million individuals in the United States of America (USA) suffered from IBD which accumulated medical costs exceeding $6 billion United Sates Dollars (USD). In Canada, during the same year, ca 200,000 individuals suffered from IBD with total direct medical costs tallying to $ 1.2 billion Canadian Dollars (CDN) (Kaplan 2015). In Europe the effects of IBD had a similar effect of healthcare costs; 2.5-3.5 million patients totaling a direct-healthcare cost of ca 4.6-5.6 billion Euros annually. While these costs capture a larger societal costs of IBD, they fail to consider the impact of IBD on an individual level; a decreased quality of life. In turn such an effect could have indirect costs, for example loss of productivity, which costs are much more difficult to measure (Kaplan 2015; Lloyd-Price *et al.* 2019).

IBD is a disease that is complicated to study as there are multiple factors that play a role in developing either CD or UC. Such factors include; microbial factors, environmental factors, genetics and nutrition (Zhang & Li 2014). These factors will be further explained below.

4.2.1 The Microbiome and IBD

The microbiome, as previously mentioned, plays an important role in homeostasis of the gut. Both UC and CD is associated with a dysbiosis in the gut microbiome and a reduction in biodiversity, though this effect is more pronounced in CD compared to UC. Another hallmark of IBD is a fluctuation of the dominating taxa in the gut; a more unstable microbiome (Zhang & Li 2014; Ananthakrishnan 2015). To further tie in the importance of the gut microbiome in development of IBD, the age at which IBD is most likely to be diagnosed reflects natural changes within the microbiome. Early onset IBD usually occurs around the age of 10, when the microbiome is still changing due to puberty, alterations in diet and illness. Late onset IBD occurs around the age of 60+, at that age there is a marked increase in instability within the microbiome (Zhang & Li 2014; Ananthakrishnan 2015; D.Kostic *et al.* 2014).

Additionally, both CD and UC is associated with a specific reduction in abundance of specific taxon and an increase in other non-commensal taxa. An example of a pathogenic

taxon which has been associated with IBD is adherent-invasive *Escherichia coli* (*E. coli* (AIEC))(Zhang & Li 2014). It has been observed in control populations that ca 6.2 % of a population has this bacteria in their gut but for patients suffering from CD this relative abundance has increased to 22 % (Zhang & Li 2014; Ananthakrishnan 2015). *E. coli* (AIEC) has the ability to not only evade macrophages but also has the ability to invade the epithelium, which has two layers of mucus protecting it from direct contact with microbes (Zhang & Li 2014; Ananthakrishnan 2015; D.Kostic *et al.* 2014). This gives *E. coli* a natural fitness against the hosts immune systems as well as possibilities to carve out new niches. Another group of bacteria that is correlated with an IBD diagnosis is the *Fusobacteria* group. Much like *E. coli* (AIEC) *Fusobacteria* have the ability to penetrate the epithelium's protective layers. *Fusobacteria* do exist in the oral microbiome of humans naturally but for patients suffering from UC this group of bacteria are also present in the gut. It has been shown that human derived *Fusobacteria* in mouse models shows colonic mucosal erosion. *Fusobacteria* are also involved in tumorigenesis of colon cancer in mouse models, suggesting a link between IBD and colon cancer (Zhang & Li 2014; Ananthakrishnan 2015; D.Kostic *et al.* 2014).

It has to be mentioned that while there are pathogenic taxa associated with IBD, commensal taxa present in the gut have protective attributes against IBD. Simply by occupying niches within the gut environment, commensal taxa hinder pathogens from colonizing the gut by outcompeting them. Commensal taxa also help to protect against inflammation of the gut through different biosynthesis pathways. Bacteria in the genera of *Bifidobacterium*, *Lactobacillicus* and *Faecalibacterium* help their human host against IBD by down-regulating pro-inflammatory cytokines while up-regulating anti-inflammatory cytokines, such as interleukin 10 (Zhang & Li 2014; Ananthakrishnan 2015). *Faecalibacterium prausnitzii* is such a bacterium; patients with lower abundance of *F. prausnitzii* in their gut have a higher chance of recurring CD and conversely patients who have managed to regain *F. prausnitzii* after a relapse are correlated with maintaining a remission status of UC. Several commensal taxa are able to ferment dietary fiber produce Short Chain Fatty Acids (SCFAs) which have multiple benefits to the human host such as: providing an energy source for epithelial cells, maintenance of intestinal barrier integrity and mucus production (Zhang & Li 2014; Ananthakrishnan 2015). Epithelial cells in turn allow for the production of T cells in the colon. Reduction in taxa digesting these dietary fibers also correlates with IBD, specifically *Odoribacter* and *Leuconostocaceae* for UC and *Phascolarctobacterium* and *Roseburia* for CD (Zhang & Li 2014; Ananthakrishnan 2015; D.Kostic *et al.* 2014).

4.2.2 Environmental Factors and IBD

One of the most widely studied risk factors correlated with the development of IBD is smoking tobacco. There is currently no robust relationship with tobacco usage and IBD as smoking has a divergent effect on CD and UC (Zhang & Li 2014; Ananthakrishnan 2015). Smoking is usually correlated with a CD diagnosis and a more aggressive diagnosis, with requirements of surgery and immunosuppression in some cases, and higher odds to relapse. On the other hand smokers with UC experience a less aggressive version of UC with less stringent requirements of surgery or the usage of medication (Zhang & Li 2014; Ananthakrishnan 2015). If these UC patients were to stop smoking then there is an elevated risk of them developing UC after 2-5 years of quitting. This elevated risk remains up to 20 years after ending the habit. Another reason as to why there is an unclear connection between IBD and smoking is the complex genetic interactions with smoking. Genetic polymorphisms contribute to nicotine metabolism which might modify susceptibility to IBD. There many interpersonal factors which impact the effect of smoking and illness, such as gender, suggesting adding another layer of complexity in the relationship between smoking and IBD (Zhang & Li 2014; Ananthakrishnan 2015).

Antibiotics and other medication usage are also tied to an IBD diagnosis. During the early years of someones life their gut microbiome is highly volatile in its composition. A case-control analysis from the University of Manitoba showed that pediatric patients who suffered from IBD were more likely to have used antibiotics within the first year of their lives (58 %) compare to controls, where only 39 % had used antibiotics in the same time frame (N. Ananthakrishnan 2013; Ananthakrishnan 2015). This association with IBD and antibiotics has been observed across all classes of antibiotics and is dose dependent. However, it is also difficult to ascertain if this link between IBD and antibiotics is causal or not. Populations in Asia, who have a more exposure to microbes in their youth due in part to their less sanitary living conditions compared to western populations, experience a protective effect from antibiotics (N. Ananthakrishnan 2013; Ananthakrishnan 2015).

Other than antibiotics there are more medications that are associated with an IBD diagnosis. Two examples would be aspirin and Nonsteroidal Anti-inflammatory Drugs (NSAIDs). Using NSAIDs in a higher dosage and for a longer time had a direct correlation to the odds of developing either CD or UC (N. Ananthakrishnan 2013; Ananthakrishnan 2015). Up to a third of patients relapses might be triggered by NSAIDs usage. A hypothesised reason as to why this may be the case is that NSAIDs are nonspecific in their inhibiting effect on Cyclooxygenase (COX) enzymes as specific COX-2 inhibitors are associated with a reduced rate of relapse. The connection between asprin and IBD is not as clear; previous studies have both found a increased chance of developing IBD while other have come to the opposite conclusion, that no association between dosage,

duration of usage and frequency of usage correlated with an increased risk of developing IBD (N. Ananthakrishnan 2013; Ananthakrishnan 2015; Zhang & Li 2014).

4.2.3 Genetics and IBD

Currently, 163 gene loci have been identified to be associated with IBD; 110 overlap between CD and UC while 30 are strictly associated with CD and 23 strictly associated with UC. These genes can be broadly divided into multiple classes of genes involved in the following metabolic processes: innate immune response, adaptive immune response, autophagy, maintenance of the epithelial barrier and more. No single gene can be described as causal for IBD but simply a key point in a much larger picture (Zhang & Li 2014; Ananthakrishnan 2015).

Nucleotide binding Oligomerization Domain containing 2 (NOD2) is a gene strongly correlated with CD and involves recognizing Muramyl Dipeptide (MDP) which is a conserved motif present in the peptidoglycan across both gram positive and gram negative bacteria. NOD2 was the first gene associated with CD, and it was characterized in 2001. Stimulation of MDP induces autophagy (N. Ananthakrishnan 2013; Ananthakrishnan 2015). Autophagy in turn is vital for intracellular homeostasis by aiding in infection defense and microbial digestion. Homozygosity at the NOD2 locus is correlated with a 20-40 fold increase in risk of developing CD. Being heterozygous at this loci reduces this to a 2-4 fold increase in the risk of developing CD. However other autophagy-related genes, such as Immunity Related GTPase M, are also related to IBD development (Zhang & Li 2014; Ananthakrishnan 2015).

Family history plays a role in the development of IBD. If a child has two parents, both of which have had or are suffering from IBD, the chance of said child developing IBD before the age of 30 is 33 % (Ananthakrishnan 2015). If a child has a family history of IBD, their odds of developing CD is between 2-14 % while 8-14 % chance to develop UC (Ananthakrishnan 2015). While a critical connection in the development of IBD, genetic factors do not account for more than 20-25 % of the heritability of IBD, suggesting a strong environmental effect of the development of IBD (Zhang & Li 2014; Ananthakrishnan 2015). This phenomena, known as the ”genetic vacuum”, is not unique for IBD but has been observed in other polygenetic diseases. A suggested reason for this relatively low level of inheritance is that interactions between genes and interactions between the products of genes could account for more of the inheritance than the genes themselves (Zhang & Li 2014; Ananthakrishnan 2015).

4.2.4 Nutrition and IBD

The diet of an individual has a vital impact upon which microbes colonize their gut and in turn which functionalities are present in said individual. For example microbial genes encoding for genes involved in synthesis of carbohydrates and amino acids will vary depending on if the gut microbiome is exposed to a carnivorous diet or a herbivore diet. Long-term dietary changes have a pronounced effect on the gut microbiomes composition. (G. Albenberg *et al.* 2012; A Gilbert *et al.* 2018).

Fiber intake from fruits and vegetables have been inversely correlated with developing CD, while having a less pronounced effect on UC. Pediatric patients who have developed CD have a lower intake of fruits and vegetables. Intake of fruits and high fiber food stuffs is negatively associated with developing CD. Conversely risks for developing UC were mitigated by consuming high amount of vegetables (N. Ananthakrishnan 2013; Ananthakrishnan 2015; G. Albenberg *et al.* 2012). Intake of vitamin D has also been associated with development of IBD as vitamin D has a immunological role in the body. Mouse models have shown that inhibition of vitamin D receptors or a deficiency in vitamin D increases the odds of developing UC. Amelioration of said deficiency suppresses expression of pro-inflammatory genes. Furthermore individuals have an increased chance of suffering from IBD if they lack vitamin D intake (N. Ananthakrishnan 2013; Ananthakrishnan 2015).

Diets involving a high intake of polyunsaturated fats, omega-6 fatty acids and meat associate with a higher risk of developing CD and UC. (G. Albenberg *et al.* 2012). IBD, specifically CD, has also been associated with intake of ultra-processed-foods. Ultra-processed-foods, according to the NOVA classification system, are ready-to-eat food items which have been extensively pre-processed before consumption (Lo *et al.* 2021). Additionally; emulsifiers, sweeteners, preservatives and other substances are added to these foodstuffs for additional qualities. Consumption of these items is associated with an increase in the risk for developing CD with a hazard ratio of 1.18. This is speculated to be due to the presence of substances such as salt, artificial sweeteners and nanoparticles which are associated with developing inflammation and expansion of pro-inflammatory taxa (Lo *et al.* 2021).

From the above description of IBD it becomes clear that to get a better and nuanced understanding of this disease a multi-omics approach must be made as each causal factor is interdependent on one another. However due to limitations in time and resources only the microbiome will be considered in this thesis.

4.3 Machine Learning

Broadly speaking Machine Learning (ML) is a subset of Artificial Intelligence technology which involves identifying patterns in data. ML can be divided into two separate approaches, supervised machine learning and unsupervised machine learning. The latter visualises general trends and patterns in data while the former tries to predict future observations by ”learning” from previous observations of a specific phenomena (Helm *et al.* 2020; Wei Khor & Yuan Ngiam 2019). Supervised approaches can be further divided into classification approaches, dealing with discrete data, and regression approaches, dealing with continuous data. More specific and niche definitions and characterizations of ML algorithms do exist, such as whether or not batch data is used or continuously streamed data is used, but for the purposes of this thesis going into more detail is not required. There are multiple available methods for each approach with no ”golden rule”; no one method is given to outperform another method for the given problem (Helm *et al.* 2020; Wei Khor & Yuan Ngiam 2019; Holzinger 2018; Zhou *et al.* 2017).

4.3.1 Machine Learning algorithms

In this project five classification algorithms have been used. These are Random Forests (RF), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Logistic Regression (LR) and Neural Networks (NN).

RF methods can be applied to both regression and classification problems. RF methods work as an ensemble of decision trees, where each tree is trained on a random subset of the training data. Each decision trees partitions the data into smaller and smaller parts, and at each partition making the resulting partitions more homogeneous with respect to their class. When a new observation is presented to the RF model all decision trees in the forest then vote on which class it belongs to, where the majority vote wins (Breiman 2001; Cutler *et al.* 2007).

KNN algorithms can be summarised with the proverb ”birds of a feather flock together”. The algorithm works by computing the distance between an unknown observation to K known samples. Whichever majority class is represented within the K-nearest neighbors, will then be applied to the unknown sample. This algorithm can be tuned using different distance metrics as well as different values of K. Generally speaking very small values of K, say 1, gives way to over fitting while larger values can cause under fitting (Kramer 2013; Mucherino *et al.* 2009).

Much like RF algorithms, SVMs can be applied to both classification and regression

problems. In the case of classification applications SVMs attempt to divide a data set, as well as possible, into two separate classes using a hyperplane, where the hyperplane is modeled using the closest, but separate, observations which are called the support vectors. The hyperplane is then the average distance between all support vectors. If the current dimensions do not allow for such a separation the algorithm will consider data in higher and higher dimensions until it is possible to separate the two classes. Future observations will then be able to be classified using a decision function which is derived from an optimal hyperplane (Hearst *et al.* 1998; Meyer 2015).

LR can be regarded as a form of linear regression but of discrete data. Unlike linear regression, which tries to find a "line-of-best-fit" through available data points in order to predict future observations, LR tries model the probability of an observation belonging to a class given a set of independent parameters. A LR model will look like the following:

$$\ln(\mathbb{P}(Y = 1)) = \beta_0 + \beta_1 * X_1 + ... + \beta_n * X_n$$

Where X is a independent feature from an observation, $\beta_1 ... \beta_n$ are weights for each feature and β_0 is the intercept term. Y signifies the class which is of interest. These β terms are tuned in order to get a better fit (LaValley 2008).

NN can be considered analogous to biological neurons where a signal is propagated through a series of neurons in order for an action to occur. A NN is divided into many layers, with each layer having a number of neurons and a bias neuron connected to it. There is a input layer, where each input feature has its own neuron, and an output layer, where a number of neurons exist, each representing a different class. At each neuron, the weighted input from neurons from the previous layer is summed up ta value. This value is then parsed to a activation function which determines if the current neuron shall send a signal to the next layer or not. How a NN learns is by tuning each weight and bias from one layer to the next (Bishop 1994).

While ML approaches show promise when applied in practical problems as well as aiding researchers in their respective fields of research, ML is not a "silver bullet"; large quantities of data are required for good performance, extensive computer resources are needed (such as HPC clusters) and each method takes time to train. Certain problems do not even need ML approaches even though they can be phrased as classification or regression problems. Such an example would be allosome determination in ancient DNA samples where pre-existing methods solve this problem by comparing fractions of DNA sampled that stem from an allosome, even though this problem can be phrased as a binary classification problem (de Flamingh *et al.* 2020). ML algorithms are also at the mercy of human error as the data used can still be parsed incorrectly by humans or mislabeled, in the case of classification problems. It is also the case that certain classifications and labels might not be relevant in the future as new knowledge or revised perspectives creates

the need to update current ML models. These errors can severely impact how well a ML algorithm performs on a given task (Wei Khor & Yuan Ngiam 2019; Holzinger 2018; Zhou *et al.* 2017). Finally some ML approaches are difficult to interpret as they can come to conclusions in dimensions that humans find difficult to visualise or understand, as dimensions beyond 3 become difficult to visualise. This can make the conclusions drawn my ML approaches opaque and difficult to understand (Wei Khor & Yuan Ngiam 2019; Holzinger 2018; Zhou *et al.* 2017).

4.3.2 Assessment of Machine Learning algorithms performance

In order to compare different ML methods to one another techniques to asses a models performance have to be used. One such technique is to plot a models False Positive Rate (FPR) against the models True Positive Rate (TPR), a so called Receiver Operating Characteristic (ROC) curve to classify any given observation correctly (Ball *et al.* 2004). This curve compares a models classification capabilities to random chance; is the model better at distinguishing two classes than a simple coin toss? The FPR is equal to 1-specificity. Specificity is the proportion of the ”negative” class that are correctly identified as such. A ”negative” class could for example be the proportion of patients who are classified as healthy when a model tries to predict a disease as a ”positive” class. Mathematically specificity is presented as a fraction of the True Negative (TN) counts divided by the sum of False Positive (FP) and TN counts: TN/(TN+FP). TPR is equal to a models sensitivity. Sensitivity is the proportion of observations correctly assigned to the ”positive” class. Sensitivity is represented as the fraction of TP divided by the sum of TN and (TP): TP/(TP+TN). A perfect classifier would be able to classify all ”positive” observations correctly without making any mistakes. A terrible classifier would be no better than a coin toss to classify an observation, that is it has a 50 % chance to classify any given observation correctly (Ball *et al.* 2004).

In order to get a quantifiable metric from the ROC curves, that does not involve manually inspecting two curves against one another which is prone to human error, the Area Under the Curve (AUC) of a ROC curve can be calculated. Using the AUROC, two separate ML methods can be directly compared to each other. The AUROC of a ”perfect” method would be equal to 1 while the AUROC of a poor method would be equal to, or less than, 0.5, since it is no better than a coin toss in assigning the correct class (Ball *et al.* 2004).

An alternative approach to measuring the performance of a ML algorithm is to consider the algorithms Precision Recall (PR) performance. This can be done through plotting an algorithms precision against its recall. Precision is a measurement on how many of the predicted ”positive” observations were truly positive and is defined as the following: TP/(TP+FP). Recall is a measurement on how many total ”positive” observations were identified (Cook & Ramadas 2020). Recall is defined as follows: TP/(TP+FN). A perfect

PR curve would be, much like a mirrored ROC curve, start at the coordinates (1,0) and move straight to (1,1). A model that can perfectly predict the number of positive classes independent of how many positive observations were identified. On the other hand a poor PR curve would be a straight line from (0.5, 0) to (0.5, 1) as this would mean the algorithm guesses on the class label for a given observation even though the algorithm succeeds in identifying more positive observations (Cook & Ramadas 2020).

Much like with Area Under the Curve ROC (AUCROC), a quantifiable metric can be found by computing the Area Under the Curve PR (AUCPR). This metric allows direct comparisons between different kinds of ML approaches. Both AUCPR and AUCROC can be used when comparing ML models performances where one can be used as a second opinion to the other. Such an approach is utilised in this project where the AUCROC and AUCPR will be considered in said order.

4.3.3 Curse of Dimensionality

The microbiome data used in this project is inherently high dimensional. High dimensional data is data that contains more features than observations. In the case of this project, each patient has up to 461 taxa associated to them but only 180 patients were used for analysis; thus the data becomes high dimensional.

The ”curse of dimensionality” is a phenomena where previously thought of truths in data analysis fail to hold up in higher dimensions. Standard statistical methods have been designed to work on 2-or-3 dimensional spaces which humans are able to visualise. These methods are also based on intuition in these dimensions (Verleysen & François 2005; Li 2015). However when more dimensions are added these intuitions fail to hold and it becomes difficult for humans to visualise and understand what kind of effect occurs in dimensions exceeding 3. Furthermore high dimensional data breaks apart given truths in lower dimensional spaces. An example would be a Gaussian distribution, in which a majority of the volume falls to the tails of the distribution as it approaches 30 dimensions. When working with high dimensional data these features have to be considered in order not to draw erroneous conclusions. Preferably these dimensions should be reduced in order to mitigate these problems (Verleysen & François 2005; Li 2015).

4.4 The KOLBIBAKT cohort

The data used for this project is taken from a much larger gastrointestinal cohort available at Center for Translational Microbiome Research (CTMR). Between 2016 and 2019, 2395 random patients at Danderyds Hospital, who were to undergo a colonoscopy, were

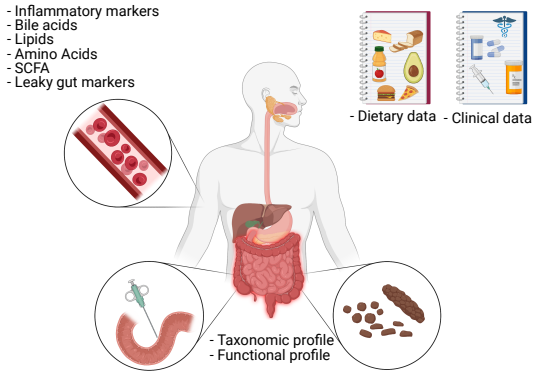

Figure 1: Overview of datasets available from the KOLBIBAKT cohort and from where the datasets are derieved. From the provided blood samples the following datasets were derived: bile acid profile, lipid profile, SCFA profile, amino acid profile, leaky gut marker profile and a lipid profile. From the biopsy and fecal samples both a metagenomic and functional profile were derived. A dietary diary and clinical records are also available. Figure was produced using Biorender

asked if they wished to participate in the study. Out of these 1259 patients agreed to participate. Furthermore 1165 submitted a stool sample, 1253 submitted a biopsy and a blood sample and 1247 completed a questionnaire on health, diet habits as well as medication usage. These results are compiled into the KOBLIBAKT cohort. Each patient was assigned a diagnosis depending on the outcome of their colonoscopy. These diagnoses included: Diverticulosis, Present cancer, Former cancer, IBD, Polyps and Clean colon. Relevant to this projects are the IBD and Clean colon sub-cohorts.

Colonoscopy finding	N	%	Age (SD)	Gender M/F	Abx >3 mon	Any medication	Cortisone	PPI
Diverticulosis	403	32.1	66.2 (9.6)	213/190	61 (15.1)	357 (88.6)	41 (10.2)	111 (27.5)
Present cancer	14	1.1	67.1 (8.1)	9/5	2 (14.3)	13 (92.9)	3 (21.4)	2 (14.3)
Former cancer	97	7.7	68.3 (10.4)	56/41	6 (6.2)	87 (89.7)	12 (12.4)	15 (15.5)
IBD	**279**	**22.2**	**48.8 (16.1)**	**163/116**	**35 (12.5)**	**265 (95.0)**	**74 (26.5)**	**42 (15.1)**
Polyps	595	47.3	65.6 (10.7)	314/281	65 (10.9)	512 (86.1)	62 (10.4)	140 (23.5)
Clean colon	**214**	**17.0**	**56.6 (15.0)**	**94/120**	**40 (18.7)**	**166 (77.6)**	**22 (10.3)**	**23 (10.7)**

Figure 2: Overview of illnesses covered in the KOLBIBAKT cohort and meta data concerning each illness. Note IBD and Clean colon cohorts are highlited

The Clean Colon cohort is a reference cohort, where no infection or illness was found during the colonoscopy. The IBD cohort contains two sub-cohorts: CD and UC cohort. These cohorts can also be further divided into active and remission cohorts. In this project only the CD cohort was considered.

4.5 Microbiome data

4.5.1 Sequencing microbiome samples

There are two general ways in which microbe communities can be sequenced; through a targeted (amplicon) and a shotgun (untargeted) approach. Each approach differs from sequencing to post-processing analysis which makes the decision to choose which method to use a non-trivial one. Amplicon sequencing works by sequencing specific marker gene(s) present in the genomes of interest, this could be the 16S Ribosomal RNA (rRNA) region in prokaryotic and archea for example, which has a highly ubiquitous distribution and is a relatively stable. It codes for the small subunit in prokaryotic ribosomes (J. Sharpton 2014). Results from the sequencing are reads specifically tailored to match sequences of interest. These reads are then used to determine which taxa are present in said sample as well as how abundant each taxa is. Since only a specific portion of the genome is sequenced during this procedure, all other genomic information is lost. This includes genomic reads which could be used to infer biological function making amplicon sequencing less applicable if a functional profile is desired. Another limitation to amplicon sequencing is that the technology often only reaches a genus-level resolution of the present sample, unless several regions of 16S are sequenced (Bai *et al.* 2021; J. Sharpton 2014).

Shotgun sequencing does not limit the sequencing of samples to a specific genetic marker sequence(s). Instead all DNA present in the sample is sequenced, after being sheared, and available for post processing. Through shotgun sequencing both genetic markers and whole genome sequences are available, if sequenced deeply enough, for downstream analysis which, in turn, allows for inference of biological functions available to the sample as well as deeper characterization of the microbiome complexity (J. Sharpton 2014; Laudadio *et al.* 2018). However shotgun sequencing is not without flaws. A few limitations of shotgun sequencing would include:

- Presence of DNA not from microbe community
- More sequencing data required for analysis
- Not trivial to determine from which taxa individual reads originate from

Once sequenced, there are a few processes from which shotgun data can be transformed into taxanomic abundance profiles: binning, marker-based annotation methods, k-mer-based annotation methods and assembly of reads into genomes. Marker-based annotation methods compare generated reads to a reference database where reads are annotated

depending on their similarity to said markers. Markers used can be anything from protein coding genes to rRNA genes, but they need to be specific in order to work (Bai *et al.* 2021). Marker based approaches also limit what can be seen in the sample as novel genomes, without any known markers, would not be detected. Choice of marker database is vital as well since databases curated for a particular kind of source would be more proficient at detecting microbes from said source and less useful when working with microbes from different sources. For example some parts of the human body are more well suited for obligate anaerobes, such as the gut, while other parts are not, such as the skin. Thus if a database derived from skin samples were to be used during the taxanomic profiling of fecal samples, some taxa would not be detected which should be there (J. Sharpton 2014; Laudadio *et al.* 2018; Bai *et al.* 2021).

K-mer methods are another metagenomic approach to produce a taxanomic profile. Kraken is such an approach. Kraken has an internal database that matches k-mers to a Last Common Ancestor (LCA). Once a sample has been sequenced Kraken creates a classification tree with the LCA at the root. Each branch contains a specific taxon, and its ancestors. Each node then has a weight assigned to it which equals to the amount of k-mers that supports said taxon. Each root-to-leaf path is then scored by the cumulative weights of all nodes in said path. The leaf-node from the highest scoring path is then the predicted taxa (Wood & Salzberg 2014). Problems with k-mer approaches are, much like a marker-based approach, that k-mers that do not exist in the software internal database will not be detected. K-mer approaches are also susceptible to over predicting taxa which are not present in the sample (J. Sharpton 2014; Bai *et al.* 2021).

Two other processes used to acquire taxanomic abundance from shotgun sequencing include binning and assembly. Binning comprise of methods grouping reads on their intrinsic features, such as grouping reads based on GC content. Binning allows for reduction in data complexity, allowing further analysis to be performed on bins of interest, and can detect novel genomes. Limitations to binning approaches include limitations surrounding convergent evolution characteristics, such as horizontal gene transfer events, would cause type 1 errors whilst grouping reads. Additionally there exists a trade-off between the amount of reads binned and the taxanomic specificity of each bin (J. Sharpton 2014; Bai *et al.* 2021).

Lastly an assembly approach is available for taxonomic profiling. Assembly approaches work by trying to reconstruct the genomes of all taxa from the underlying sample by merging reads into a single continuous sequence. Performing an assembly of reads could be beneficial for downstream analysis however it is biased towards more abundant taxa present in the sample as these taxa are easier to reconstruct than others (J. Sharpton 2014). Assembly approaches also susceptible to chimeric reads, artificial contigs generated by similarities between different reads from different genomes during Polymerase

Chain reaction (PCR) as they incorrectly align to each other. Users of this approach must also bear in mind to keep track of coverage, amount of reads that align to the average base in the contig, and depth, average number of times a given base has been covered by sequencing, in order to verify the integrity of a final assembled genome (J. Sharpton 2014).

4.5.2 Compositional data and challenges

Compositional data is regarded as data that only displays a portion of a whole due to an arbitrary limit. Microbiome data is such an example; inherent to the sequencing technology applied there is a maximum capacity of reads that can be sequenced, thus the reads generated by a machine are not an absolute representation of the microbes present in a sample but only a glimpse of the whole. In strict mathematical terms; compositional data is defined as a vector with an uninformative sum and with strictly positive real numbers (Gloor *et al.* 2017):

$$x = (x_1, x_2, .., x_n); x_i > 0, i \in \{1, .., n\} \tag{1}$$

These limitations reflect a Dirichlet distribution which is difficult to work with; Dirichlet fail to model variability of data within said distribution. Another detrimental aspect is the distributions built-in interdependence to its definition that it becomes an inconvenient class to model compositional data. Due to these limitations of working within a Dirichlet distribution, standard methods in statistical analysis can not be used with compositional data (Gloor *et al.* 2017; Calle 2019).

In order to make compositional data more applicable to standard statistical methodologies it has to be transformed. Ratio transformations allow compositional data to be used in standard statistical models as it captures relationships between the features in a dataset and breaks free from the Dirichlet distribution into a real-space distribution(Gloor *et al.* 2017).

There are different transformations available, each more or less suited for a given problem. Below, 4 example logratio transformations are presented, which at one time or another were relevant to this project (Tolosana-Delgado *et al.* 2019), the Centered Log-Ratio, Pairwise Log-Ratio, Isometric Log-Ratio and the Addative Log-Ratio, each trying to ameliorate said limitation of a Dirichlet distribution:

- clr(x) = $\ln[\frac{x_1}{x_D}, \frac{x_2}{x_D}, ..., \frac{x_{D-1}}{x_D}]$, $g(x) = \sqrt[D]{x_1 \cdot x_2 \cdot ... \cdot x_D}$
- pwlr(x) = $[\xi_{i,j}; i < j = 1, 2, ..., D]$, $\xi_{i,j} = \ln x^*_{i,j}$, $x^*_{i,j} = \frac{x_i}{x_j}$
- ilr(x) = $clr(x) * V$, $V^t * V = I_{D-1}$

- alr(x) = $\ln[\frac{x_1}{x_D}, \frac{x_2}{x_D}, ..., \frac{x_{D-1}}{x_D}]$

Ratio transformations alleviate problems caused by compositional data. However, it is vital to keep in mind that transformations only make the data easier to work with but do not fundamentally change the data from being compositional to absolute, and thus limiting which conclusions can be drawn from the data.

4.5.3 Covariates analysis

Before using abundance data for the purpose of training ML algorithms, covariates have to be determined. If this is not performed further analysis will be biased by said covariates. Covariate determination was performed by performing a PERMANOVA test. PERMANOVA is a non-parametric multivariate analysis of variance test. The test determines how much a given measurement of the data influences the variation in the data (Anderson 2017). PERMANOVA has been previously used to find statistically significant covariates as well as their effect on the data set in other microbiome studies (Fazlollahi *et al.* 2018; Al Alam *et al.* 2020). The dissimilarity measure chosen for the PERMANOVA analysis was the Bray-Curtis dissimilarity. The Bray-Curtis dissimilarity can be expressed as follows (Calle 2019):

$$d(p_1, p_2) = \frac{\sum_{n=1}^{k} |p_{1i} - p_{2i}|}{\sum_{n=1}^{k} |p_{1i} + p_{2i}|}$$

Where $p_1 = (p_{11}, \ldots, p_{1k})$ and $p_2 = (p_{21}, \ldots, p_{2k})$ denote the relative abundance of species between two different sites. The Bray-Curtis is a common dissimilarity measure and it is a measure on how different species are between two different sites. It scales from 0, the sites are identical, and 1, the sites are completely different (Calle 2019).

5 Methods and materials

The data for this project is derived from the KOLBIBAKT gastrointestinal cohort. Each patient also provided a detailed dietary diary, a stool sample collected prior to bowel preparation for the colonoscopy and filled in a questionnaire concerning information about their lifestyle. Medical records and clinical data were also available, detailing other medical conditions and medication usage. This provided us with large amounts of metadata for each patient. All patient's stool samples were sequenced using MGI

DNBSEQ-T7 machines with a minimum read length setting of 100 base pairs and a with the mean number of reads for all samples being 102.1386 Million reads.

After sequencing all samples were demultiplexed using an in-house script and further processed using Stag-mwc, a pipeline written in SnakeMake, developed by the CTMR to analyse metagenomic samples (Boulund 2018). Stag-mwc performs adapter removal and read quality filtering using fastp. Reads aligning to the human genome (version) using Kraken2 (Wood *et al.* 2019) are removed. STAG-mwc can perform other kinds of data processing, such as binning, but relevant to this project are its ability to produce taxanomic and functional profiles from raw reads, using MetaPhlAn3 or Kraken2. For this project it was decided upon to use MetaPhlAn3 (Beghini *et al.* 2021) for the taxanomic profiling after reviewing a benchmarking paper (Ye *et al.* 2019). MetaPhlan3 is a marker-based method while Krakren2 uses a k-mer approach to determine relative abundance. Kraken2 tended to overpredict the pressence of taxa as a function of sequencing depth, i.e. the more you sequenced the more taxa would be predicted, even though certain taxons were not present in the original sample. MetaPhlAn3 did not have this problem, although it was limited to its database. Another benefit to MetaPhlAn3 was the possibility to use HUMANn3 to produce a functional profile.

MetaPhlAn3 used the ”mpa_v30_CHOCOPhlAn_201901” database and was run with the following commands ”–unknown_estimation –index latest”. All other parameters were left as default. HUMANn3 had the following settings: ”community” as the normalization mode and ”cpm” as the normalization scheme. The following database was used for the HUMANn3 profiling: ”uniref90_201901b_full.dmnd”. All other parameters were left as default

Before performing any form of analyses on the data it was filtered on two criteria. In order for a species to be kept for further analysis it had to:

- Be present in at least two samples
- Have an abundance of more than 0.0001 %

These parameters reduced the amount of taxa present from 845 to 461. It was discovered that more stringent filtering, either increasing abundance requirements or presence in number of samples or both, only resulted in removing marginal amounts of taxa. Thus no more stringent parameters were chosen as it would have diminishing returns in terms of filtering away redundant abundances.

Covariate analysis was performed using the adonis function in the ”Vegan” R-package, version 2.5-7 using a Bray-Curtis dissimilarity matrix. The ALR transformation was

created from the ”alr()” method using the R-package ”Compositions” v. 2.0-4. CLR transformation was performed with the package MixOmics v 6.18.1. All ML algorithms were implemented using the R-package caret v. 6.0-91. The following ML algorithms were called:

- ”rf”: A random forest algoritm
- ”LR”: A Linear Regression algorithm
- ”knn”: A K-nearest-neighbor algorithm
- ”svmRadial”: A support vector machine algorithm
- ”pcaNNet”: A Neural Network algorithm

A balanced dataset, 90 healthy individuals and 90 ill individuals, was split randomly into either a train set and a test set. The data set came from the CD sub-cohort. 70 % of the data became the training set and 30 % of the data was used for the test set. The 30 % would later be used to validate the models performances on novel data. The pipeline ran using both an ALR and CLR transformation of the microbiome data. Splitting of the data was performed after a transformation. The models used a Leave-One-Out Cross-Validation (LOOCV) during their training in order for the models to be tuned during their training. Each transformation output was compared to one another. Which ever model had the highest AUCROC score was used for further testing.

Once a satisfactory model had been found it was selected for further tuning. The ”rf” implementation of the random forest algorithm was discovered to be satisfactory in this case; outperforming all other algorithms on the test set data. The ”rf” model was then further tuned by adjusting the following parameters until a model with the highest AUCROC value had been found:

- Test training data split: 70/30, 75/25, 80/20, 85/15 were tested
- Number of trees: 2501, 5001, 7501, 10001, 12501 were tested
- Node sizes: 1, 3, 5, 10, 15 were tested
- Sample sizes: 1, 3, 5, 10, 15 were tested
- Random and Grid search settings were used to tune the ”mtry” parameter.

Each iteration was tested using repeated CV, 10 repeats and a K value which divides each fold to an equivalent size of the test set. Furthermore the top 54 taxa were chosen for further work, in order to reduce the dimensions of the data. This was done without performing a new transformation; the most significant features were extracted from an already transformed dataframe. These results are then later compared to contemporary literature, to see if any novel taxa have been discovered by said algorithms.

6 Results

6.1 Covariate analysis

Figure 3 presents the most significant covariates which PERMANOVA identified. It is clear that the diagnosis of a patient has a large impact on the observed variations between healthy and sick individuals. A cut of point was made at ca 1/3 of the R2 score of diagnosis in order to reduce the number of covariates considered. After discussion with clinical partners it was determined that the only confounding effect present in the data was age, and thus this factor has to be matched before using ML algorithms. All other factors were related to or consequences of an IBD diagnosis and would be expected to rank highly. These factors include Diagnosis, Medscore, Bristol_categories, Childhood_environment, Education and Years_of_IBD. Lane_ID is associated with which plate on the MGI machine the samples were sequenced on. These results do show a batch effect but it is not significant and when investigated further it is not seen, suggesting that it can be disregarded. When looking into the positive controls present on each plate no mishaps were noticed, the machine managed to sequence all positive controls correctly. Furthermore a NMDS plot was performed on all samples, accounting for their sequencing plates. No outliers were detected and there was a great amount of overlap between all samples suggesting a lack of batch effect. Finally a chi-square test was performed which concluded that there was no statistically significant correlation between the Lane_ID and diagnosis. between the plate Please refer to the appendix ”Analysis of Batch Effect” for plots and tables.

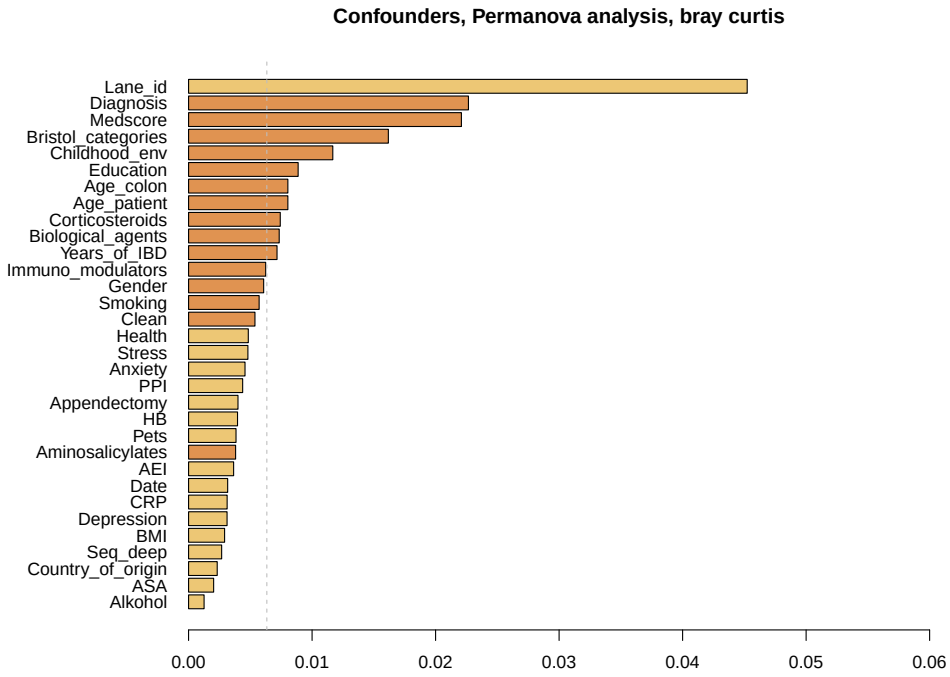

Figure 3: Results of PERMANOVA analysis. On the Y-axis covariates of interest are presented. On the X-axis the R2 value is displayed, how much each covariate explains the variance in the data. Significant covariates, $p < 0.05$, are displayed in orange. Non-significant covariets displayed in dark yellow.

6.2 Spearman results

The Spearman test showed clear signs of inter dependencies between a majority of the taxa. From these results it could be inferred that uni variate statistical methods will not be applicable and instead multivariate methods, such as machine learning must be applied. Refer to Figure 12 for a larger heatmap of the results.

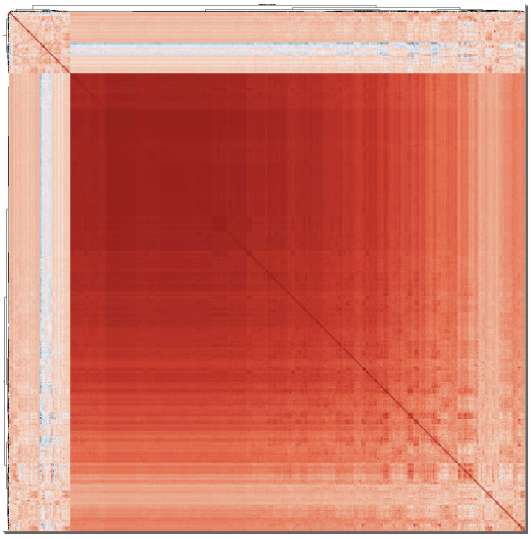

Figure 4: Spearman analysis heatmap. Red signifies a positive correlation while blue signifies a negative correlation. In each axis all 461 taxa are compared against one another.

6.3 Machine learning algorithms performance

Initial performance showed that the transformation of compositional data had a small effect on the performance of algorithms trained on the different data. Figure 5 shows three barplots of all five algorithms results after being run on either a ALR or a CLR transformed dataset, and their respective AUCROC, sensitivity and specificity scores. For further work the CLR was chosen as it was easier to interpret and had the same effect as ALR. It became clear that the RF algorithm performed better than the other algorithms on the test data. Due to this observation it was decided upon to further tune the RF algorithm to increase its classification ability.

Algorithm	CLR	ALR
SVM	0.62	0.62
PLR	0.59	0.62
NN	0.57	0.58
KNN	0.54	0.51
RF	**0.78**	**0.86**

Table 1: Resulting AUCROC values for each algorithm from a CLR and ALR transformation. It becomes apparent that RF method outperforms all others. This test was performed on the training dataset.

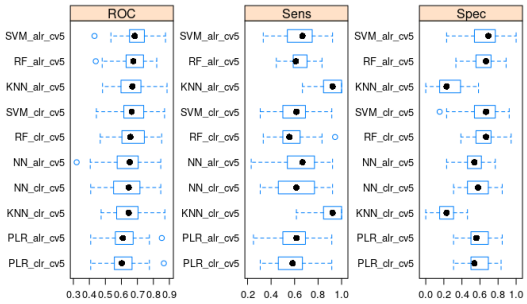

Figure 5: ALR and CLR results compared to each other. Note these results are from the train set and using K-fold cross validation as an example.

After both transformations had been assessed, the most important taxa for each model was assessed to determine if certain taxa could be removed from further analysis. Removing taxa would reduce the dimensionality of the data as well as improve the speed of the RF algorithm. Importance is measured by scaling the prediction accuracy when a feature is removed from sampling when producing a RF classifier. If a feature has high importance it has a larger impact on the classification ability of the model than a feature with low importance. Features below and at an importance threshold of 40 were removed. This threshold was decided upon since there was asymptote around the importance value of 40, meaning a lot of features had a relatively similar effect on the classification ability of the RF classifier. After filtering away taxa at or below the importance threshold of 40, the ALR transformation had 62 features remaining while the CLR transformation had 68 features remaining. To further reduce the dimensionality of the data, the intersecting taxa between both transformation was used to produce the final RF model, as this decreased the number of features to only 54.

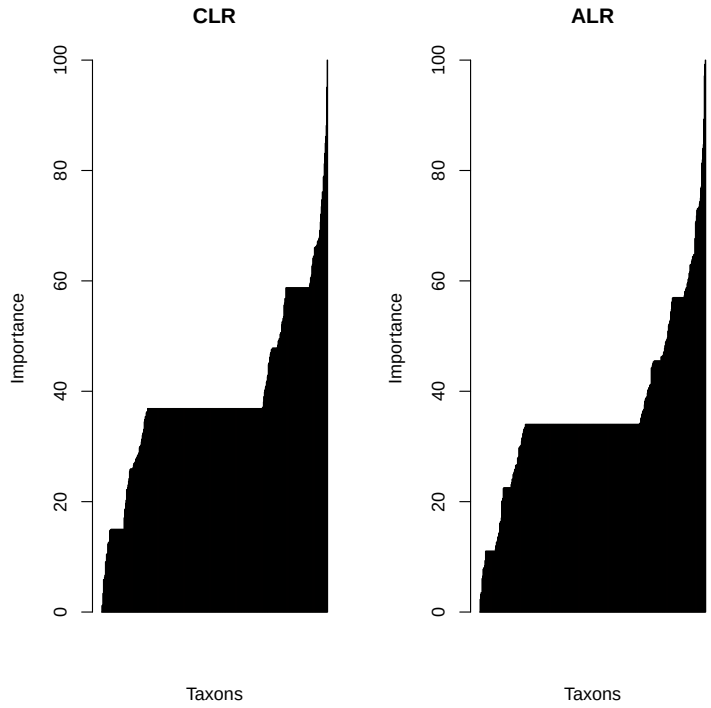

Figure 6: ALR and CLR importance compared to each other. Note in both plots there is an asymptote around the importance value of 40.

6.3.1 Random Forest Performance

After testing, the following values for each hyper parameter created a RF model with the best performance ability: node size: 3, sample size: 10, data split: 85% train, 15% test, K-fold value of 6, 54 out of the initial 461 taxa kept and an mtry value of 52.

Train/Test split	K-value	Number of trees	Node Size	Sample Size
70/30	2	2501	1	1
75/25	3	5001	**3**	3
80/20	4	7501	5	5
85/15	**6**	10001	10	**10**
-	-	12501	15	15

Table 2: A table summarising the different kinds of hyperparameters tested during tuning. Values which resulted in the best performance are highlighted. The 1st column displays which data split was used for the training and testing respectively. 2nd column displays which K value was used during the K-fold cross validation. 3rd column displays the number of trees present in the ensemble. 4th column displays the maximum size of the nodes at each decision tree. 5th column displays how many samples were used to construct each tree in the ensemble.

The number of trees had little impact on the classification performance. The parameters which had a larger impact were the node sizes, the split of test and training data, sample sizes and which K value was used for cross validation. For more information please refer to appendix 6.3 for further information. These optimal parameters resulted in an RF classifier that had an AUCROC value of 0.82 on testing data and a AUCPR value of 0.79 on testing data. These curves are displayed in Figure 7, and Figure 8 respectively.

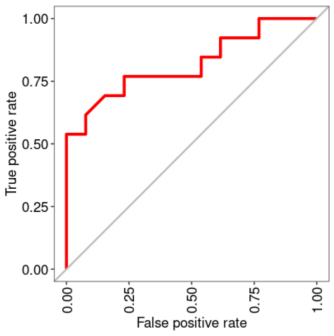

Figure 7: ROC curve. AUC is 0.82

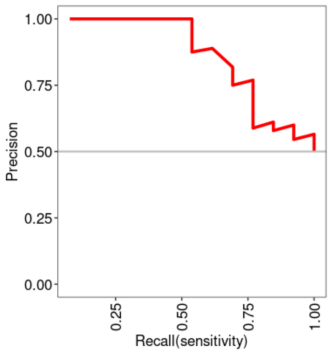

Figure 8: PR curve. AUC is 0.79

6.4 Taxa associated with IBD

In Figure 9 the most important taxa are ranked in a descending order. These results are from the optimal RF algorithm model presented in the section above ”Random Forest Performance”. It is not apparent from these results which taxa is associated with which

class, i.e. is a lower abundance of *Akkermansia muciniphila* associated with a healthy or a sick individual? This has to be verified. For further discussion, the top 5 taxa will be considered: *Ruminococcaceae bacterium* D5, *Akkermansia muciniphila*, *Streptococcus parasanguinis*, *Flavonifractor plautii* and *Bifidobacterium bifidum*.

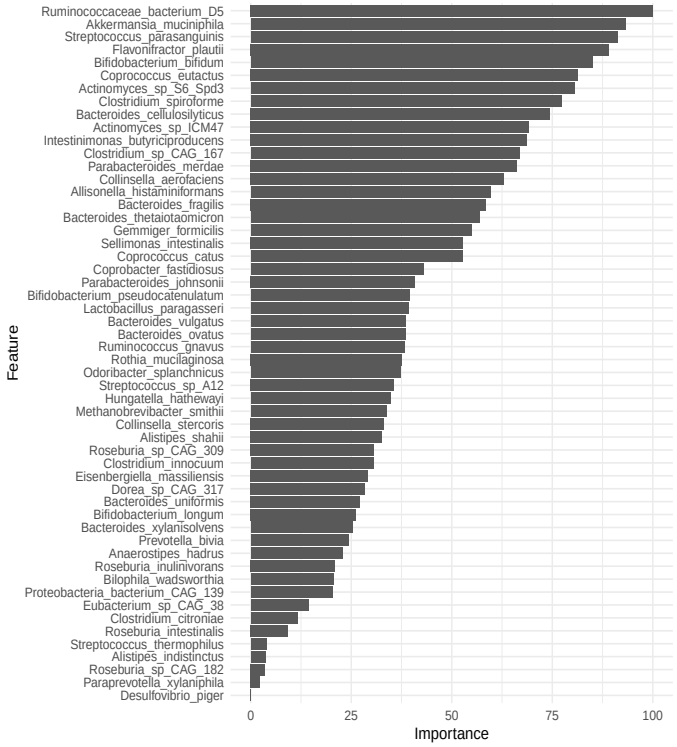

Figure 9: Taxa ranked by importance from the best performing ML algorithm.

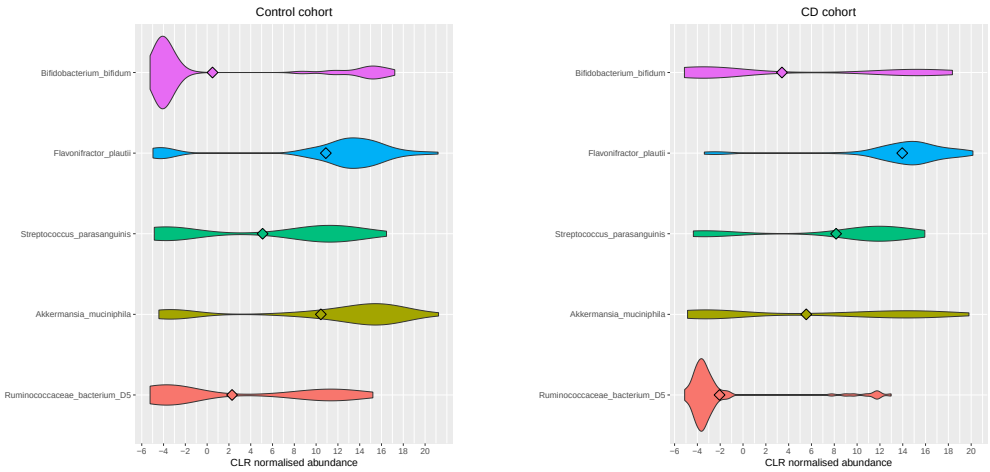

Figure 10: The relative abundances of the top 5 taxa from the sick and healthy cohort respectfully. Mean relative-abundance is displayed as a small diamond in each plot. From these plots it is visible that B. bifidum is more present in the CD cohort compare to the control (healthy) cohort. This is true for F. plautii and S. parasanguinis. R. bacterium D5 and A. muciniphila abundances are reduced in the CD cohort while having more presence in the healthy cohort.

7 Discussion

7.1 Choice of Compositional Data Transformation

A prerequisite to work with compositional data is to perform a logratio transformation of it. CLR and ALR transformations were chosen as these were already established in the field of microbiome research (Gloor *et al.* 2017). Tolosana-Delgado et al. in their Conference Paper ”On machine learning algorithms and compositional data” presented PWLRs as the best performing transformation in the application of Random Forest ML techniques, only tied with a combination of PWLRs, raw data, ILRs, and CLRs (Tolosana-Delgado *et al.* 2019). While this is a promising case for PWLRs it has to be noted that only Random Forest classifiers were used in the assessment and that other transformations did perform amicably for an increase in size of dataset used in training. When practically working with PWLR transformed data in R it became apparent that the dataset became to large for R to work with. From 460 features to 106,030 features, when training algorithms on this dataset Stack Overflow became a common problem and the dimensionality of the data was substantially increased. No convenient solution to this problem could be found in a timely manner, thus this transformation was not assessed.

ILR transformation has been argued for being the theoretically correct way to handle compositional data analysis as they have mathematically interesting features, such as forming orthonormal basis of the compositional data vectors (Greenacre & Grunsky 2019). However a key fault of ILR is that, while it has mathematically interesting properties, it is hard to practically evaluate what it means to apply such a transformation as alternative transformations do exist which manage to capture similar trends as ILRs but are easier to interpret. Furthermore ILRs can present erroneous correlations in data; either correlations that do not exist or correlations that should not exist (Greenacre & Grunsky 2019).

Due to the fact that the CLR has been recommended by microbiome literature (Gloor *et al.* 2017) and that the ALR was simpler to implement and to understand, compared to PWLR and ILR respectively, these transformations were chosen for benchmarking.

7.2 Relevant Taxa

Clooney et al. in 2021 performed a similar study to this project where the objective was to investigate how different lifestyle factors and environmental factors impacted the compositionality of the gut microbiome in IBD patients (Clooney *et al.* 2021). Clooney et al. study cohort had a total of 303 CD patients, 228 UC patients and 161 controls in it. Each patient had been diagnosed by conventional and investigative criteria and provide a sample at 3 time points ca 16 weeks apart. 283 of the patients came from Cork, Ireland while the remaining 409 were from Manitoba, Canada (Clooney *et al.* 2021). Clooney et al. further used PERMANOVA to determine the effect of external factors, such as geography drug usage and diet, on of the fecal microbiota composition. In their methodology Clooney et al. used 16s sequencing and applied the ”XGBoost” algorithm to classify patients into either healthy or sick. Comparing the taxa, which they concluded was associated with CD specifically, to the results produced by the RF algorithm there were a total of 18 taxa which intersected. These are displayed in table 3.

Taxa	RF Importance	Abundance
Flavonifractor plautii	89.10069	Increased in CD
Bifidobacterium bifidum	84.97957	Increased in CD
Coprococcus eutactus	81.38829	Decreased in CD
Clostridium spiroforme	77.29859	Decreased in CD
Intestinimonas butyriciproducens	68.75818	Increased in CD
Collinsella aerofaciens	63.00948	Increased in CD
Bacteroides fragilis	58.49208	Increased in CD
Gemmiger formicilis	54.84827	Decreased in CD
Coprococcus catus	52.61209	Decreased in CD
Bacteroides ovatus	38.48099	Increased in CD
Ruminococcus gnavus	38.21020	Increased in CD
Rothia mucilaginosa	37.64197	Decreased in CD
Alistipes shahii	32.69517	Increased in CD
Clostridium innocuum	30.52067	Increased in CD
Bifidobacterium longum	26.21996	Increased in CD
Bacteroides xylanisolvens	25.24975	Decreased in CD
Roseburia inulinivorans	20.96199	Increased in CD
Bilophila wadsworthia	20.62030	Increased in CD

Table 3: Table summarising results from Cloony et al. and taxa which the optimal RF algorithm predicted was important. In the first column the taxons name is displayed. In the second column the importance from the RF algorithm is shown. In the third column it is showed whether presence of said taxon is associated with an increased risk of developing CD or a decreased risk, based on Clooney et al. results. Clooney et al. determined that a given taxon is associated with CD by using a metagonmic seq analysis to determine which taxa are deferentially abundant between the two classes.

Clooney et al. found a total of 101 species to be associated with CD (Clooney *et al.* 2021); either increasing or decreasing the odds of developing CD. While these results do not mirror those of this report entirely, only 54 taxa were deemed to be associated with CD and only 18 taxa were identified to be associated with CD in both reports, this does show credence to the methodology applied in this project; by applying ML methods to microbiome data while taking into account covariates and confounders which could bias the results. Differences in the importance of taxa between this report and Clooney et al. could attributed to multiple sources. These sources would include:

- Different sequencing technologies used: Illumina technologies compared to MGI technologies
- Different sequencing technique used: amplicon sequencing compared to shotgun sequencing
- Different ML algorithm used: XGBoost compared to RF
- Different cohort and study design

A review by Lacroix et al. in 2021 was performed to characterize the taxonomic composition of patients suffering from CD. In this review Lacroix et al. identified a total of 12 taxa associated with CD, these results are presented in the table 4 (Lacroix *et al.* 2021):

Taxa	Abundance	After FS	Before FS
Faecalibacterium prausnitzii	Decreased in CD		✓
Bifidobacterium adolescentis	Decreased in CD		✓
Dialister invisus	Decreased in CD		✓
Roseburia inulinivorans	Decreased in CD	✓	✓
Clostridium XIVa	Decreased in CD		
Ruminococcus torques	Decreased in CD		✓
Clostridium lavalense	Decreased in CD		✓
Bacteroides uniformis	Decreased in CD	✓	✓
Clostridium coccoides	Decreased in CD		
Clostridium leptum	Decreased in CD		✓
Escherichia coli	Increased in CD		✓
Ruminococcus gnavus	Increased in CD	✓	✓

Table 4: Table summarising results from Lacroix et al. and how their results compare to this project. In the first column the name of the taxa is displayed. In the second column Lacroix et al. results are displayed showing if the presence of said taxa increases or decreases for a CD diagnosis. In the third column it is displayed if said taxa made it past the feature selection step performed in this project. The final fourth column shows if said taxa was present in the initial 461 taxa used in this project. FS: Feature Selection.

Three of the taxa Lacroix et al. identified are in agreement with the results of this project. These are *R. inulinivorans*, *R. gnavus* and *B. uniformis*. 7 more were present in the sample used for the initial RF runs but were not deemed important enough for further classification as they were equal to or below the importance threshold set up. *R. inulinivorans*

has an importance of 13.048814 while *R. gnavus* and *B. uniformis* have importance values of 17.694902 and 20.440685 respectively. These importance values suggest that, while these taxa were not filtered away in the initial feature selection, these taxa are not particularly important in the RF models ability to discriminate between CD and healthy observations.

R. gnavus has been previously characterised to have a ”robust association” with CD (Henke *et al.* 2019). *R. gnavus* is a gram positive, anerobic commensal taxa that usually has a relative abundance of ca 0.1%. During sever flares of CD this abundance can increase up to 69%. *R. gnavus* not only produces an inflammatory glucorhamnan polysaccharide that in turn cause dendritic cells to produce TNF α, which is an inflammatory cytokine, but also colonises the mucosal layer of the epithelium. Here it uses sialic acid from mucin glycans as a carbon source thus damaging gut-barrier functionality (Henke *et al.* 2019).

The taxa within the *Bacteroides* genus are known for their probiotic capabilities. These taxa are bile resistant, gram negative anerobic commensal taxa. Members with probiotic effects include *B. fragilis*, which produces polysaccharide A that play a role in activating T cells, and *B. acidifaciens*, which aids in protection against obesity by promoting activation within certain metabolomic pathways. *B. uniformis* has been know to alleviate immunological dysfunctions and metabolic disorders related to obese mice. However no exact mechanism has been proposed (Dahiya *et al.* 2019).

R. inulinivorans is a gram negative anaerobic taxon which is a part of the *Roseburia* genus (KELLERMAYER 2019). This genus is associated with a healthy gut microbiome as members are capable to produce not only butyrate but also propinate, a different kind of SCFA. Presence of *R. inulinivorans* has been associated with remission for patients with CD and UC as well as regulating cell cycle control, perhaps performing some form of tumor suppresion. Furthermore *R. inulinivorans*, when patients are treated with fecal microbiota transplantation, has been positively correlated with a successful outcome of a fecal microbiota transplantation (KELLERMAYER 2019).

The top 5 most important taxa predicted by the random forest algorithm were: *Ruminococcaceae bacterium* D5, *Akkermansia muciniphila*, *Streptococcus parasanguinis*, *Flavonifractor plautii* and *Bifidobacterium bifidum*.

The genus *Ruminococcaceae* are gram-positive, anaerobic microbes whose members an be found as commensal bacteria in the gut. It has been previously identified that *Ruminococcaceae* are less abundant in patients suffering from IBD (Lo Presti *et al.* 2019). Due to their ability to produce SCFAs, such as butyrate, it has been hypothesised that *Ruminococcaceae* have a protective effect against inflammation. As mentioned before

SCFAs have a multitude of different health benefits to the gut, such as being an energy source to the epithelial cells (Maukonen *et al.* 2015; Lo Presti *et al.* 2019).

A. muciniphila is a commensal, gram-negative anaerobic bacteria that has been previously identified to have inverse correlation to a multitude of illnesses relative to its abundance. These illnesses include obesity, diabetes, inflammation and other metabolic disorders where lower abundances of the taxon has been observed. *A. muciniphila* has been described as common in the human gut, with it representing ca 3-5% of the microbial community, and the exact mechanisms behind *A. muciniphila* health effect is currently unknown. A hypothesis is that *A. muciniphila* helps regulate the mucus layer around epithelial cells, making it harder for noxious agents and pathogens from reaching the epithelial cells. Ironically *A. muciniphila* uses mucin, a key component of the mucus layer, as its primary energy source. By digesting older mucin *A. muciniphila* frees up SCFA for the synthesis of new mucus, thus decreasing the turnover rate of new mucus production. In turn abundance of *A. muciniphila* is correlated by a lower presence of serum LPS, an indicator of gut permeability. Additionally presence of *Bifidobacterium animalis* as also been correlated with an increase of *A. muciniphila* in fecal matter in mouse models (Zhou 2017).

S. parasanguinis is a gram-positive bacterium that is commensal in the oral microbiome. It is associated with a healthy oral microbiome as it protects against caries and periodontopathogens by producing hydrogen peroxide (Chen *et al.* 2019; Corby *et al.* 2005). It has also been identified to be an early colonizer of the human gut, as it is present in breast milk, where it tunes and matures the infants immune system (Chen *et al.* 2019). However presence of *S. parasanguinis* in the gut has also been associated with cardiovascular disease, such as unstable angina, and promotion of oral cavities, suggesting *S. parasanguinis* could behave as a opportunistic pathogen (Liu *et al.* 2022).

F. plautii is a anaerobic, rod-shaped gram positive bacteria. Due to the fact that *F. plautii* is difficult to isolate there is currently not a lot of data concerning its clinical significance (Berger *et al.* 2018). It has been reported that *F. plautii* facilitates the metabolism of catechins, a set of antioxidants, in the human gut. A study by Mikami et al. showed that in mouse models, which were exposed to catechins orally, an increased intake of catechins correlated with an increased abundance of *F. plautii* in the stool of the mice. Moreover, mice which expressed IBD-symptoms had their symptoms alleviated once they were fed catechins. This suggests that *F. plautii* has a protective element against IBD. It has been hypothesised that *F. plautii* has a multitude of abilities that gives it a protective role against IBD. The taxon can produce butyrate, a for of SCFA, which in turn can be utilised by other taxa or used in the production of new mucus for the gut. *F. plautii* mediates IL-17, a pro-inflammatory cytokine, which can damage the mucosal layer of the gut if over expressed (Mikami *et al.* 2021).

B. bifidum is a commensal, rod-shaped, gram positive bacteria that is highly abundant in the gut in infancy but decreases while aging. *B. bifidum* has been described as tuning the immune system of infants by promoting a pro-inflammatory response while down regulating certain cytokines and other elements. Once in adulthood *B. bifidum* acts in a similar fashion to *A. muciniphila*, by using mucin as its primary energy source, thus promoting production of more mucus making the epithelium thicker, increasing its protective ability. In mouse models, *B. bifidum* has been shown to have a plethora of probiotic effects. In mouse models *B. bifidum* has been shown to down regulate certain families of miRNA which are linked with colitis. Moreover *B. bifidum* also alliveate the severity of colitis in mice and can do so post hoc by regulating the expression of pro-inflammatory cytokines such as IL-1β as well as increasing the colon length, a symptom of colitis (Din *et al.* 2020; Turroni *et al.* 2014).

7.3 Possible improvements

Feature selection, also known as variable elimination, is a data wrangling technique in which features are screened before using machine learning algorithms in order to reduce the dimensionality of the data used as well as finding features which improve prediction performance. Reducing features also improves computation time as well as reducing the noise of the data (Chandrashekar & Sahin 2014).

The feature selection that was applied in the current project was the following:

1. Run all algorithms
2. Select algorithm that has performed the best
3. Rank all features taxa in descending order and select those features above an asymptotic threshold
4. Use said features for further analysis with the best performing algorithm

While straight forward and easy to implement no comparisons were made to other alternative approaches. Performing a filtration step before classification before could be an approach; ranking each feature by utilising methods such as Pearson correlation coefficient tests in order to filter away less informative features could be an alternative approach. Alternatively wrapper methods could be used. These methods attempt to find the best features by suing optimisation algorithms, such as particle swarm optimisation, after training a method. Features that succeed in reducing a models objective function

are then selected as being more informative ones (Chandrashekar & Sahin 2014). There are also stochastic ML feature selection methodologies available which have also proven to be good in discriminating useful features from less informative ones. MCFS is such an method, which bases a features importance depending on how well multiple randomly constructed decision trees perform when said feature is present in them (Dramiński *et al.* 2007). Feature selection methodologies should be investigated in future applications as it could alleviate the curse of dimensionality.

Applying ”clear-box” ML algorithms would be another approach to this project; applying algorithms in which it becomes apparent how the algorithm ”thinks” in order to come to its conclusions. All approaches in this paper were ”black-box”; as users of these models, it can not be said how the algorithms came to their conclusions, only that they did. ”Clear-box” approaches might lend more insight as to how different taxa might correlate to each other, perhaps an increase in abundance of one taxon could inversely relate to the abundance of another and thus be negatively correlated with IBD.

It must be mentioned that only prokaryotes were considered in this project. No attempt was made to characterise either fungal or viruses that are associated with Crohn’s Disease. These microbes could have interactions between the presented prokaryotes in this project.

7.4 Future work

These results are specifically for the CD form of IBD and do not take into account UC. Future work will need to be performed on the UC cohort in order to determine which taxa are associated with UC. Furthermore, patients with remission should be compared to controls, for both the UC and CD cohort, in order to infer whether or not there are taxa associated with IBD remission. No distinction was made in this work between patients with remission or active CD.

Given the highly personalized nature of the gut microbiome, it could be the case that viewing CD through a taxonomic lens severely limits which conclusions can be made about the disease, as some taxa could share similar roles in the biome of the gut. This would not be apparent by looking at a taxonomic profile alone. Casimiro-Soriguer et al. used an explainable AI model to predict colorectal cancer in patients. In their work they discovered that by using a functional profile of their data, instead of a taxonomic profile, the machine could distinguish between colorectal cancer and adenoma, a form of precursor to a tumor. They concluded that their results suggest that what is changing in the microbiomes functional profile is more indicative of tumor formation than the taxa

present. The functional profile was also more interpretable according to the authors since ”features are informative by themselves”. Taxanomic features have to be determined posteriori while functional features by themselves are simpler to interpret (Casimiro-Soriguer *et al.* 2022). These results suggest credence to Loyd-Price et al. hypothesis that a core of functions available in a microbiome is a more correct manner to describe a healthy microbiome, rather than the taxa present (Loyd-Price *et al.* 2016). Analysing the functional profile of both CD and UC could perhaps gain more insights to what kind of pathways in the microbiome could be associated with CD and UC. Perhaps, in the long future, these functional profiles could be used as targets for a more personalized healthcare approach.

Since IBD is a complicated disease, with many environmental, genetic and biological factors at play, it is relevant to try and contextualise these results with other data sets available from the KOLBIBAKT cohort. Relevant data sets to compare to could be perhaps the Olink inflammation panel results or SCFAs analysis results. Combined these results might reveal more indepth and detailed associations between IBD and the multitude of factors which can cause it.

Finally it can be said that investigating ensemble ML techniques could create models with better discriminatory ability than the current RF model. Ensemble ML techniques work by combining multiple ML techniques into one classifier. For example an ensemble technique could be the combination of a NN classifier, a RF classifier and a KNN classifier. Each classifier is trained on the same data and each classifier ”votes” on which class a new observation would belong to (Ardabili *et al.* 2020; Hosni *et al.* 2019). This is the general methodology of an ensemble algorithm. These algorithms can outperform standard ML approaches using only a single ML algorithm, which is the intended goal of an ensemble approach, however ensemble techniques do take longer time to train than a single model (Ganaie *et al.* 2021).

8 Acknowledgements

I would like to wholeheartedly thank my supervisor, Stefanie Prast-Nielsen for this unique and wonderful opportunity as well as her guidance during this project. I would also wish to thank Sergi Sayols who provided insightful commentary of my work and conclusions. Finally I would like to thank the CTMR bioinformatics team for aiding me through this project.

9 Appendix

9.1 Batch effect analysis

P-value	Statistic	Groups
0.615	10.989	CD-rem vs CC
0.447	13.002	CD-act vs CC
0.426	13.284	UC-rem vs CC
0.478	12.612	UC-act vs CC

Table 5: Chi-square test results comparing the Lane_ID to the different diagnoses in the IBD cohort. Note that no relationship is considered significant.

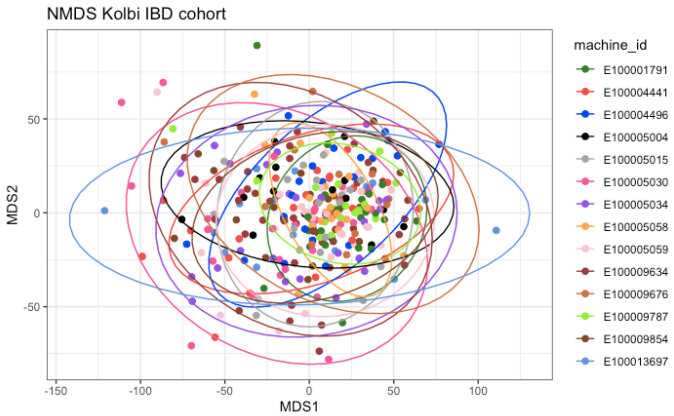

Figure 11: NMDS plot of IBD cohort with 95% confidence interval displayed as as ellipses. Color refers to which plate the samples were sequenced on- Note the large amount of overlap between the plates, suggesting that all plates are very similar.

9.2 Hyper parameter tuning results

"Hyperparameters"	"AUC-ROC"	"AUC-PR"
"ntrees: 2501 nodesize: 1 sampSize 1"	0.58	0.56
"ntrees: 5001 nodesize: 1 sampSize 1"	0.58	0.56
"ntrees: 7501 nodesize: 1 sampSize 1"	0.58	0.56
"ntrees: 10001 nodesize: 1 sampSize 1"	0.58	0.56
"ntrees: 12501 nodesize: 1 sampSize 1"	0.58	0.56
"ntrees: 2501 nodesize: 3 sampSize 1"	0.58	0.56
"ntrees: 5001 nodesize: 3 sampSize 1"	0.58	0.56
"ntrees: 7501 nodesize: 3 sampSize 1"	0.58	0.56
"ntrees: 10001 nodesize: 3 sampSize 1"	0.58	0.56
"ntrees: 12501 nodesize: 3 sampSize 1"	0.58	0.56
"ntrees: 2501 nodesize: 5 sampSize 1"	0.58	0.56
"ntrees: 5001 nodesize: 5 sampSize 1"	0.58	0.56
"ntrees: 7501 nodesize: 5 sampSize 1"	0.58	0.56
"ntrees: 10001 nodesize: 5 sampSize 1"	0.58	0.56
"ntrees: 12501 nodesize: 5 sampSize 1"	0.58	0.56
"ntrees: 2501 nodesize: 10 sampSize 1"	0.58	0.56
"ntrees: 5001 nodesize: 10 sampSize 1"	0.58	0.56
"ntrees: 7501 nodesize: 10 sampSize 1"	0.58	0.56
"ntrees: 10001 nodesize: 10 sampSize 1"	0.58	0.56
"ntrees: 12501 nodesize: 10 sampSize 1"	0.58	0.56
"ntrees: 2501 nodesize: 15 sampSize 1"	0.58	0.56
"ntrees: 5001 nodesize: 15 sampSize 1"	0.58	0.56
"ntrees: 7501 nodesize: 15 sampSize 1"	0.58	0.56
"ntrees: 10001 nodesize: 15 sampSize 1"	0.58	0.56
"ntrees: 12501 nodesize: 15 sampSize 1"	0.58	0.56
"ntrees: 2501 nodesize: 1 sampSize 3"	0.65	0.56
"ntrees: 5001 nodesize: 1 sampSize 3"	0.65	0.56
"ntrees: 7501 nodesize: 1 sampSize 3"	0.65	0.56
"ntrees: 10001 nodesize: 1 sampSize 3"	0.65	0.56
"ntrees: 12501 nodesize: 1 sampSize 3"	0.65	0.56
"ntrees: 2501 nodesize: 3 sampSize 3"	0.65	0.56
"ntrees: 5001 nodesize: 3 sampSize 3"	0.65	0.56
"ntrees: 7501 nodesize: 3 sampSize 3"	0.65	0.56
"ntrees: 10001 nodesize: 3 sampSize 3"	0.65	0.56
"ntrees: 12501 nodesize: 3 sampSize 3"	0.65	0.56
"ntrees: 2501 nodesize: 5 sampSize 3"	0.64	0.56
"ntrees: 5001 nodesize: 5 sampSize 3"	0.64	0.56
"ntrees: 7501 nodesize: 5 sampSize 3"	0.64	0.56
"ntrees: 10001 nodesize: 5 sampSize 3"	0.64	0.56
"ntrees: 12501 nodesize: 5 sampSize 3"	0.64	0.56
"ntrees: 2501 nodesize: 10 sampSize 3"	0.64	0.56
"ntrees: 5001 nodesize: 10 sampSize 3"	0.64	0.56
"ntrees: 7501 nodesize: 10 sampSize 3"	0.64	0.56
"ntrees: 10001 nodesize: 10 sampSize 3"	0.64	0.56
"ntrees: 12501 nodesize: 10 sampSize 3"	0.64	0.56
"ntrees: 2501 nodesize: 15 sampSize 3"	0.64	0.56
"ntrees: 5001 nodesize: 15 sampSize 3"	0.64	0.56
"ntrees: 7501 nodesize: 15 sampSize 3"	0.64	0.56
"ntrees: 10001 nodesize: 15 sampSize 3"	0.64	0.56
"ntrees: 12501 nodesize: 15 sampSize 3"	0.64	0.56
"ntrees: 2501 nodesize: 1 sampSize 5"	0.65	0.57
"ntrees: 5001 nodesize: 1 sampSize 5"	0.65	0.57
"ntrees: 7501 nodesize: 1 sampSize 5"	0.65	0.57
"ntrees: 10001 nodesize: 1 sampSize 5"	0.65	0.57
"ntrees: 12501 nodesize: 1 sampSize 5"	0.65	0.57
"ntrees: 2501 nodesize: 3 sampSize 5"	0.64	0.57
"ntrees: 5001 nodesize: 3 sampSize 5"	0.64	0.57
"ntrees: 7501 nodesize: 3 sampSize 5"	0.64	0.57
"ntrees: 10001 nodesize: 3 sampSize 5"	0.64	0.57
"ntrees: 12501 nodesize: 3 sampSize 5"	0.64	0.57
"ntrees: 2501 nodesize: 5 sampSize 5"	0.63	0.55
"ntrees: 5001 nodesize: 5 sampSize 5"	0.63	0.55
"ntrees: 7501 nodesize: 5 sampSize 5"	0.63	0.55
"ntrees: 10001 nodesize: 5 sampSize 5"	0.63	0.55
"ntrees: 12501 nodesize: 5 sampSize 5"	0.63	0.55
"ntrees: 2501 nodesize: 10 sampSize 5"	0.64	0.55
"ntrees: 5001 nodesize: 10 sampSize 5"	0.64	0.55
"ntrees: 7501 nodesize: 10 sampSize 5"	0.64	0.55
"ntrees: 10001 nodesize: 10 sampSize 5"	0.64	0.55
"ntrees: 12501 nodesize: 10 sampSize 5"	0.64	0.55
"ntrees: 2501 nodesize: 15 sampSize 5"	0.64	0.55
"ntrees: 5001 nodesize: 15 sampSize 5"	0.64	0.55
"ntrees: 7501 nodesize: 15 sampSize 5"	0.64	0.55
"ntrees: 10001 nodesize: 15 sampSize 5"	0.64	0.55
"ntrees: 12501 nodesize: 15 sampSize 5"	0.64	0.55
"ntrees: 2501 nodesize: 1 sampSize 10"	0.61	0.55
"ntrees: 5001 nodesize: 1 sampSize 10"	0.61	0.55
"ntrees: 7501 nodesize: 1 sampSize 10"	0.61	0.55
"ntrees: 10001 nodesize: 1 sampSize 10"	0.61	0.55
"ntrees: 12501 nodesize: 1 sampSize 10"	0.61	0.55
"ntrees: 2501 nodesize: 3 sampSize 10"	0.62	0.54
"ntrees: 5001 nodesize: 3 sampSize 10"	0.62	0.54
"ntrees: 7501 nodesize: 3 sampSize 10"	0.62	0.54
"ntrees: 10001 nodesize: 3 sampSize 10"	0.62	0.54
"ntrees: 12501 nodesize: 3 sampSize 10"	0.62	0.54
"ntrees: 2501 nodesize: 5 sampSize 10"	0.64	0.57
"ntrees: 5001 nodesize: 5 sampSize 10"	0.64	0.57
"ntrees: 7501 nodesize: 5 sampSize 10"	0.64	0.57
"ntrees: 10001 nodesize: 5 sampSize 10"	0.64	0.57
"ntrees: 12501 nodesize: 5 sampSize 10"	0.64	0.57
"ntrees: 2501 nodesize: 10 sampSize 10"	0.62	0.55
"ntrees: 5001 nodesize: 10 sampSize 10"	0.62	0.55
"ntrees: 7501 nodesize: 10 sampSize 10"	0.62	0.55
"ntrees: 10001 nodesize: 10 sampSize 10"	0.62	0.55
"ntrees: 12501 nodesize: 10 sampSize 10"	0.62	0.55
"ntrees: 2501 nodesize: 15 sampSize 10"	0.63	0.55
"ntrees: 5001 nodesize: 15 sampSize 10"	0.63	0.55
"ntrees: 7501 nodesize: 15 sampSize 10"	0.63	0.55
"ntrees: 10001 nodesize: 15 sampSize 10"	0.63	0.55
"ntrees: 12501 nodesize: 15 sampSize 10"	0.63	0.55
"ntrees: 2501 nodesize: 1 sampSize 15"	0.61	0.54
"ntrees: 5001 nodesize: 1 sampSize 15"	0.61	0.54
"ntrees: 7501 nodesize: 1 sampSize 15"	0.61	0.54
"ntrees: 10001 nodesize: 1 sampSize 15"	0.61	0.54
"ntrees: 12501 nodesize: 1 sampSize 15"	0.61	0.54
"ntrees: 2501 nodesize: 3 sampSize 15"	0.58	0.52
"ntrees: 5001 nodesize: 3 sampSize 15"	0.58	0.52
"ntrees: 7501 nodesize: 3 sampSize 15"	0.58	0.52
"ntrees: 10001 nodesize: 3 sampSize 15"	0.58	0.52
"ntrees: 12501 nodesize: 3 sampSize 15"	0.58	0.52
"ntrees: 2501 nodesize: 5 sampSize 15"	0.61	0.54
"ntrees: 5001 nodesize: 5 sampSize 15"	0.61	0.54
"ntrees: 7501 nodesize: 5 sampSize 15"	0.61	0.54
"ntrees: 10001 nodesize: 5 sampSize 15"	0.61	0.54
"ntrees: 12501 nodesize: 5 sampSize 15"	0.61	0.54
"ntrees: 2501 nodesize: 10 sampSize 15"	0.63	0.56
"ntrees: 5001 nodesize: 10 sampSize 15"	0.63	0.56
"ntrees: 7501 nodesize: 10 sampSize 15"	0.63	0.56
"ntrees: 10001 nodesize: 10 sampSize 15"	0.63	0.56
"ntrees: 12501 nodesize: 10 sampSize 15"	0.63	0.56
"ntrees: 2501 nodesize: 15 sampSize 15"	0.59	0.54
"ntrees: 5001 nodesize: 15 sampSize 15"	0.59	0.54
"ntrees: 7501 nodesize: 15 sampSize 15"	0.59	0.54
"ntrees: 10001 nodesize: 15 sampSize 15"	0.59	0.54
"ntrees: 12501 nodesize: 15 sampSize 15"	0.59	0.54

Table 6: Results from a 70/30 data split with grid tuning of the Mtry parameter

"Hyperparameters"	"AUC-ROC"	"AUC-PR"
"ntrees: 2501 nodesize: 1 sampSize 1"	0.61	0.53
"ntrees: 5001 nodesize: 1 sampSize 1"	0.61	0.53
"ntrees: 7501 nodesize: 1 sampSize 1"	0.61	0.53
"ntrees: 10001 nodesize: 1 sampSize 1"	0.61	0.53
"ntrees: 12501 nodesize: 1 sampSize 1"	0.61	0.53
"ntrees: 2501 nodesize: 3 sampSize 1"	0.61	0.53
"ntrees: 5001 nodesize: 3 sampSize 1"	0.61	0.53
"ntrees: 7501 nodesize: 3 sampSize 1"	0.61	0.53
"ntrees: 10001 nodesize: 3 sampSize 1"	0.61	0.53
"ntrees: 12501 nodesize: 3 sampSize 1"	0.61	0.53
"ntrees: 2501 nodesize: 5 sampSize 1"	0.61	0.53
"ntrees: 5001 nodesize: 5 sampSize 1"	0.61	0.53
"ntrees: 7501 nodesize: 5 sampSize 1"	0.61	0.53
"ntrees: 10001 nodesize: 5 sampSize 1"	0.61	0.53
"ntrees: 12501 nodesize: 5 sampSize 1"	0.61	0.53
"ntrees: 2501 nodesize: 10 sampSize 1"	0.61	0.53
"ntrees: 5001 nodesize: 10 sampSize 1"	0.61	0.53
"ntrees: 7501 nodesize: 10 sampSize 1"	0.61	0.53
"ntrees: 10001 nodesize: 10 sampSize 1"	0.61	0.53
"ntrees: 12501 nodesize: 10 sampSize 1"	0.61	0.53
"ntrees: 2501 nodesize: 15 sampSize 1"	0.61	0.53
"ntrees: 5001 nodesize: 15 sampSize 1"	0.61	0.53
"ntrees: 7501 nodesize: 15 sampSize 1"	0.61	0.53
"ntrees: 10001 nodesize: 15 sampSize 1"	0.61	0.53
"ntrees: 12501 nodesize: 15 sampSize 1"	0.61	0.53
"ntrees: 2501 nodesize: 1 sampSize 3"	0.63	0.56
"ntrees: 5001 nodesize: 1 sampSize 3"	0.63	0.56
"ntrees: 7501 nodesize: 1 sampSize 3"	0.63	0.56
"ntrees: 10001 nodesize: 1 sampSize 3"	0.63	0.56
"ntrees: 12501 nodesize: 1 sampSize 3"	0.63	0.56
"ntrees: 2501 nodesize: 3 sampSize 3"	0.63	0.56
"ntrees: 5001 nodesize: 3 sampSize 3"	0.63	0.56
"ntrees: 7501 nodesize: 3 sampSize 3"	0.63	0.56
"ntrees: 10001 nodesize: 3 sampSize 3"	0.63	0.56
"ntrees: 12501 nodesize: 3 sampSize 3"	0.63	0.56
"ntrees: 2501 nodesize: 5 sampSize 3"	0.67	0.58
"ntrees: 5001 nodesize: 5 sampSize 3"	0.67	0.58
"ntrees: 7501 nodesize: 5 sampSize 3"	0.67	0.58
"ntrees: 10001 nodesize: 5 sampSize 3"	0.67	0.58
"ntrees: 12501 nodesize: 5 sampSize 3"	0.67	0.58
"ntrees: 2501 nodesize: 10 sampSize 3"	0.67	0.58
"ntrees: 5001 nodesize: 10 sampSize 3"	0.67	0.58
"ntrees: 7501 nodesize: 10 sampSize 3"	0.67	0.58
"ntrees: 10001 nodesize: 10 sampSize 3"	0.67	0.58
"ntrees: 12501 nodesize: 10 sampSize 3"	0.67	0.58
"ntrees: 2501 nodesize: 15 sampSize 3"	0.67	0.58
"ntrees: 5001 nodesize: 15 sampSize 3"	0.67	0.58
"ntrees: 7501 nodesize: 15 sampSize 3"	0.67	0.58
"ntrees: 10001 nodesize: 15 sampSize 3"	0.67	0.58
"ntrees: 12501 nodesize: 15 sampSize 3"	0.67	0.58
"ntrees: 2501 nodesize: 1 sampSize 5"	0.66	0.59
"ntrees: 5001 nodesize: 1 sampSize 5"	0.66	0.59
"ntrees: 7501 nodesize: 1 sampSize 5"	0.66	0.59
"ntrees: 10001 nodesize: 1 sampSize 5"	0.66	0.59
"ntrees: 12501 nodesize: 1 sampSize 5"	0.66	0.59
"ntrees: 2501 nodesize: 3 sampSize 5"	0.65	0.58
"ntrees: 5001 nodesize: 3 sampSize 5"	0.65	0.58
"ntrees: 7501 nodesize: 3 sampSize 5"	0.65	0.58
"ntrees: 10001 nodesize: 3 sampSize 5"	0.65	0.58
"ntrees: 12501 nodesize: 3 sampSize 5"	0.65	0.58
"ntrees: 2501 nodesize: 5 sampSize 5"	0.65	0.58
"ntrees: 5001 nodesize: 5 sampSize 5"	0.65	0.58
"ntrees: 7501 nodesize: 5 sampSize 5"	0.65	0.58
"ntrees: 10001 nodesize: 5 sampSize 5"	0.65	0.58
"ntrees: 12501 nodesize: 5 sampSize 5"	0.65	0.58
"ntrees: 2501 nodesize: 10 sampSize 5"	0.62	0.55
"ntrees: 5001 nodesize: 10 sampSize 5"	0.62	0.55
"ntrees: 7501 nodesize: 10 sampSize 5"	0.62	0.55
"ntrees: 10001 nodesize: 10 sampSize 5"	0.62	0.55
"ntrees: 12501 nodesize: 10 sampSize 5"	0.62	0.55
"ntrees: 2501 nodesize: 15 sampSize 5"	0.62	0.55
"ntrees: 5001 nodesize: 15 sampSize 5"	0.62	0.55
"ntrees: 7501 nodesize: 15 sampSize 5"	0.62	0.55
"ntrees: 10001 nodesize: 15 sampSize 5"	0.62	0.55
"ntrees: 12501 nodesize: 15 sampSize 5"	0.62	0.55
"ntrees: 2501 nodesize: 1 sampSize 10"	0.63	0.55
"ntrees: 5001 nodesize: 1 sampSize 10"	0.63	0.55
"ntrees: 7501 nodesize: 1 sampSize 10"	0.63	0.55
"ntrees: 10001 nodesize: 1 sampSize 10"	0.63	0.55
"ntrees: 12501 nodesize: 1 sampSize 10"	0.63	0.55
"ntrees: 2501 nodesize: 3 sampSize 10"	0.66	0.58
"ntrees: 5001 nodesize: 3 sampSize 10"	0.66	0.58
"ntrees: 7501 nodesize: 3 sampSize 10"	0.66	0.58
"ntrees: 10001 nodesize: 3 sampSize 10"	0.66	0.58
"ntrees: 12501 nodesize: 3 sampSize 10"	0.66	0.58
"ntrees: 2501 nodesize: 5 sampSize 10"	0.63	0.56
"ntrees: 5001 nodesize: 5 sampSize 10"	0.63	0.56
"ntrees: 7501 nodesize: 5 sampSize 10"	0.63	0.56
"ntrees: 10001 nodesize: 5 sampSize 10"	0.63	0.56
"ntrees: 12501 nodesize: 5 sampSize 10"	0.63	0.56
"ntrees: 2501 nodesize: 10 sampSize 10"	0.61	0.55
"ntrees: 5001 nodesize: 10 sampSize 10"	0.61	0.55
"ntrees: 7501 nodesize: 10 sampSize 10"	0.61	0.55
"ntrees: 10001 nodesize: 10 sampSize 10"	0.61	0.55
"ntrees: 12501 nodesize: 10 sampSize 10"	0.61	0.55
"ntrees: 2501 nodesize: 15 sampSize 10"	0.62	0.55
"ntrees: 5001 nodesize: 15 sampSize 10"	0.62	0.55
"ntrees: 7501 nodesize: 15 sampSize 10"	0.62	0.55
"ntrees: 10001 nodesize: 15 sampSize 10"	0.62	0.55
"ntrees: 12501 nodesize: 15 sampSize 10"	0.62	0.55
"ntrees: 2501 nodesize: 1 sampSize 15"	0.65	0.59
"ntrees: 5001 nodesize: 1 sampSize 15"	0.65	0.59
"ntrees: 7501 nodesize: 1 sampSize 15"	0.65	0.59
"ntrees: 10001 nodesize: 1 sampSize 15"	0.65	0.59
"ntrees: 12501 nodesize: 1 sampSize 15"	0.65	0.59
"ntrees: 2501 nodesize: 3 sampSize 15"	0.61	0.55
"ntrees: 5001 nodesize: 3 sampSize 15"	0.61	0.55
"ntrees: 7501 nodesize: 3 sampSize 15"	0.61	0.55
"ntrees: 10001 nodesize: 3 sampSize 15"	0.61	0.55
"ntrees: 12501 nodesize: 3 sampSize 15"	0.61	0.55
"ntrees: 2501 nodesize: 5 sampSize 15"	0.59	0.51
"ntrees: 5001 nodesize: 5 sampSize 15"	0.59	0.51
"ntrees: 7501 nodesize: 5 sampSize 15"	0.59	0.51
"ntrees: 10001 nodesize: 5 sampSize 15"	0.59	0.51
"ntrees: 12501 nodesize: 5 sampSize 15"	0.59	0.51
"ntrees: 2501 nodesize: 10 sampSize 15"	0.59	0.54
"ntrees: 5001 nodesize: 10 sampSize 15"	0.59	0.54
"ntrees: 7501 nodesize: 10 sampSize 15"	0.59	0.54
"ntrees: 10001 nodesize: 10 sampSize 15"	0.59	0.54
"ntrees: 12501 nodesize: 10 sampSize 15"	0.59	0.54
"ntrees: 2501 nodesize: 15 sampSize 15"	0.6	0.55
"ntrees: 5001 nodesize: 15 sampSize 15"	0.6	0.55
"ntrees: 7501 nodesize: 15 sampSize 15"	0.6	0.55
"ntrees: 10001 nodesize: 15 sampSize 15"	0.6	0.55
"ntrees: 12501 nodesize: 15 sampSize 15"	0.6	0.55

Table 7: Results from a 70/30 data split with random tuning of the Mtry parameter

"Hyperparameters"	"AUC-ROC"	"AUC-PR"
"ntrees: 2501 nodesize: 1 sampSize 1"	0.79	0.66
"ntrees: 5001 nodesize: 1 sampSize 1"	0.79	0.66
"ntrees: 7501 nodesize: 1 sampSize 1"	0.79	0.66
"ntrees: 10001 nodesize: 1 sampSize 1"	0.79	0.66
"ntrees: 12501 nodesize: 1 sampSize 1"	0.79	0.66
"ntrees: 2501 nodesize: 3 sampSize 1"	0.79	0.66
"ntrees: 5001 nodesize: 3 sampSize 1"	0.79	0.66
"ntrees: 7501 nodesize: 3 sampSize 1"	0.79	0.66
"ntrees: 10001 nodesize: 3 sampSize 1"	0.79	0.66
"ntrees: 12501 nodesize: 3 sampSize 1"	0.79	0.66
"ntrees: 2501 nodesize: 5 sampSize 1"	0.79	0.66
"ntrees: 5001 nodesize: 5 sampSize 1"	0.79	0.66
"ntrees: 7501 nodesize: 5 sampSize 1"	0.79	0.66
"ntrees: 10001 nodesize: 5 sampSize 1"	0.79	0.66
"ntrees: 12501 nodesize: 5 sampSize 1"	0.79	0.66
"ntrees: 2501 nodesize: 10 sampSize 1"	0.79	0.66
"ntrees: 5001 nodesize: 10 sampSize 1"	0.79	0.66
"ntrees: 7501 nodesize: 10 sampSize 1"	0.79	0.66
"ntrees: 10001 nodesize: 10 sampSize 1"	0.79	0.66
"ntrees: 12501 nodesize: 10 sampSize 1"	0.79	0.66
"ntrees: 2501 nodesize: 15 sampSize 1"	0.79	0.66
"ntrees: 5001 nodesize: 15 sampSize 1"	0.79	0.66
"ntrees: 7501 nodesize: 15 sampSize 1"	0.79	0.66
"ntrees: 10001 nodesize: 15 sampSize 1"	0.79	0.66
"ntrees: 12501 nodesize: 15 sampSize 1"	0.79	0.66
"ntrees: 2501 nodesize: 1 sampSize 3"	0.81	0.74
"ntrees: 5001 nodesize: 1 sampSize 3"	0.81	0.74
"ntrees: 7501 nodesize: 1 sampSize 3"	0.81	0.74
"ntrees: 10001 nodesize: 1 sampSize 3"	0.81	0.74
"ntrees: 12501 nodesize: 1 sampSize 3"	0.81	0.74
"ntrees: 2501 nodesize: 3 sampSize 3"	0.76	0.69
"ntrees: 5001 nodesize: 3 sampSize 3"	0.76	0.69
"ntrees: 7501 nodesize: 3 sampSize 3"	0.76	0.69
"ntrees: 10001 nodesize: 3 sampSize 3"	0.76	0.69
"ntrees: 12501 nodesize: 3 sampSize 3"	0.76	0.69
"ntrees: 2501 nodesize: 5 sampSize 3"	0.8	0.73
"ntrees: 5001 nodesize: 5 sampSize 3"	0.8	0.73
"ntrees: 7501 nodesize: 5 sampSize 3"	0.8	0.73
"ntrees: 10001 nodesize: 5 sampSize 3"	0.8	0.73
"ntrees: 12501 nodesize: 5 sampSize 3"	0.8	0.73
"ntrees: 2501 nodesize: 10 sampSize 3"	0.8	0.73
"ntrees: 5001 nodesize: 10 sampSize 3"	0.8	0.73
"ntrees: 7501 nodesize: 10 sampSize 3"	0.8	0.73
"ntrees: 10001 nodesize: 10 sampSize 3"	0.8	0.73
"ntrees: 12501 nodesize: 10 sampSize 3"	0.8	0.73
"ntrees: 2501 nodesize: 15 sampSize 3"	0.8	0.73
"ntrees: 5001 nodesize: 15 sampSize 3"	0.8	0.73
"ntrees: 7501 nodesize: 15 sampSize 3"	0.8	0.73
"ntrees: 10001 nodesize: 15 sampSize 3"	0.8	0.73
"ntrees: 12501 nodesize: 15 sampSize 3"	0.8	0.73
"ntrees: 2501 nodesize: 1 sampSize 5"	0.79	0.71
"ntrees: 5001 nodesize: 1 sampSize 5"	0.79	0.71
"ntrees: 7501 nodesize: 1 sampSize 5"	0.79	0.71
"ntrees: 10001 nodesize: 1 sampSize 5"	0.79	0.71
"ntrees: 12501 nodesize: 1 sampSize 5"	0.79	0.71
"ntrees: 2501 nodesize: 3 sampSize 5"	0.79	0.71
"ntrees: 5001 nodesize: 3 sampSize 5"	0.79	0.71
"ntrees: 7501 nodesize: 3 sampSize 5"	0.79	0.71
"ntrees: 10001 nodesize: 3 sampSize 5"	0.79	0.71
"ntrees: 12501 nodesize: 3 sampSize 5"	0.79	0.71
"ntrees: 2501 nodesize: 5 sampSize 5"	0.75	0.67
"ntrees: 5001 nodesize: 5 sampSize 5"	0.75	0.67
"ntrees: 7501 nodesize: 5 sampSize 5"	0.75	0.67
"ntrees: 10001 nodesize: 5 sampSize 5"	0.75	0.67
"ntrees: 12501 nodesize: 5 sampSize 5"	0.75	0.67
"ntrees: 2501 nodesize: 10 sampSize 5"	0.76	0.67
"ntrees: 5001 nodesize: 10 sampSize 5"	0.76	0.67
"ntrees: 7501 nodesize: 10 sampSize 5"	0.76	0.67
"ntrees: 10001 nodesize: 10 sampSize 5"	0.76	0.67
"ntrees: 12501 nodesize: 10 sampSize 5"	0.76	0.67
"ntrees: 2501 nodesize: 15 sampSize 5"	0.76	0.67
"ntrees: 5001 nodesize: 15 sampSize 5"	0.76	0.67
"ntrees: 7501 nodesize: 15 sampSize 5"	0.76	0.67
"ntrees: 10001 nodesize: 15 sampSize 5"	0.76	0.67
"ntrees: 12501 nodesize: 15 sampSize 5"	0.76	0.67
"ntrees: 2501 nodesize: 1 sampSize 10"	0.8	0.7
"ntrees: 5001 nodesize: 1 sampSize 10"	0.8	0.7
"ntrees: 7501 nodesize: 1 sampSize 10"	0.8	0.7
"ntrees: 10001 nodesize: 1 sampSize 10"	0.8	0.7
"ntrees: 12501 nodesize: 1 sampSize 10"	0.8	0.7
"ntrees: 2501 nodesize: 3 sampSize 10"	0.78	0.69
"ntrees: 5001 nodesize: 3 sampSize 10"	0.78	0.69
"ntrees: 7501 nodesize: 3 sampSize 10"	0.78	0.69
"ntrees: 10001 nodesize: 3 sampSize 10"	0.78	0.69
"ntrees: 12501 nodesize: 3 sampSize 10"	0.78	0.69
"ntrees: 2501 nodesize: 5 sampSize 10"	0.79	0.7
"ntrees: 5001 nodesize: 5 sampSize 10"	0.79	0.7
"ntrees: 7501 nodesize: 5 sampSize 10"	0.79	0.7
"ntrees: 10001 nodesize: 5 sampSize 10"	0.79	0.7
"ntrees: 12501 nodesize: 5 sampSize 10"	0.79	0.7
"ntrees: 2501 nodesize: 10 sampSize 10"	0.79	0.69
"ntrees: 5001 nodesize: 10 sampSize 10"	0.79	0.69
"ntrees: 7501 nodesize: 10 sampSize 10"	0.79	0.69
"ntrees: 10001 nodesize: 10 sampSize 10"	0.79	0.69
"ntrees: 12501 nodesize: 10 sampSize 10"	0.79	0.69
"ntrees: 2501 nodesize: 15 sampSize 10"	0.78	0.7
"ntrees: 5001 nodesize: 15 sampSize 10"	0.78	0.7
"ntrees: 7501 nodesize: 15 sampSize 10"	0.78	0.7
"ntrees: 10001 nodesize: 15 sampSize 10"	0.78	0.7
"ntrees: 12501 nodesize: 15 sampSize 10"	0.78	0.7
"ntrees: 2501 nodesize: 1 sampSize 15"	0.78	0.66
"ntrees: 5001 nodesize: 1 sampSize 15"	0.78	0.66
"ntrees: 7501 nodesize: 1 sampSize 15"	0.78	0.66
"ntrees: 10001 nodesize: 1 sampSize 15"	0.78	0.66
"ntrees: 12501 nodesize: 1 sampSize 15"	0.78	0.66
"ntrees: 2501 nodesize: 3 sampSize 15"	0.78	0.67
"ntrees: 5001 nodesize: 3 sampSize 15"	0.78	0.67
"ntrees: 7501 nodesize: 3 sampSize 15"	0.78	0.67
"ntrees: 10001 nodesize: 3 sampSize 15"	0.78	0.67
"ntrees: 12501 nodesize: 3 sampSize 15"	0.78	0.67
"ntrees: 2501 nodesize: 5 sampSize 15"	0.77	0.65
"ntrees: 5001 nodesize: 5 sampSize 15"	0.77	0.65
"ntrees: 7501 nodesize: 5 sampSize 15"	0.77	0.65
"ntrees: 10001 nodesize: 5 sampSize 15"	0.77	0.65
"ntrees: 12501 nodesize: 5 sampSize 15"	0.77	0.65
"ntrees: 2501 nodesize: 10 sampSize 15"	0.78	0.69
"ntrees: 5001 nodesize: 10 sampSize 15"	0.78	0.69
"ntrees: 7501 nodesize: 10 sampSize 15"	0.78	0.69
"ntrees: 10001 nodesize: 10 sampSize 15"	0.78	0.69
"ntrees: 12501 nodesize: 10 sampSize 15"	0.78	0.69
"ntrees: 2501 nodesize: 15 sampSize 15"	0.79	0.7
"ntrees: 5001 nodesize: 15 sampSize 15"	0.79	0.7
"ntrees: 7501 nodesize: 15 sampSize 15"	0.79	0.7
"ntrees: 10001 nodesize: 15 sampSize 15"	0.79	0.7
"ntrees: 12501 nodesize: 15 sampSize 15"	0.79	0.7

Table 8: Results from a 75/25 data split with grid tuning of the Mtry parameter

"Hyperparameters"	"AUC-ROC"	"AUC-PR"
"ntrees: 2501 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 5001 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 7501 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 10001 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 12501 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 2501 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 5001 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 7501 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 10001 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 12501 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 2501 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 5001 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 7501 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 10001 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 12501 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 2501 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 5001 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 7501 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 10001 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 12501 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 2501 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 5001 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 7501 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 10001 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 12501 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 2501 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 5001 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 7501 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 10001 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 12501 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 2501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 5001 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 7501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 10001 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 12501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 2501 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 5001 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 7501 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 10001 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 12501 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 2501 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 5001 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 7501 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 10001 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 12501 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 2501 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 5001 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 7501 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 10001 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 12501 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 2501 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 5001 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 7501 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 10001 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 12501 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 2501 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 5001 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 7501 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 10001 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 12501 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 2501 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 5001 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 7501 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 10001 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 12501 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 2501 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 5001 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 7501 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 10001 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 12501 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 2501 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 5001 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 7501 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 10001 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 12501 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 2501 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 5001 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 7501 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 10001 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 12501 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 2501 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 5001 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 7501 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 10001 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 12501 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 2501 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 5001 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 7501 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 10001 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 12501 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 2501 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 5001 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 7501 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 10001 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 12501 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 2501 nodesize: 15 sampSize 15"	0.77	0.66
"ntrees: 5001 nodesize: 15 sampSize 15"	0.77	0.66
"ntrees: 7501 nodesize: 15 sampSize 15"	0.77	0.66
"ntrees: 10001 nodesize: 15 sampSize 15"	0.77	0.66
"ntrees: 12501 nodesize: 15 sampSize 15"	0.77	0.66

Table 9: Results from a 75/25 data split with random tuning of the Mtry parameter

"Hyperparameters"	"AUC-ROC"	"AUC-PR"
"ntrees: 2501 nodesize: 1 sampSize 1"	0.79	0.72
"ntrees: 5001 nodesize: 1 sampSize 1"	0.79	0.72
"ntrees: 7501 nodesize: 1 sampSize 1"	0.79	0.72
"ntrees: 10001 nodesize: 1 sampSize 1"	0.79	0.72
"ntrees: 12501 nodesize: 1 sampSize 1"	0.79	0.72
"ntrees: 2501 nodesize: 3 sampSize 1"	0.79	0.72
"ntrees: 5001 nodesize: 3 sampSize 1"	0.79	0.72
"ntrees: 7501 nodesize: 3 sampSize 1"	0.79	0.72
"ntrees: 10001 nodesize: 3 sampSize 1"	0.79	0.72
"ntrees: 12501 nodesize: 3 sampSize 1"	0.79	0.72
"ntrees: 2501 nodesize: 5 sampSize 1"	0.79	0.72
"ntrees: 5001 nodesize: 5 sampSize 1"	0.79	0.72
"ntrees: 7501 nodesize: 5 sampSize 1"	0.79	0.72
"ntrees: 10001 nodesize: 5 sampSize 1"	0.79	0.72
"ntrees: 12501 nodesize: 5 sampSize 1"	0.79	0.72
"ntrees: 2501 nodesize: 10 sampSize 1"	0.79	0.72
"ntrees: 5001 nodesize: 10 sampSize 1"	0.79	0.72
"ntrees: 7501 nodesize: 10 sampSize 1"	0.79	0.72
"ntrees: 10001 nodesize: 10 sampSize 1"	0.79	0.72
"ntrees: 12501 nodesize: 10 sampSize 1"	0.79	0.72
"ntrees: 2501 nodesize: 15 sampSize 1"	0.79	0.72
"ntrees: 5001 nodesize: 15 sampSize 1"	0.79	0.72
"ntrees: 7501 nodesize: 15 sampSize 1"	0.79	0.72
"ntrees: 10001 nodesize: 15 sampSize 1"	0.79	0.72
"ntrees: 12501 nodesize: 15 sampSize 1"	0.79	0.72
"ntrees: 2501 nodesize: 1 sampSize 3"	0.77	0.69
"ntrees: 5001 nodesize: 1 sampSize 3"	0.77	0.69
"ntrees: 7501 nodesize: 1 sampSize 3"	0.77	0.69
"ntrees: 10001 nodesize: 1 sampSize 3"	0.77	0.69
"ntrees: 12501 nodesize: 1 sampSize 3"	0.77	0.69
"ntrees: 2501 nodesize: 3 sampSize 3"	0.77	0.69
"ntrees: 5001 nodesize: 3 sampSize 3"	0.77	0.69
"ntrees: 7501 nodesize: 3 sampSize 3"	0.77	0.69
"ntrees: 10001 nodesize: 3 sampSize 3"	0.77	0.69
"ntrees: 12501 nodesize: 3 sampSize 3"	0.77	0.69
"ntrees: 2501 nodesize: 5 sampSize 3"	0.77	0.67
"ntrees: 5001 nodesize: 5 sampSize 3"	0.77	0.67
"ntrees: 7501 nodesize: 5 sampSize 3"	0.77	0.67
"ntrees: 10001 nodesize: 5 sampSize 3"	0.77	0.67
"ntrees: 12501 nodesize: 5 sampSize 3"	0.77	0.67
"ntrees: 2501 nodesize: 10 sampSize 3"	0.77	0.67
"ntrees: 5001 nodesize: 10 sampSize 3"	0.77	0.67
"ntrees: 7501 nodesize: 10 sampSize 3"	0.77	0.67
"ntrees: 10001 nodesize: 10 sampSize 3"	0.77	0.67
"ntrees: 12501 nodesize: 10 sampSize 3"	0.77	0.67
"ntrees: 2501 nodesize: 15 sampSize 3"	0.77	0.67
"ntrees: 5001 nodesize: 15 sampSize 3"	0.77	0.67
"ntrees: 7501 nodesize: 15 sampSize 3"	0.77	0.67
"ntrees: 10001 nodesize: 15 sampSize 3"	0.77	0.67
"ntrees: 12501 nodesize: 15 sampSize 3"	0.77	0.67
"ntrees: 2501 nodesize: 1 sampSize 5"	0.82	0.71
"ntrees: 5001 nodesize: 1 sampSize 5"	0.82	0.71
"ntrees: 7501 nodesize: 1 sampSize 5"	0.82	0.71
"ntrees: 10001 nodesize: 1 sampSize 5"	0.82	0.71
"ntrees: 12501 nodesize: 1 sampSize 5"	0.82	0.71
"ntrees: 2501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 5001 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 7501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 10001 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 12501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 2501 nodesize: 5 sampSize 5"	0.8	0.67
"ntrees: 5001 nodesize: 5 sampSize 5"	0.8	0.67
"ntrees: 7501 nodesize: 5 sampSize 5"	0.8	0.67
"ntrees: 10001 nodesize: 5 sampSize 5"	0.8	0.67
"ntrees: 12501 nodesize: 5 sampSize 5"	0.8	0.67
"ntrees: 2501 nodesize: 10 sampSize 5"	0.79	0.72
"ntrees: 5001 nodesize: 10 sampSize 5"	0.79	0.72
"ntrees: 7501 nodesize: 10 sampSize 5"	0.79	0.72
"ntrees: 10001 nodesize: 10 sampSize 5"	0.79	0.72
"ntrees: 12501 nodesize: 10 sampSize 5"	0.79	0.72
"ntrees: 2501 nodesize: 15 sampSize 5"	0.79	0.72
"ntrees: 5001 nodesize: 15 sampSize 5"	0.79	0.72
"ntrees: 7501 nodesize: 15 sampSize 5"	0.79	0.72
"ntrees: 10001 nodesize: 15 sampSize 5"	0.79	0.72
"ntrees: 12501 nodesize: 15 sampSize 5"	0.79	0.72
"ntrees: 2501 nodesize: 1 sampSize 10"	0.79	0.68
"ntrees: 5001 nodesize: 1 sampSize 10"	0.79	0.68
"ntrees: 7501 nodesize: 1 sampSize 10"	0.79	0.68
"ntrees: 10001 nodesize: 1 sampSize 10"	0.79	0.68
"ntrees: 12501 nodesize: 1 sampSize 10"	0.79	0.68
"ntrees: 2501 nodesize: 3 sampSize 10"	0.75	0.65
"ntrees: 5001 nodesize: 3 sampSize 10"	0.75	0.65
"ntrees: 7501 nodesize: 3 sampSize 10"	0.75	0.65
"ntrees: 10001 nodesize: 3 sampSize 10"	0.75	0.65
"ntrees: 12501 nodesize: 3 sampSize 10"	0.75	0.65
"ntrees: 2501 nodesize: 5 sampSize 10"	0.81	0.7
"ntrees: 5001 nodesize: 5 sampSize 10"	0.81	0.7
"ntrees: 7501 nodesize: 5 sampSize 10"	0.81	0.7
"ntrees: 10001 nodesize: 5 sampSize 10"	0.81	0.7
"ntrees: 12501 nodesize: 5 sampSize 10"	0.81	0.7
"ntrees: 2501 nodesize: 10 sampSize 10"	0.76	0.64
"ntrees: 5001 nodesize: 10 sampSize 10"	0.76	0.64
"ntrees: 7501 nodesize: 10 sampSize 10"	0.76	0.64
"ntrees: 10001 nodesize: 10 sampSize 10"	0.76	0.64
"ntrees: 12501 nodesize: 10 sampSize 10"	0.76	0.64
"ntrees: 2501 nodesize: 15 sampSize 10"	0.78	0.68
"ntrees: 5001 nodesize: 15 sampSize 10"	0.78	0.68
"ntrees: 7501 nodesize: 15 sampSize 10"	0.78	0.68
"ntrees: 10001 nodesize: 15 sampSize 10"	0.78	0.68
"ntrees: 12501 nodesize: 15 sampSize 10"	0.78	0.68
"ntrees: 2501 nodesize: 1 sampSize 15"	0.76	0.64
"ntrees: 5001 nodesize: 1 sampSize 15"	0.76	0.64
"ntrees: 7501 nodesize: 1 sampSize 15"	0.76	0.64
"ntrees: 10001 nodesize: 1 sampSize 15"	0.76	0.64
"ntrees: 12501 nodesize: 1 sampSize 15"	0.76	0.64
"ntrees: 2501 nodesize: 3 sampSize 15"	0.75	0.64
"ntrees: 5001 nodesize: 3 sampSize 15"	0.75	0.64
"ntrees: 7501 nodesize: 3 sampSize 15"	0.75	0.64
"ntrees: 10001 nodesize: 3 sampSize 15"	0.75	0.64
"ntrees: 12501 nodesize: 3 sampSize 15"	0.75	0.64
"ntrees: 2501 nodesize: 5 sampSize 15"	0.82	0.71
"ntrees: 5001 nodesize: 5 sampSize 15"	0.82	0.71
"ntrees: 7501 nodesize: 5 sampSize 15"	0.82	0.71
"ntrees: 10001 nodesize: 5 sampSize 15"	0.82	0.71
"ntrees: 12501 nodesize: 5 sampSize 15"	0.82	0.71
"ntrees: 2501 nodesize: 10 sampSize 15"	0.78	0.67
"ntrees: 5001 nodesize: 10 sampSize 15"	0.78	0.67
"ntrees: 7501 nodesize: 10 sampSize 15"	0.78	0.67
"ntrees: 10001 nodesize: 10 sampSize 15"	0.78	0.67
"ntrees: 12501 nodesize: 10 sampSize 15"	0.78	0.67
"ntrees: 2501 nodesize: 15 sampSize 15"	0.77	0.67
"ntrees: 5001 nodesize: 15 sampSize 15"	0.77	0.67
"ntrees: 7501 nodesize: 15 sampSize 15"	0.77	0.67
"ntrees: 10001 nodesize: 15 sampSize 15"	0.77	0.67
"ntrees: 12501 nodesize: 15 sampSize 15"	0.77	0.67

Table 10: Results from a 80/20 data split with grid tuning of the Mtry parameter

"Hyperparameters"	"AUC-ROC"	"AUC-PR"
"ntrees: 2501 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 1 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 3 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 5 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 10 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 5001 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 7501 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 10001 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 12501 nodesize: 15 sampSize 1"	0.73	0.69
"ntrees: 2501 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 5001 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 7501 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 10001 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 12501 nodesize: 1 sampSize 3"	0.81	0.72
"ntrees: 2501 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 5001 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 7501 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 10001 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 12501 nodesize: 3 sampSize 3"	0.78	0.66
"ntrees: 2501 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 5001 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 7501 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 10001 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 12501 nodesize: 5 sampSize 3"	0.8	0.68
"ntrees: 2501 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 5001 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 7501 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 10001 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 12501 nodesize: 10 sampSize 3"	0.8	0.68
"ntrees: 2501 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 5001 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 7501 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 10001 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 12501 nodesize: 15 sampSize 3"	0.8	0.68
"ntrees: 2501 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 5001 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 7501 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 10001 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 12501 nodesize: 1 sampSize 5"	0.78	0.67
"ntrees: 2501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 5001 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 7501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 10001 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 12501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 2501 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 5001 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 7501 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 10001 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 12501 nodesize: 5 sampSize 5"	0.79	0.69
"ntrees: 2501 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 5001 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 7501 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 10001 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 12501 nodesize: 10 sampSize 5"	0.77	0.69
"ntrees: 2501 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 5001 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 7501 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 10001 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 12501 nodesize: 15 sampSize 5"	0.77	0.69
"ntrees: 2501 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 5001 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 7501 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 10001 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 12501 nodesize: 1 sampSize 10"	0.77	0.69
"ntrees: 2501 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 5001 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 7501 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 10001 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 12501 nodesize: 3 sampSize 10"	0.78	0.67
"ntrees: 2501 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 5001 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 7501 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 10001 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 12501 nodesize: 5 sampSize 10"	0.79	0.69
"ntrees: 2501 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 5001 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 7501 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 10001 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 12501 nodesize: 10 sampSize 10"	0.8	0.7
"ntrees: 2501 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 5001 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 7501 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 10001 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 12501 nodesize: 15 sampSize 10"	0.79	0.69
"ntrees: 2501 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 5001 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 7501 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 10001 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 12501 nodesize: 1 sampSize 15"	0.79	0.68
"ntrees: 2501 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 5001 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 7501 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 10001 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 12501 nodesize: 3 sampSize 15"	0.79	0.67
"ntrees: 2501 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 5001 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 7501 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 10001 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 12501 nodesize: 5 sampSize 15"	0.78	0.67
"ntrees: 2501 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 5001 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 7501 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 10001 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 12501 nodesize: 10 sampSize 15"	0.76	0.65
"ntrees: 2501 nodesize: 15 sampSize 15"	0.77	0.66
"ntrees: 5001 nodesize: 15 sampSize 15"	0.77	0.66
"ntrees: 7501 nodesize: 15 sampSize 15"	0.77	0.66
"ntrees: 10001 nodesize: 15 sampSize 15"	0.77	0.66
"ntrees: 12501 nodesize: 15 sampSize 15"	0.77	0.66

Table 11: Results from a 80/20 data split with random tuning of the Mtry parameter

"Hyperparameters"	"AUC-ROC"	"AUC-PR"
"ntrees: 2501 nodesize: 1 sampSize 1"	0.63	0.64
"ntrees: 5001 nodesize: 1 sampSize 1"	0.63	0.64
"ntrees: 7501 nodesize: 1 sampSize 1"	0.63	0.64
"ntrees: 10001 nodesize: 1 sampSize 1"	0.63	0.64
"ntrees: 12501 nodesize: 1 sampSize 1"	0.63	0.64
"ntrees: 2501 nodesize: 3 sampSize 1"	0.63	0.64
"ntrees: 5001 nodesize: 3 sampSize 1"	0.63	0.64
"ntrees: 7501 nodesize: 3 sampSize 1"	0.63	0.64
"ntrees: 10001 nodesize: 3 sampSize 1"	0.63	0.64
"ntrees: 12501 nodesize: 3 sampSize 1"	0.63	0.64
"ntrees: 2501 nodesize: 5 sampSize 1"	0.63	0.64
"ntrees: 5001 nodesize: 5 sampSize 1"	0.63	0.64
"ntrees: 7501 nodesize: 5 sampSize 1"	0.63	0.64
"ntrees: 10001 nodesize: 5 sampSize 1"	0.63	0.64
"ntrees: 12501 nodesize: 5 sampSize 1"	0.63	0.64
"ntrees: 2501 nodesize: 10 sampSize 1"	0.63	0.64
"ntrees: 5001 nodesize: 10 sampSize 1"	0.63	0.64
"ntrees: 7501 nodesize: 10 sampSize 1"	0.63	0.64
"ntrees: 10001 nodesize: 10 sampSize 1"	0.63	0.64
"ntrees: 12501 nodesize: 10 sampSize 1"	0.63	0.64
"ntrees: 2501 nodesize: 15 sampSize 1"	0.63	0.64
"ntrees: 5001 nodesize: 15 sampSize 1"	0.63	0.64
"ntrees: 7501 nodesize: 15 sampSize 1"	0.63	0.64
"ntrees: 10001 nodesize: 15 sampSize 1"	0.63	0.64
"ntrees: 12501 nodesize: 15 sampSize 1"	0.63	0.64
"ntrees: 2501 nodesize: 1 sampSize 3"	0.68	0.57
"ntrees: 5001 nodesize: 1 sampSize 3"	0.68	0.57
"ntrees: 7501 nodesize: 1 sampSize 3"	0.68	0.57
"ntrees: 10001 nodesize: 1 sampSize 3"	0.68	0.57
"ntrees: 12501 nodesize: 1 sampSize 3"	0.68	0.57
"ntrees: 2501 nodesize: 3 sampSize 3"	0.68	0.57
"ntrees: 5001 nodesize: 3 sampSize 3"	0.68	0.57
"ntrees: 7501 nodesize: 3 sampSize 3"	0.68	0.57
"ntrees: 10001 nodesize: 3 sampSize 3"	0.68	0.57
"ntrees: 12501 nodesize: 3 sampSize 3"	0.68	0.57
"ntrees: 2501 nodesize: 5 sampSize 3"	0.7	0.6
"ntrees: 5001 nodesize: 5 sampSize 3"	0.7	0.6
"ntrees: 7501 nodesize: 5 sampSize 3"	0.7	0.6
"ntrees: 10001 nodesize: 5 sampSize 3"	0.7	0.6
"ntrees: 12501 nodesize: 5 sampSize 3"	0.7	0.6
"ntrees: 2501 nodesize: 10 sampSize 3"	0.7	0.6
"ntrees: 5001 nodesize: 10 sampSize 3"	0.7	0.6
"ntrees: 7501 nodesize: 10 sampSize 3"	0.7	0.6
"ntrees: 10001 nodesize: 10 sampSize 3"	0.7	0.6
"ntrees: 12501 nodesize: 10 sampSize 3"	0.7	0.6
"ntrees: 2501 nodesize: 15 sampSize 3"	0.7	0.6
"ntrees: 5001 nodesize: 15 sampSize 3"	0.7	0.6
"ntrees: 7501 nodesize: 15 sampSize 3"	0.7	0.6
"ntrees: 10001 nodesize: 15 sampSize 3"	0.7	0.6
"ntrees: 12501 nodesize: 15 sampSize 3"	0.7	0.6
"ntrees: 2501 nodesize: 1 sampSize 5"	0.73	0.63
"ntrees: 5001 nodesize: 1 sampSize 5"	0.73	0.63
"ntrees: 7501 nodesize: 1 sampSize 5"	0.73	0.63
"ntrees: 10001 nodesize: 1 sampSize 5"	0.73	0.63
"ntrees: 12501 nodesize: 1 sampSize 5"	0.73	0.63
"ntrees: 2501 nodesize: 3 sampSize 5"	0.72	0.6
"ntrees: 5001 nodesize: 3 sampSize 5"	0.72	0.6
"ntrees: 7501 nodesize: 3 sampSize 5"	0.72	0.6
"ntrees: 10001 nodesize: 3 sampSize 5"	0.72	0.6
"ntrees: 12501 nodesize: 3 sampSize 5"	0.72	0.6
"ntrees: 2501 nodesize: 5 sampSize 5"	0.72	0.64
"ntrees: 5001 nodesize: 5 sampSize 5"	0.72	0.64
"ntrees: 7501 nodesize: 5 sampSize 5"	0.72	0.64
"ntrees: 10001 nodesize: 5 sampSize 5"	0.72	0.64
"ntrees: 12501 nodesize: 5 sampSize 5"	0.72	0.64
"ntrees: 2501 nodesize: 10 sampSize 5"	0.72	0.67
"ntrees: 5001 nodesize: 10 sampSize 5"	0.72	0.67
"ntrees: 7501 nodesize: 10 sampSize 5"	0.72	0.67
"ntrees: 10001 nodesize: 10 sampSize 5"	0.72	0.67
"ntrees: 12501 nodesize: 10 sampSize 5"	0.72	0.67
"ntrees: 2501 nodesize: 15 sampSize 5"	0.72	0.67
"ntrees: 5001 nodesize: 15 sampSize 5"	0.72	0.67
"ntrees: 7501 nodesize: 15 sampSize 5"	0.72	0.67
"ntrees: 10001 nodesize: 15 sampSize 5"	0.72	0.67
"ntrees: 12501 nodesize: 15 sampSize 5"	0.72	0.67
"ntrees: 2501 nodesize: 1 sampSize 10"	0.73	0.71
"ntrees: 5001 nodesize: 1 sampSize 10"	0.73	0.71
"ntrees: 7501 nodesize: 1 sampSize 10"	0.73	0.71
"ntrees: 10001 nodesize: 1 sampSize 10"	0.73	0.71
"ntrees: 12501 nodesize: 1 sampSize 10"	0.73	0.71
"ntrees: 2501 nodesize: 3 sampSize 10"	0.82	0.79
"ntrees: 5001 nodesize: 3 sampSize 10"	0.82	0.79
"ntrees: 7501 nodesize: 3 sampSize 10"	0.82	0.79
"ntrees: 10001 nodesize: 3 sampSize 10"	0.82	0.79
"ntrees: 12501 nodesize: 3 sampSize 10"	0.82	0.79
"ntrees: 2501 nodesize: 5 sampSize 10"	0.75	0.69
"ntrees: 5001 nodesize: 5 sampSize 10"	0.75	0.69
"ntrees: 7501 nodesize: 5 sampSize 10"	0.75	0.69
"ntrees: 10001 nodesize: 5 sampSize 10"	0.75	0.69
"ntrees: 12501 nodesize: 5 sampSize 10"	0.75	0.69
"ntrees: 2501 nodesize: 10 sampSize 10"	0.74	0.68
"ntrees: 5001 nodesize: 10 sampSize 10"	0.74	0.68
"ntrees: 7501 nodesize: 10 sampSize 10"	0.74	0.68
"ntrees: 10001 nodesize: 10 sampSize 10"	0.74	0.68
"ntrees: 12501 nodesize: 10 sampSize 10"	0.74	0.68
"ntrees: 2501 nodesize: 15 sampSize 10"	0.76	0.71
"ntrees: 5001 nodesize: 15 sampSize 10"	0.76	0.71
"ntrees: 7501 nodesize: 15 sampSize 10"	0.76	0.71
"ntrees: 10001 nodesize: 15 sampSize 10"	0.76	0.71
"ntrees: 12501 nodesize: 15 sampSize 10"	0.76	0.71
"ntrees: 2501 nodesize: 1 sampSize 15"	0.74	0.72
"ntrees: 5001 nodesize: 1 sampSize 15"	0.74	0.72
"ntrees: 7501 nodesize: 1 sampSize 15"	0.74	0.72
"ntrees: 10001 nodesize: 1 sampSize 15"	0.74	0.72
"ntrees: 12501 nodesize: 1 sampSize 15"	0.74	0.72
"ntrees: 2501 nodesize: 3 sampSize 15"	0.73	0.67
"ntrees: 5001 nodesize: 3 sampSize 15"	0.73	0.67
"ntrees: 7501 nodesize: 3 sampSize 15"	0.73	0.67
"ntrees: 10001 nodesize: 3 sampSize 15"	0.73	0.67
"ntrees: 12501 nodesize: 3 sampSize 15"	0.73	0.67
"ntrees: 2501 nodesize: 5 sampSize 15"	0.79	0.76
"ntrees: 5001 nodesize: 5 sampSize 15"	0.79	0.76
"ntrees: 7501 nodesize: 5 sampSize 15"	0.79	0.76
"ntrees: 10001 nodesize: 5 sampSize 15"	0.79	0.76
"ntrees: 12501 nodesize: 5 sampSize 15"	0.79	0.76
"ntrees: 2501 nodesize: 10 sampSize 15"	0.78	0.73
"ntrees: 5001 nodesize: 10 sampSize 15"	0.78	0.73
"ntrees: 7501 nodesize: 10 sampSize 15"	0.78	0.73
"ntrees: 10001 nodesize: 10 sampSize 15"	0.78	0.73
"ntrees: 12501 nodesize: 10 sampSize 15"	0.78	0.73
"ntrees: 2501 nodesize: 15 sampSize 15"	0.73	0.73
"ntrees: 5001 nodesize: 15 sampSize 15"	0.73	0.73
"ntrees: 7501 nodesize: 15 sampSize 15"	0.73	0.73
"ntrees: 10001 nodesize: 15 sampSize 15"	0.73	0.73
"ntrees: 12501 nodesize: 15 sampSize 15"	0.73	0.73

Table 12: Results from a 85/15 data split with grid tuning of the Mtry parameter

"Hyperparameters"	"AUC-ROC"	"AUC-PR"
"ntrees: 2501 nodesize: 1 sampSize 1"	0.75	0.66
"ntrees: 5001 nodesize: 1 sampSize 1"	0.75	0.66
"ntrees: 7501 nodesize: 1 sampSize 1"	0.75	0.66
"ntrees: 10001 nodesize: 1 sampSize 1"	0.75	0.66
"ntrees: 12501 nodesize: 1 sampSize 1"	0.75	0.66
"ntrees: 2501 nodesize: 3 sampSize 1"	0.75	0.66
"ntrees: 5001 nodesize: 3 sampSize 1"	0.75	0.66
"ntrees: 7501 nodesize: 3 sampSize 1"	0.75	0.66
"ntrees: 10001 nodesize: 3 sampSize 1"	0.75	0.66
"ntrees: 12501 nodesize: 3 sampSize 1"	0.75	0.66
"ntrees: 2501 nodesize: 5 sampSize 1"	0.75	0.66
"ntrees: 5001 nodesize: 5 sampSize 1"	0.75	0.66
"ntrees: 7501 nodesize: 5 sampSize 1"	0.75	0.66
"ntrees: 10001 nodesize: 5 sampSize 1"	0.75	0.66
"ntrees: 12501 nodesize: 5 sampSize 1"	0.75	0.66
"ntrees: 2501 nodesize: 10 sampSize 1"	0.75	0.66
"ntrees: 5001 nodesize: 10 sampSize 1"	0.75	0.66
"ntrees: 7501 nodesize: 10 sampSize 1"	0.75	0.66
"ntrees: 10001 nodesize: 10 sampSize 1"	0.75	0.66
"ntrees: 12501 nodesize: 10 sampSize 1"	0.75	0.66
"ntrees: 2501 nodesize: 15 sampSize 1"	0.75	0.66
"ntrees: 5001 nodesize: 15 sampSize 1"	0.75	0.66
"ntrees: 7501 nodesize: 15 sampSize 1"	0.75	0.66
"ntrees: 10001 nodesize: 15 sampSize 1"	0.75	0.66
"ntrees: 12501 nodesize: 15 sampSize 1"	0.75	0.66
"ntrees: 2501 nodesize: 1 sampSize 3"	0.69	0.6
"ntrees: 5001 nodesize: 1 sampSize 3"	0.69	0.6
"ntrees: 7501 nodesize: 1 sampSize 3"	0.69	0.6
"ntrees: 10001 nodesize: 1 sampSize 3"	0.69	0.6
"ntrees: 12501 nodesize: 1 sampSize 3"	0.69	0.6
"ntrees: 2501 nodesize: 3 sampSize 3"	0.69	0.6
"ntrees: 5001 nodesize: 3 sampSize 3"	0.69	0.6
"ntrees: 7501 nodesize: 3 sampSize 3"	0.69	0.6
"ntrees: 10001 nodesize: 3 sampSize 3"	0.69	0.6
"ntrees: 12501 nodesize: 3 sampSize 3"	0.69	0.6
"ntrees: 2501 nodesize: 5 sampSize 3"	0.69	0.6
"ntrees: 5001 nodesize: 5 sampSize 3"	0.69	0.6
"ntrees: 7501 nodesize: 5 sampSize 3"	0.69	0.6
"ntrees: 10001 nodesize: 5 sampSize 3"	0.69	0.6
"ntrees: 12501 nodesize: 5 sampSize 3"	0.69	0.6
"ntrees: 2501 nodesize: 10 sampSize 3"	0.69	0.6
"ntrees: 5001 nodesize: 10 sampSize 3"	0.69	0.6
"ntrees: 7501 nodesize: 10 sampSize 3"	0.69	0.6
"ntrees: 10001 nodesize: 10 sampSize 3"	0.69	0.6
"ntrees: 12501 nodesize: 10 sampSize 3"	0.69	0.6
"ntrees: 2501 nodesize: 15 sampSize 3"	0.69	0.6
"ntrees: 5001 nodesize: 15 sampSize 3"	0.69	0.6
"ntrees: 7501 nodesize: 15 sampSize 3"	0.69	0.6
"ntrees: 10001 nodesize: 15 sampSize 3"	0.69	0.6
"ntrees: 12501 nodesize: 15 sampSize 3"	0.69	0.6
"ntrees: 2501 nodesize: 1 sampSize 5"	0.69	0.65
"ntrees: 5001 nodesize: 1 sampSize 5"	0.69	0.65
"ntrees: 7501 nodesize: 1 sampSize 5"	0.69	0.65
"ntrees: 10001 nodesize: 1 sampSize 5"	0.69	0.65
"ntrees: 12501 nodesize: 1 sampSize 5"	0.69	0.65
"ntrees: 2501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 5001 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 7501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 10001 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 12501 nodesize: 3 sampSize 5"	0.78	0.68
"ntrees: 2501 nodesize: 5 sampSize 5"	0.71	0.63
"ntrees: 5001 nodesize: 5 sampSize 5"	0.71	0.63
"ntrees: 7501 nodesize: 5 sampSize 5"	0.71	0.63
"ntrees: 10001 nodesize: 5 sampSize 5"	0.71	0.63
"ntrees: 12501 nodesize: 5 sampSize 5"	0.71	0.63
"ntrees: 2501 nodesize: 10 sampSize 5"	0.76	0.66
"ntrees: 5001 nodesize: 10 sampSize 5"	0.76	0.66
"ntrees: 7501 nodesize: 10 sampSize 5"	0.76	0.66
"ntrees: 10001 nodesize: 10 sampSize 5"	0.76	0.66
"ntrees: 12501 nodesize: 10 sampSize 5"	0.76	0.66
"ntrees: 2501 nodesize: 15 sampSize 5"	0.76	0.66
"ntrees: 5001 nodesize: 15 sampSize 5"	0.76	0.66
"ntrees: 7501 nodesize: 15 sampSize 5"	0.76	0.66
"ntrees: 10001 nodesize: 15 sampSize 5"	0.76	0.66
"ntrees: 12501 nodesize: 15 sampSize 5"	0.76	0.66
"ntrees: 2501 nodesize: 1 sampSize 10"	0.78	0.7
"ntrees: 5001 nodesize: 1 sampSize 10"	0.78	0.7
"ntrees: 7501 nodesize: 1 sampSize 10"	0.78	0.7
"ntrees: 10001 nodesize: 1 sampSize 10"	0.78	0.7
"ntrees: 12501 nodesize: 1 sampSize 10"	0.78	0.7
"ntrees: 2501 nodesize: 3 sampSize 10"	0.77	0.73
"ntrees: 5001 nodesize: 3 sampSize 10"	0.77	0.73
"ntrees: 7501 nodesize: 3 sampSize 10"	0.77	0.73
"ntrees: 10001 nodesize: 3 sampSize 10"	0.77	0.73
"ntrees: 12501 nodesize: 3 sampSize 10"	0.77	0.73
"ntrees: 2501 nodesize: 5 sampSize 10"	0.75	0.61
"ntrees: 5001 nodesize: 5 sampSize 10"	0.75	0.61
"ntrees: 7501 nodesize: 5 sampSize 10"	0.75	0.61
"ntrees: 10001 nodesize: 5 sampSize 10"	0.75	0.61
"ntrees: 12501 nodesize: 5 sampSize 10"	0.75	0.61
"ntrees: 2501 nodesize: 10 sampSize 10"	0.73	0.65
"ntrees: 5001 nodesize: 10 sampSize 10"	0.73	0.65
"ntrees: 7501 nodesize: 10 sampSize 10"	0.73	0.65
"ntrees: 10001 nodesize: 10 sampSize 10"	0.73	0.65
"ntrees: 12501 nodesize: 10 sampSize 10"	0.73	0.65
"ntrees: 2501 nodesize: 15 sampSize 10"	0.73	0.68
"ntrees: 5001 nodesize: 15 sampSize 10"	0.73	0.68
"ntrees: 7501 nodesize: 15 sampSize 10"	0.73	0.68
"ntrees: 10001 nodesize: 15 sampSize 10"	0.73	0.68
"ntrees: 12501 nodesize: 15 sampSize 10"	0.73	0.68
"ntrees: 2501 nodesize: 1 sampSize 15"	0.67	0.56
"ntrees: 5001 nodesize: 1 sampSize 15"	0.67	0.56
"ntrees: 7501 nodesize: 1 sampSize 15"	0.67	0.56
"ntrees: 10001 nodesize: 1 sampSize 15"	0.67	0.56
"ntrees: 12501 nodesize: 1 sampSize 15"	0.67	0.56
"ntrees: 2501 nodesize: 3 sampSize 15"	0.78	0.74
"ntrees: 5001 nodesize: 3 sampSize 15"	0.78	0.74
"ntrees: 7501 nodesize: 3 sampSize 15"	0.78	0.74
"ntrees: 10001 nodesize: 3 sampSize 15"	0.78	0.74
"ntrees: 12501 nodesize: 3 sampSize 15"	0.78	0.74
"ntrees: 2501 nodesize: 5 sampSize 15"	0.78	0.72
"ntrees: 5001 nodesize: 5 sampSize 15"	0.78	0.72
"ntrees: 7501 nodesize: 5 sampSize 15"	0.78	0.72
"ntrees: 10001 nodesize: 5 sampSize 15"	0.78	0.72
"ntrees: 12501 nodesize: 5 sampSize 15"	0.78	0.72
"ntrees: 2501 nodesize: 10 sampSize 15"	0.75	0.64
"ntrees: 5001 nodesize: 10 sampSize 15"	0.75	0.64
"ntrees: 7501 nodesize: 10 sampSize 15"	0.75	0.64
"ntrees: 10001 nodesize: 10 sampSize 15"	0.75	0.64
"ntrees: 12501 nodesize: 10 sampSize 15"	0.75	0.64
"ntrees: 2501 nodesize: 15 sampSize 15"	0.76	0.65
"ntrees: 5001 nodesize: 15 sampSize 15"	0.76	0.65
"ntrees: 7501 nodesize: 15 sampSize 15"	0.76	0.65
"ntrees: 10001 nodesize: 15 sampSize 15"	0.76	0.65
"ntrees: 12501 nodesize: 15 sampSize 15"	0.76	0.65

Table 13: Results from a 85/15 data split with random tuning of the Mtry parameter

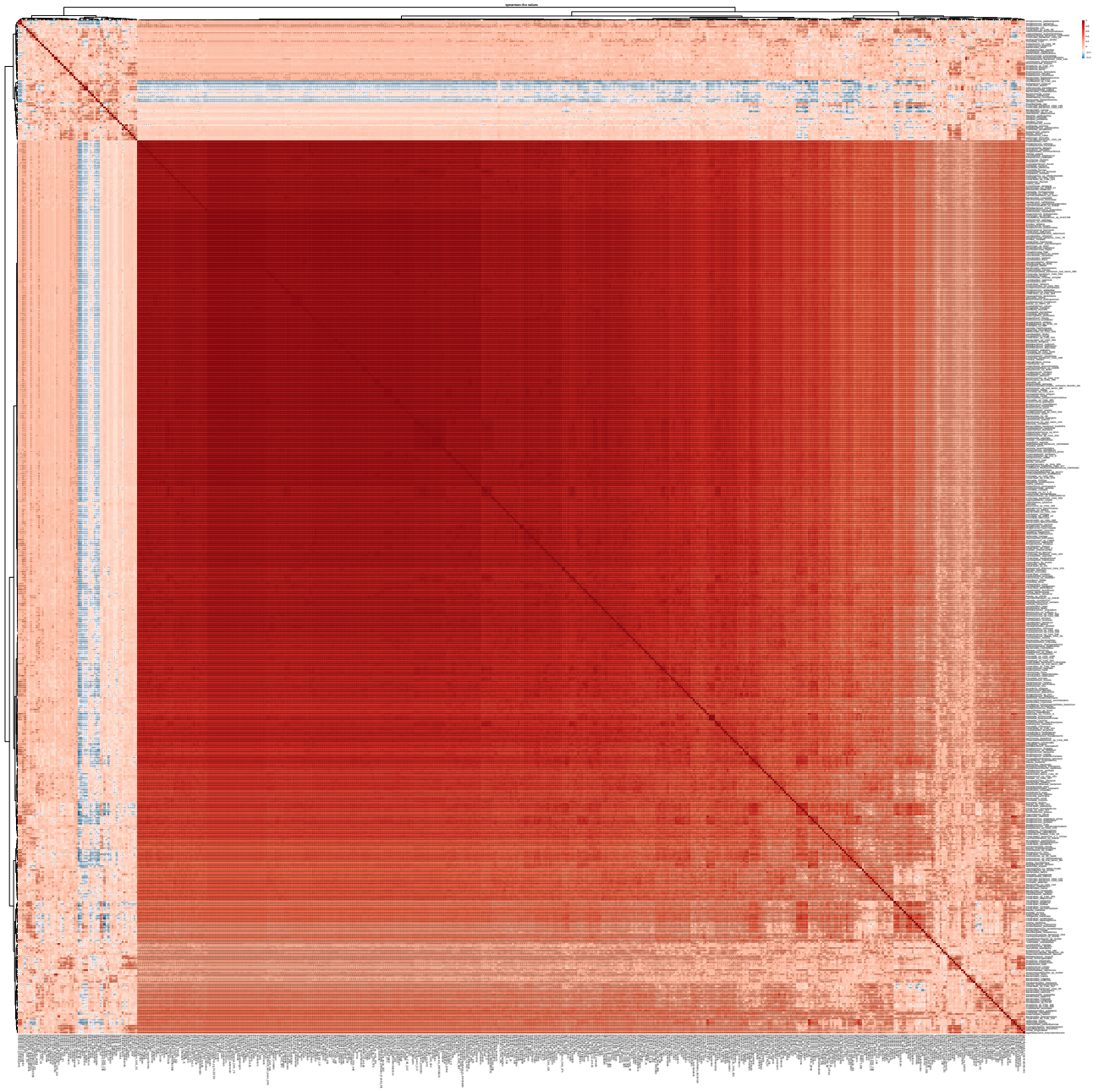

Figure 12: Spearman analysis heatmap. Red signifies a positive correlation while blue signifies a negative correlation. In each axis all 461 taxa are compared against one another.

References

A Gilbert J, J Blaser M, Caporaso G, K Jansson J, V Lynch S, Knight R. 2018. Current understanding of the human microbiome. Nature Medicine 24: 392–400.

Ahlawat S, Asha, Sharma K. 2020. Gut–organ axis: A microbial outreach and networking. Letters in Applied Microbiology 72: 636–668.

Al Alam D, Danopoulos S, Grubbs B, Ali NA, MacAogain M, Chotirmall SH, Warburton D, Gaggar A, Ambalavanan N, Lal CV, et al. 2020. Human fetal lungs harbor a microbiome signature. American Journal of Respiratory and Critical Care Medicine 201: 1002–1006.

Ananthakrishnan AN. 2015. Epidemiology and risk factors for IBD. Nat Rev Gastroenterol Hepatol 205–217.

Anderson MJ. 2017. Permutational Multivariate Analysis of Variance (PERMANOVA), John Wiley Sons, Ltd, 1–15.

Ardabili S, Mosavi A, Várkonyi-Kóczy AR. 2020. Advances in machine learning modeling reviewing hybrid and ensemble methods. Várkonyi-Kóczy AR, editor, Engineering for Sustainable Future. Springer International Publishing, Cham, 215–227.

Bai Y, Guo X, Qian X, Lu M, Chen T, Qin Y, Liu YX. 2021. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell 315–330.

Ball J, Cheek L, Bewick V. 2004. Statistics review 13: Receiver operating characteristic curves. Critical Care 8.

Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N. 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. eLife 10: e65088.

Berger FK, Schwab N, Glanemann M, Bohle RM, Gärtner B, Groesdonk HV. 2018. Flavonifractor (eubacterium) plautii bloodstream infection following acute cholecystitis. IDCases 14.

Bishop CM. 1994. Neural networks and their applications. Review of Scientific Instruments 65: 1803–1832.

Boulund F. 2018. boulund/stag-mwc: Stag-mwc v0.3.0-beta.

Breiman L. 2001. Random Forests. Springer 45: 5–32.

Calle ML. 2019. Statistical Analysis of Metagenomics Data. Genomics & Informatics 17.

Casimiro-Soriguer CS, Loucera C, Peña-Chilet M, Dopazo J. 2022. Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer. Scientific Reports 12.

Chandrashekar G, Sahin F. 2014. A survey on feature selection methods. Computers Electrical Engineering 40: 16–28. 40th-year commemorative issue.

Chen Q, Wu G, Chen H, Li H, Li S, Zhang C, Pang X, Wang L, Zhao L, Shen J. 2019. Quantification of human oral and fecal streptococcus parasanguinis by use of quantitative real-time pcr targeting the groel gene. Frontiers in Microbiology 10.

Clooney AG, Eckenberger J, Laserna-Mendieta E, Sexton KA, Bernstein MT, Vagianos K, Sargent M, Ryan FJ, Moran C, Sheehan D, Sleator RD, Targownik LE, Bernstein CN, Shanahan F, Claesson MJ. 2021. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. Gut 70: 499–510.

Cook J, Ramadas V. 2020. When to consult precision-recall curves. The Stata Journal 20: 131–148.

Corby PM, Lyons-Weiler J, Bretz WA, Hart TC, Aas JA, Boumenna T, Goss J, Corby AL, Junior HM, Weyant RJ, et al. 2005. Microbial risk indicators of early childhood caries. Journal of Clinical Microbiology 43: 5753–5759.

Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. 2007. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. Ecological Society of America 88: 2783–2792.

Dahiya DK, Renuka, Dangi AK, Shandilya UK, Puniya AK, Shukla P. 2019. Chapter 44 - new-generation probiotics: Perspectives and applications. Faintuch J, Faintuch S, editors, Microbiome and Metabolome in Diagnosis, Therapy, and other Strategic Applications, Academic Press, 417–424.

Din AU, Hassan A, Zhu Y, Zhang K, Wang Y, Li T, Wang Y, Wang G. 2020. Inhibitory effect of bifidobacterium bifidum atcc 29521 on colitis and its mechanism. The Journal of Nutritional Biochemistry 79: 108353.

DKostic A, J Xavier R, Gevers D. 2014. The Microbiome in Inflammatory Bowel Disease: Current Status and the Future Ahead. Gastroenterology 146: 1489–1499.

Dramiński M, Rada-Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J. 2007. Monte Carlo feature selection for supervised classification. Bioinformatics 24: 110–117.

Fazlollahi M, Chun Y, Grishin A, Wood RA, Burks AW, Dawson P, Jones SM, Leung DYM, Sampson HA, Sicherer SH, Bunyavanich S. 2018. Early-life gut microbiome and egg allergy. Allergy 73: 1515–1524.

de Flamingh A, Coutu A, Roca AL, Malhi RS. 2020. Accurate sex identification of ancient elephant and other animal remains using low-coverage dna shotgun sequencing data. G3 Genes|Genomes|Genetics 10: 1427–1432.

G Albenberg L, D Lewis J, D Wu G. 2012. Food and the Gut Microbiota in IBD: A Critical Connection. Current opinion in gastroenterology 28: 314–320.

Ganaie MA, Hu M, Tanveer M, Suganthan PN. 2021. Ensemble deep learning: A review. CoRR abs/2104.02395.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology 8.

Greenacre M, Grunsky E. 2019. The isometric logratio transformation in compositional data analysis: a practical evaluation. Economics Working Paper Series .

Hearst M, Dumais S, Osuna E, Platt J, Scholkopf B. 1998. Support vector machines. IEEE Intelligent Systems and their Applications 13: 18–28.

Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, Spitzer AI, Ramkumar PN. 2020. Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. Current Reviews in Musculoskeletal Medicine 13: 69–76.

Henke MT, Kenny DJ, Cassilly CD, Vlamakis H, Xavier RJ, Clardy J. 2019. ruminococcus gnavus, a member of the human gut microbiome associated with crohn’s disease, produces an inflammatory polysaccharide. Proceedings of the National Academy of Sciences 116: 12672–12677.

Holzinger A. 2018. From Machine Learning to Explainable AI. Košice, Slovakia.

Hosni M, Abnane I, Idri A, Carrillo de Gea JM, Fernández Alemán JL. 2019. Reviewing ensemble classification methods in breast cancer. Computer Methods and Programs in Biomedicine 177: 89–112.

J Sharpton T. 2014. An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science 5.

Kaplan GG. 2015. The global burden of IBD: from 2015 to 2025. Nat Rev Gastroenterol Hepatol 12: 720–727.

KELLERMAYER R. 2019. Roseburia Species: Prime candidates for Microbial therapeutics in Inflammatory bowel Disease. Gastroenterology 157.

Kramer O. 2013. K-Nearest Neighbors, Springer Berlin Heidelberg, Berlin, Heidelberg, 13–23.

Lacroix V, Cassard A, Mas E, Barreau F. 2021. Multi-omics analysis of gut microbiota in inflammatory bowel diseases: What benefits for diagnostic, prognostic and therapeutic tools? International Journal of Molecular Sciences 22: 11255.

Laudadio I, Fulci V, Palone F, Stronati L, Cucchiara S, Carissimi C. 2018. Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. OMICS: A Journal of Integrative Biology 22.

LaValley MP. 2008. Logistic regression. Circulation 117: 2395–2399.

Li H. 2015. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annual Review of Statistics and Its Application 2: 73–94.

Liu X, Shen M, Yan H, Long P, Jiang H, Zhang Y, Zhou L, Yu K, Qiu G, Yang H, Li X, Min X, He M, Zhang X, Choi H, Wang C, Wu T. 2022. Alternations in the gut microbiota and metabolome with newly diagnosed unstable angina. Journal of Genetics and Genomics 49: 240–248.

Lloyd-Price J, Arze C, Ananthakrishnan A, Schirmer M, Avila-Pacheco J. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 569: 655–662.

Lo CH, Khandpur N, Laurini Rossato S, Lochhead P, W Lopes E, E Burke K, M Richter J, Song M, Victor Ardisson Korat A, Sun Q, Fung T, Khalili H, T Chan A, N Ananthakrishnan A. 2021. Ultra-processed Foods and Risk of Crohn's Disease and Ulcerative Colitis: A Prospective Cohort Study. Clinical Gastroenterology and Hepatology .

Lo Presti A, Zorzi F, Del Chierico F, Altomare A, Cocca S, Avola A, De Biasio F, Russo A, Cella E, Reddel S, Calabrese E, Biancone L, Monteleone G, Cicala M, Angeletti S, Ciccozzi M, Putignani L, Guarino MPL. 2019. Fecal and mucosal microbiota profiling in irritable bowel syndrome and inflammatory bowel disease. Frontiers in Microbiology 10.

Loyd-Price J, Abu-Ali G, Huttenhower C. 2016. The healthy human microbiome. Genome Medicine 8.

Martín R, Miquel S, Ulmer J, Kechaou N, Langella P, Bermúdez-Humarán LG. 2013. Role of commensal and probiotic bacteria in human health: A focus on inflammatory bowel disease - microbial cell factories.

Maukonen J, Kolho KL, Paasela M, Honkanen J, Klemetti P, Vaarala O, Saarela M. 2015. Altered Fecal Microbiota in Paediatric Inflammatory Bowel Disease. Journal of Crohn's and Colitis 9: 1088–1095.

Meyer D. 2015. Support vector machines: The interface to libsvm in package e1071.

Mikami A, Ogita T, Namai F, Shigemori S, Sato T, Shimosato T. 2021. Oral administration of flavonifractor plautii, a bacteria increased with green tea consumption, promotes recovery from acute colitis in mice via suppression of il-17. Frontiers in Nutrition 7.

Mucherino A, Papajorgji PJ, Pardalos PM. 2009. k-Nearest Neighbor Classification, Springer New York, New York, NY, 83–106.

N Ananthakrishnan A. 2013. Environmental Risk Factors for Inflammatory Bowel Disease. Gastroenterol Hepatol (N Y) 9: 367–374.

Tolosana-Delgado R, Talebi H, Khodadadzadeh M, van den Boogaart K. 2019. On machine learning algorithms and compositional data. Terrassa, Spain.

Turroni F, Duranti S, Bottacini F, Guglielmetti S, Van Sinderen D, Ventura M. 2014. Bifidobacterium bifidum as an example of a specialized human gut commensal. Frontiers in Microbiology 5.

Ursell LK, Metcalf JL, Parfrey LW, Knight R. 2012. Defining the human microbiome. Nutrition Reviews 70: S38–S44.

Verleysen M, François D. 2005. The curse of dimensionality in data mining and time series prediction. Cabestany J, Prieto A, Sandoval F, editors, Computational Intelligence and Bioinspired Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, 758–770.

Wei Khor I, Yuan Ngiam K. 2019. Big data and machine learning algorithms for healthcare delivery. The Lancet Oncology 20: 262–273.

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with kraken 2. Genome Biology 20.

Wood DE, Salzberg SL. 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15.

Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178: 763–1030.

Zhang YZ, Li YY. 2014. Inflammatory bowel disease: pathogenesis. World J Gastroenterol 20: 91–9.

Zhou K. 2017. Strategies to promote abundance of akkermansia muciniphila, an emerging probiotics in the gut, evidence from dietary intervention studies. Journal of Functional Foods 33: 194–201.

Zhou L, Pan S, Wang J, Vasilakos AV. 2017. Machine learning on big data: Opportunities and challenges. Neurocomputing 237: 350–361.