The Academic College of Tel-Aviv

THE SCHOOL OF COMPUTER SCIENCE

Predicting Fibromyalgia from Gut Microbiome using Machine Learning

Thesis submitted in partial fulfillment of the requirements for the M.Sc. degree in the School of Computer Science of the Academic College of Tel Aviv University

By

Mor Miryam Chako

The research work for the thesis has been carried out under the supervision of

Prof. Adi Shraibman and Dr. Dorit Shweiki

Acknowledgements

I would like to express my deepest appreciation to my supervisors, Prof. Adi Shraibman and Dr. Dorit Shweiki, for their guidance, support, and profound insights throughout the entire journey of this thesis. Their push towards the most impactful and interesting research has been instrumental in shaping the scope and quality of this research.

I express my sincere appreciation to Dr. Amir Minerbi for graciously sharing the dataset and offering valuable insights into his previous studies. His willingness to answer my questions has enhanced the quality of this thesis.

I am also grateful to Dr. Uri Globus, whose valuable recommendations greatly enhanced the writing process of this thesis.

A heartfelt thank you goes to my husband Nir Chako, for the support during the entire chapter of my master’s degree.

Words cannot express my gratitude to my parents. My mother, Haya Naim, may she enjoy a long and healthy life. And my father, Abrahem Naim, may he rest in peace. Thanks to their encouragement, I was able to embark on the fulfilling journey of pursuing a master's degree.

Table of Contents

1. Abstract

Fibromyalgia (FM), a widespread medical problem characterized by chronic pain, cognitive difficulties, fatigue, and sleep disturbances, poses diagnostic challenges. The diagnostic process of FM involves identifying symptoms, ruling out similar diseases, and applying the American College of Rheumatology (ACR) classification criteria, which is a subjective questionnaire.

A groundbreaking study on the relationship between gut microbiome and FM collected a dataset including gut microbiome samples from women, and by using the SVM machine learning algorithm and selecting 72 specific Operational Taxonomic Units (OTUs) achieved an AUC of 87.8%.

In our study, we tried to find relations between gut microbiome and fibromyalgia by using different types of classification algorithms and other machine learning tools on the aforementioned dataset.

We used the Select K Best algorithm to identify the 12 most influential microbiome traits associated with FM. It contributed greatly to our research, achieving 100% accuracy and 100% area under the curve (AUC) with KNN algorithm and 92% AUC with SVM.

In addition, we found that one Operational Taxonomic Unit (OTU), Bacteroides_uniformis_1, is highly significant in the diagnosis of fibromyalgia.

These results offer promising ways to understand the pathophysiology of FM, developing diagnostic aids, and exploring new treatment modalities.

2. Background

2.1.Fibromyalgia

Fibromyalgia (FM) is a chronic disorder characterized by a wide range of symptoms, including –

- Pervasive musculoskeletal pain lasting for at least three months, affecting both sides of the body and above and below the waist.
- Presence of specific tender points on the body, including areas around the neck, shoulders, chest, hips, knees, and elbows.
- Fatigue and Sleep Disturbances: Persistent fatigue, often coupled with sleep disturbances such as insomnia or non-restorative sleep.
- Cognitive difficulties such as problems with concentrating, thinking clearly, and memory (sometimes called "fibro fog") (1).

FM affects approximately 2-8% of the world's population (2) and is more common in women than in men in a proportion of 9:1(3). While the exact cause is unknown, factors such as genetics, infections, physical or emotional trauma, and hormonal changes may contribute to the development of FM (3).

2.1.1. The Current Approach for Fibromyalgia Diagnosis

In the past, the diagnosis of fibromyalgia posed significant challenges due to the lack of specific diagnostic criteria and understanding of the condition. Physicians often relied on clinical judgment based on patient-reported symptoms, such as widespread pain and tender points, which could vary widely between individuals (4). However, in 1990, the American College of Rheumatology (ACR) established diagnostic criteria that included the presence of widespread pain and tenderness at specific anatomical sites. These criteria provided a standardized framework for diagnosing fibromyalgia (5).

A significant shift occurred in 2010 when the diagnostic criteria were revised (6). The tender point count was abandoned, and greater importance was placed on patient-reported symptoms. Additionally, the 2010 criteria introduced severity scales, providing physicians with a tool to assess polysymptomatic distress on a continuous scale (4). Its focus is to try and define the patient's widespread pain index (WPI) by self-reporting on the severity of pain in 19 body areas. Alongside the WPI report, there are other symptoms that the patient is required to define, such as - cognitive difficulties and sleep disturbances (7). This modification allowed healthcare professionals, even those skeptical of the fibromyalgia concept, to diagnose and evaluate patients using an alternative approach (4).

The ACR is widely accepted as the best FM diagnostic tool. Yet, the field of medicine, as a scientific field, prefers to rely on clear evidence from laboratory tests and not on a subjective questionnaire. Relying on patient-reported outcomes introduces a level of uncertainty and may lead to misdiagnoses or delayed diagnosis (7).

2.1.2. Fibromyalgia Diagnosis After Eliminating Alternatives

Medical professionals will often follow a systematic approach to diagnose FM. The patient will go through an extensive testing process to rule out other disorders with similar symptoms, such as rheumatoid arthritis, lupus, and thyroid disorders. This process can lead to a lengthy and frustrating testing period for patients, who may undergo many tests and consultations before fibromyalgia is even considered a possibility (8).

This diagnostic process places fibromyalgia as a diagnosis of exclusion, where it becomes a last resort after other potential illnesses have been ruled out. This approach not only lengthens the diagnostic timeline but adds to the burden on the veterinary services and also adds to the frustration experienced by patients who may feel that their symptoms are not being properly treated or understood.

2.1.3. The Treatment for Fibromyalgia

FM treatment today revolves around symptom management rather than focusing on the underlying cause. Treatment includes a combination of medications such as pain relievers, antidepressants, and anti-seizure medications to relieve the variety of symptoms experienced by people with fibromyalgia.

In addition, patients are offered lifestyle changes, such as regular exercise, stress management and adequate sleep as an important part of the treatment plan.

Another management strategy is supportive therapies, such as physical therapy, counseling, and support groups, which can provide additional assistance in managing the challenges associated with fibromyalgia.

This approach aims to improve quality of life by treating pain, sleep disorders and other associated challenges. Given the variation in symptom severity between patients, a personalized and holistic treatment strategy remains essential to navigating the complexity of this condition (9).

2.2. Gut Microbiome

The gut microbiome refers to the complex and diverse community of microorganisms, including bacteria, viruses, fungi, and other microbes, found in our digestive tract. The microbial population residing in the gastrointestinal (GI) tract is estimated to surpass 10^{14} microorganisms, comprising approximately tenfold more bacterial cells than human cells and over a hundredfold greater genomic content (microbiome) compared to the human genome (10).

In the picture below you can see the abundance of microorganisms in the gut:

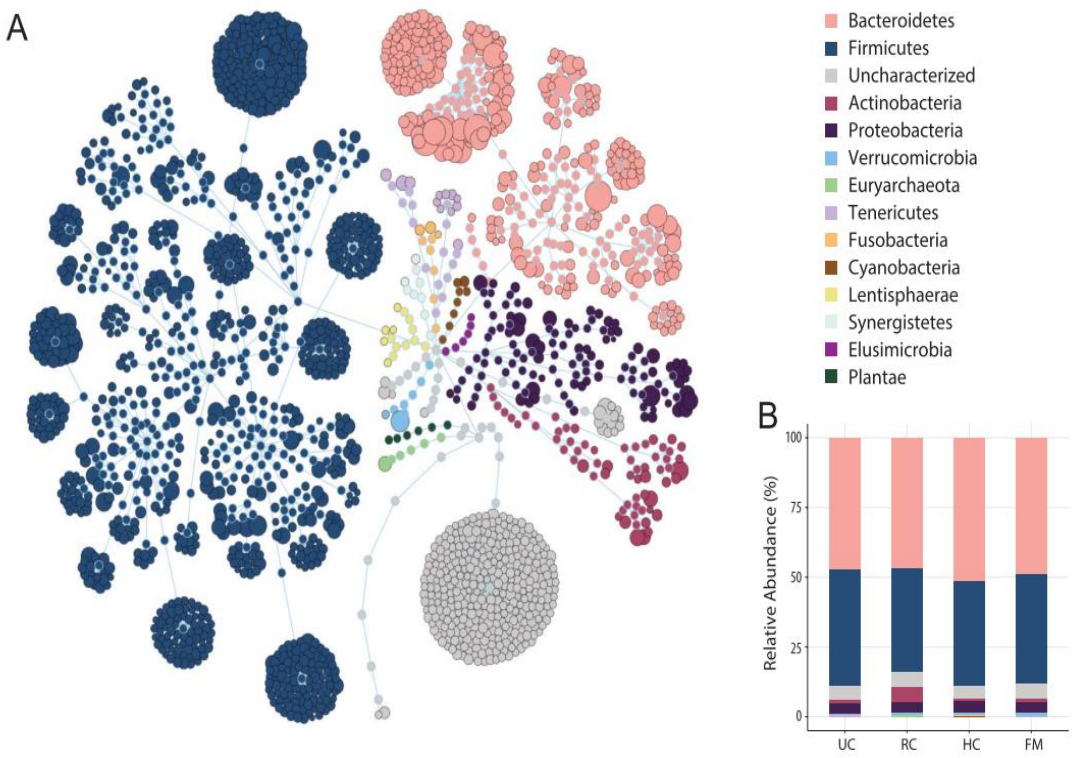


Figure 1. Flower diagram of 1620 Operational Taxonomic Units colour coded by phyla. (Source: Altered Microbiome Composition in Individuals with Fibromyalgia [11])

2.2.1. Factors Influencing the Gut Microbiome

Changes in the composition of the gut microbiome are mainly influenced by diet but also by ageing, diet, antibiotic usage, the administration of drugs, prebiotic and probiotic supplementations, surgeries and non-surgical treatments, pregnancy, the length of the gestational period, sex and sexual preference, post menopause, exposure to dust and chemicals, circadian rhythm, smoking, geographical origin, heritability, and area of residence (12). In addition, the method of delivery was also found to be influential, Infants delivered vaginally had higher amounts of bacteria in their gut compared to infants delivered by Cesarean section (13).

2.2.2. The Gut-Brain Axis

In recent years, medical research has made significant progress in unraveling the intricate interplay between the gut and the brain. It's been discovered that gut bacteria produce bioactive molecules and metabolites that directly influence brain function and the nervous system. This connection, known as the gut-brain axis, serves as a vital two-way communication system between the digestive system and the central nervous system, regulating a wide array of physiological processes and behaviors.

This sophisticated network comprises multiple pathways, including neural, hormonal, and immunological signals, facilitating interactions among the gut microbiota, the enteric nervous system (ENS), and the brain. Signals travel bidirectionally through

nerves, hormones, and neurotransmitters within the gut itself. Research has shown that the health of the intestines, the hormones they produce, and the gut microbiota can impact mood, stress levels, fatigue, immune function, hunger, satiety, and overall well-being. Conversely, the brain can influence intestinal function; stress signals from the brain, for instance, can affect gut functionality.

Emerging studies continue to underscore the profound influence of the gut microbiota on brain function and behavior. Changes in gut bacteria composition have been linked to various neurological and psychiatric disorders, including depression, anxiety, and autism spectrum disorders. This growing body of research highlights the bidirectional nature of gut-brain communication, shedding light on its diagnostic and therapeutic potential in enhancing human health and well-being. (14).

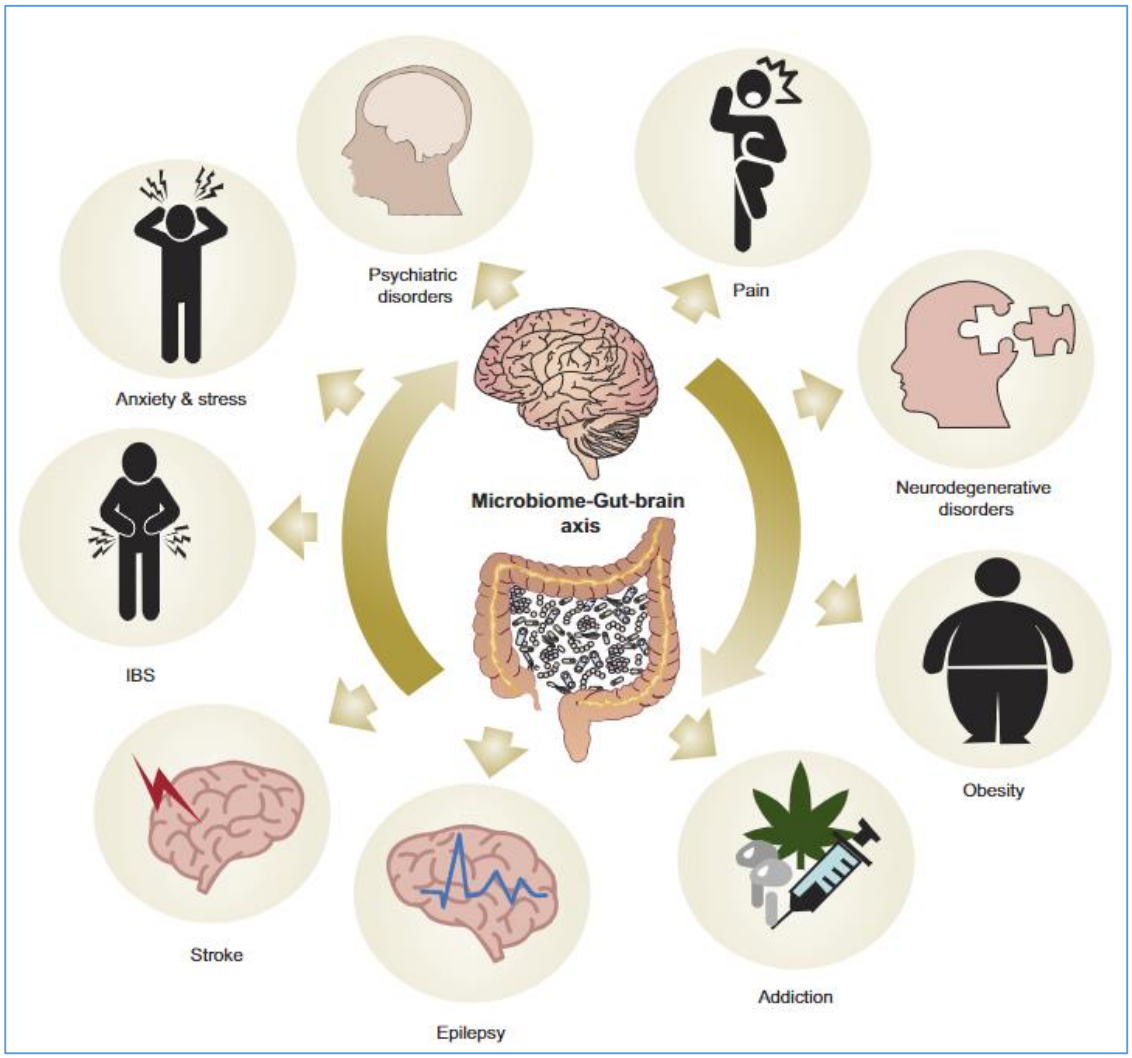


Figure 2. An outline illustrating the variety of disease and disease processes the microbiota are currently implicated in; examples include psychiatric and neurodegenerative disorders, pain, stress, irritable bowel syndrome (IBS), stroke, addiction, and obesity (Source: The Microbiota-gut-brain Axis [14])

2.2.3. The influence of the microbiome on health

The gut microbiome plays a key role in everything related to digestion, metabolism, and immune system modulation. Due to the gut-brain axis, the connection between the digestive system and the central nervous system, recent studies show that the gut microbiome can have an impact on the broader aspects of health (15), including mental health such as depression, anxiety, and fatigue (13), immune system function, and chronic diseases (16).

Studies have found a connection between the gutStudies have found a link between the gut microbiome and the risk of stomach cancer, breast cancer and prostate cancer, which highlights the complex relationship between cancer and the human microbiota. In addition, changes in the intestinal microbiome may endanger the integrity of the barrier of the digestive system in inflammatory bowel diseases (IBD), affect tight junctions between epithelial cells and disrupt barrier function. Intestinal dysbiosis has also been found to be associated with cardiovascular disease (17).

Since the composition of the gut microbiome is mainly related to the choice of diet, recognition of the relationship between the gut microbiome and various diseases opens the possibility of targeted treatments by modulating the composition of the microbiome through specific nutrition.

The following diagram shows the effect of the gut microbiome on chronic diseases:

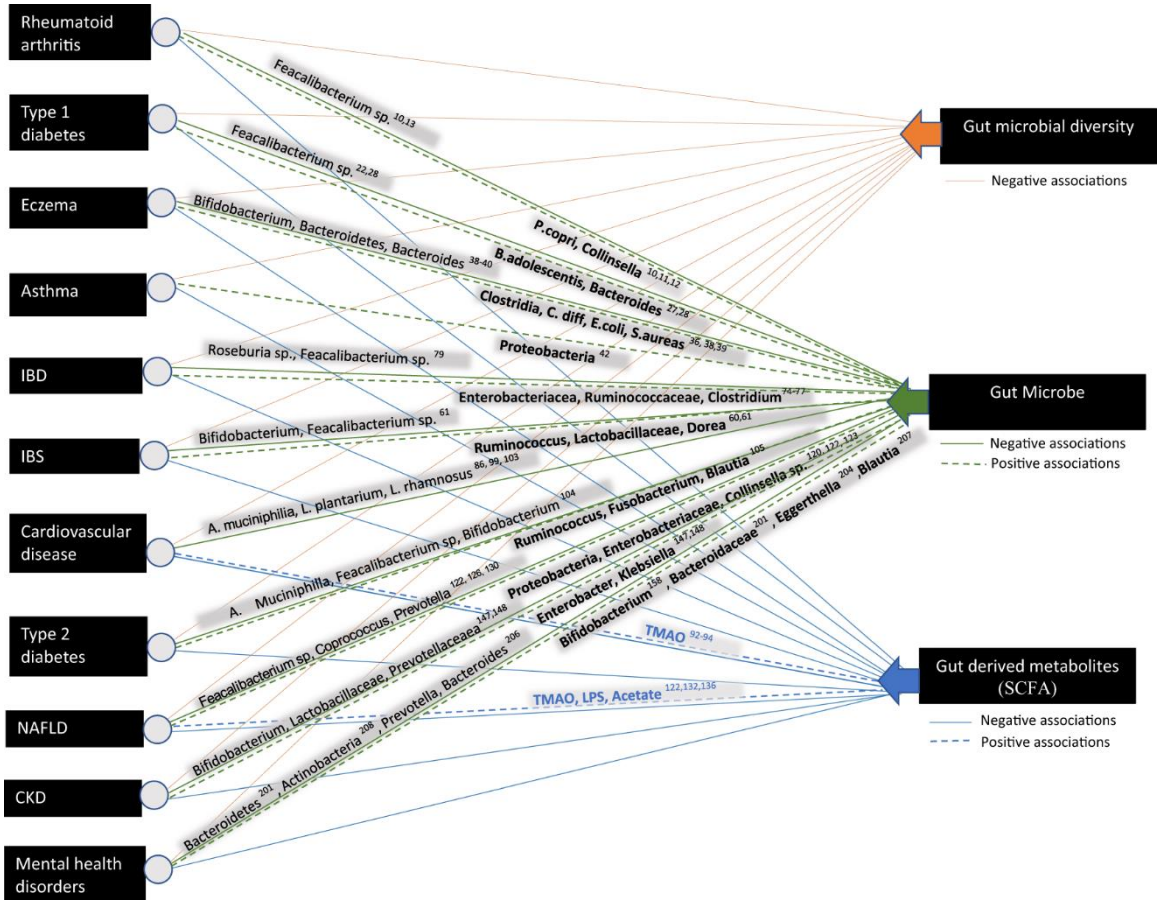


Figure 3. Schematic representation of the association of the composition of the gut microbiome and gut-derived metabolites with chronic diseases (Source: Role of the gut microbiome in chronic diseases: a narrative review [16])

2.3. Machine Learning

Machine learning (ML) is one of the many branches of AI. ML is the science of developing computational algorithms and statistical models that computer systems use to perform complex tasks without explicit instructions, but instead rely on patterns and inference. These algorithms and models are designed to imitate human intelligence by learning from the surrounding environment and developing the ability to deal with various challenges.

Traditionally, there are three main approaches to perform a machine learning procedure – Supervised, unsupervised, and reinforcement learning.

Supervised learning is where the computer is presented with a guided learning path by receiving examples with input and their outputs, and by using this data, the computer should learn how to infer the output for future inputs.

Unsupervised learning deals with unlabeled data. The algorithm explores the data's inherent structure, patterns, or relationships without explicit guidance.

Reinforcement learning is where a computer tries to achieve a specific goal. This approach involves training an agent to make decisions by receiving rewards or penalties, applicable in sequential decision-making tasks such as game playing and autonomous systems (18).

2.4.Machine Learning for Health Care

There is a significant rise in the popularity of utilizing ML techniques for health care (19). Around 86% of healthcare organizations incorporate ML solutions, and over 80% of healthcare organization leaders have formulated an artificial intelligence (AI) strategy. ML offers advantages stemming from machine capabilities surpassing those of humans, enabling algorithms to derive medical insights beyond traditional data analysis methods. Unlike traditional hypothesis-driven statistical analysis, ML prioritizes predictive model accuracy. Furthermore, human error, often attributed to limited short-term memory, underscores the need for ML's systematic approach. (20)

Three of the most common ML applications for medical needs:

2.4.1. Machine learning of Medical Images

Modern medical images, which are digital, present challenges in their effective use in healthcare. Despite these challenges, medical imaging techniques provide visual representations of the human body's interior, aiding diagnosis, analysis, and medical interventions. This approach helps avoid or minimize the need for exploratory surgery, reducing associated risks such as infections and strokes. While traditionally assessed by trained professionals, such as physicians or radiologists, this clinical standard is prone to human error and expensive, often requiring years or decades of experience to achieve a level of understanding which can consistently assess these images.

Machine learning's demonstrated capabilities, as showcased by Andrew Ng, have led to its early adoption in healthcare, particularly in the assessment of medical images.

Medical imaging is now the preferred tool for initial diagnosis in the clinical setting, specifically in the detection of lesions such as those commonly found in mammograms, brain scans, and other body scans (19).

2.4.2. Natural language processing of medical documents and literature

Electronic medical records (EMR) have become standard in hospitals, requiring intricate digital infrastructure to unify health data and improve hospital efficiency and patient outcomes. Yet, transitioning from physical to electronic documentation, particularly with historical records, presents challenges, necessitating laborious and costly manual inputting. Natural language processing (NLP), a type of machine learning, offers a solution by rapidly scanning documents and extracting information from free text, including handwritten notes. While structured forms ease language processing, challenges like missing or inaccurately categorized data persist. Similarly, developing an enhanced clinical decision support (CDS) system using old patient records aims to leverage medical knowledge for individual patient care, requiring integration of specialized NLP systems. Compiling scientific research into centralized repositories also faces challenges due to the overwhelming volume of papers across multiple journals, potentially leading to overlooked promising treatments (19).

2.4.3. Machine learning in genetics for the prediction and understanding of complex disease

The rapid growth of genetic information and technologies since 2008 has presented significant challenges in handling exponentially increasing data. Advances in genetic sequencing, particularly Next-Generation Sequencing (NGS) technologies, have accelerated the speed and reduced the cost of sequencing whole human genomes. Despite this progress, deciphering the complexities of the human genome, with its interconnected structure and variations between individuals, remains a challenge.

Machine learning has emerged as a powerful tool to uncover patterns and trends in genetic data. By leveraging vast amounts of genetic information, machine learning holds the potential to predict disease risks accurately, including cancer and Alzheimer's disease. Additionally, it offers insights into genetic links to mental illnesses like schizophrenia and bipolar disorder (19).

2.5. Classification Algorithms for Medical Diagnosis

The medical diagnosis process can be regarded as a classification problem, which includes the identification and classification of a set of medical characteristics for specific medical diagnoses. This can be achieved by supervised learning with a dataset that includes features and appropriate predictions, or results. The result of the training will be a model that can predict the result given the medical characteristics of a new patient (21).

There are several different algorithms that implement this approach, in this study we used the following classification algorithms: K-Nearest Neighbors (KNN), Support Vector Machines (SVM), XGboost, CatBoost, Extra Trees Classifier and Logistic Regression.

2.5.1. K-Nearest Neighbors (KNN)

The base concept of KNN is the assumption that similar results can be found next to each other based on their features. There is an active model which will divide the different points on a graph into different classes based on the training phase.

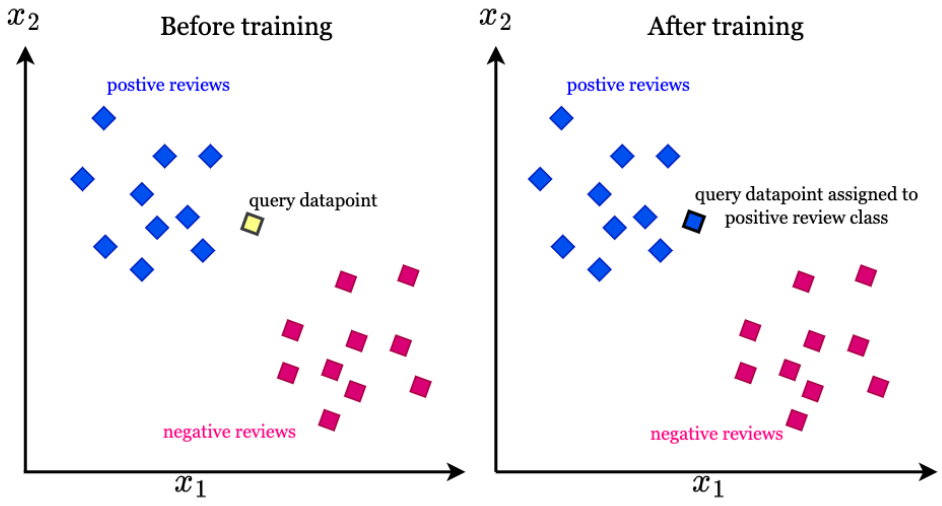


Figure 4. KNN query datapoint before and after training (Source: K-Nearest Neighbors Algorithm [22])

In KNN the "K" is a variable that determines the number of nearest neighbors takes into account when making a prediction. A smaller "K" might lead to a more sensitive classification model, while a larger value might offer a more subtle decision threshold with the potential overlook of finer details in the datapoint.

The similarities between the datapoint to the class are calculated using the Euclidean distance. In 2 dimensions the distance between point A (x_1,y_1) and B (x_2,y_2) will be –

$$D = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2}$$

Closer points will be considered as more similar and as a result will be classified as part of the same category (23).

2.5.2. Support Vector Machines (SVM)

SVM seeks to find the most effective separation between classes. It does it by putting an emphasis on the data points that are more crucial to the decision making. The "Support Vectors" are the ones that define the separations between the classes and are essentially the optimal hyperplanes in the space to make this classification (24).

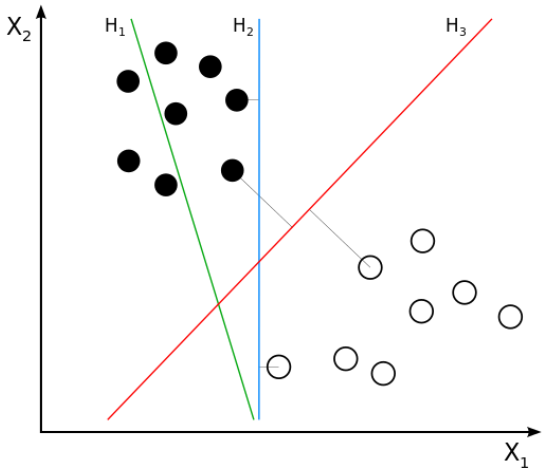


Figure 5. SVM hyperplanes (Source: Wikipedia Support Vector Machine [25])

In the above diagram, H1 does not separate the classes. H2 does, but only by a small margin. The SVM algorithm strives to identify the optimal decision boundary, which is H3 that separates the classes with the maximal margin.

2.5.3. eXtreme Gradient Boosting (XGBoost)

XGBoost is a code library that includes optimization of the distributed gradient boosting. XGBoost is designed to be as flexible and efficient.

It relies on the Gradient Boosting framework and provides a parallel tree boosting. Another essential feature of XGBoost is that it is supports missing values by default. In tree algorithms, branch directions for missing values are learned during training (26).

2.5.4. Categorical Boosting (CatBoost)

CatBoost is a powerful open-source gradient boosting library, designed by Yandex for efficient handling of categorical features in machine learning. With innovative techniques optimizing training speed and accuracy, CatBoost is a robust choice for diverse predictive modeling tasks. Supporting both classification and regression, it minimizes the need for extensive preprocessing, and its default parameters often yield competitive results without intensive hyperparameter tuning (27).

2.5.5. Extra Trees Classifier

The Extra Trees Classifier, a variant of the Random Forest algorithm, stands out for its high-performance and efficiency in handling diverse datasets. By introducing additional randomness during tree building, it enhances generalization and minimizes overfitting. An ensemble learning method, it combines multiple decision trees to deliver robust predictions, making it a valuable tool for various classification tasks (28).

2.5.6. Logistic Regression

A fundamental algorithm in machine learning that excels in binary classification tasks. Despite its name, it is used for classification, not regression. Leveraging a logistic function, it estimates the probability of an instance belonging to a particular class. Widely adopted for its simplicity and interpretability, Logistic Regression serves as a go-to method for understanding relationships between features and predicting outcomes in diverse fields (29).

2.6. Feature Selection

Feature selection in ML is a method being used to find the most relevant features from a larger dataset, aiming to enhance the algorithm's performance and prevent overfitting by reducing the number of features considered.

There are a few different types of feature selection methods (30), including –

- Filter Methods: Selection by assessing if a feature is relevant based on statistical evaluations.
- Wrapper Methods: Selection by evaluating the performance of an ML algorithm using different assembly of features.
- Embedded Methods: Selection as part of the model training itself.

2.6.1. Select K Best

Select K best is at the top of the most used feature selection methods. It's part of the "Filter Methods" family, which means that the selection of the best K features is done regardless of any specific ML algorithm. It relies on statistical metrics to assign the different features a score and rank them.

The statistical evaluation of the relevance of each feature is done by different measures such as Analysis of Variance (ANOVA) and chi-squared test (chi^2). The "K" stands for the number of features with the highest score associated with relevance to the classification process (31).

2.7. ROC - AUC

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the True positive Rate (TPR) against False Positive Rate (FPR) at various threshold values and essentially separates the 'signal' from the 'noise' (32).

The TPR is defined as follows (where TP-True Positive, and FN-False Negative):

$$TPR = \frac{TP}{TP + FN}$$

The FPR is defined as follows (where FP-False Positive, and TN-True Negative):

$$FPR = \frac{FP}{FP + TN}$$

In other words, it shows the performance of a classification model at all classification thresholds. By looking at this graph, we can understand how good the model is and choose the threshold that gives us the right balance between correct and incorrect predictions.

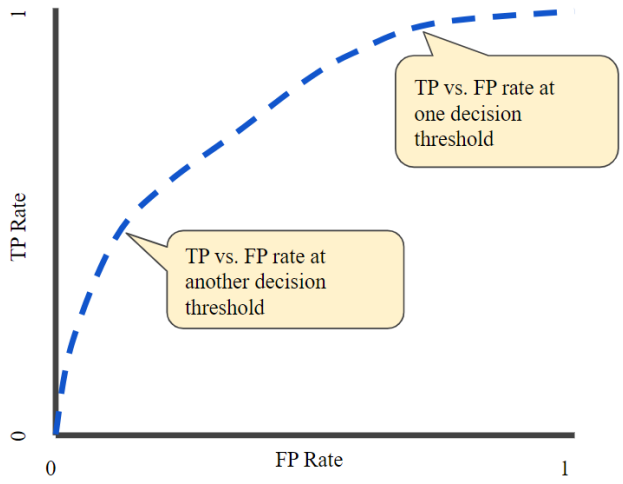


Figure 6. ROC curve - TP vs. FP rate at different classification thresholds (Source: Classification: ROC Curve and AUC [32])

The Area Under the Curve (AUC) is the measure of the ability of a binary classifier to distinguish between classes and is used as a summary of the ROC curve.

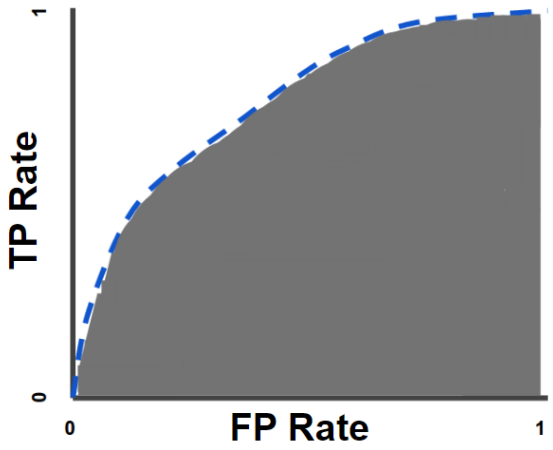


Figure 7. Area under the curve - AUC (Source: Classification: ROC Curve and AUC [31])

Another way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example (32).

3. Problem Description

The diagnosis of FM remains a challenge in the medical field, mainly due to the lack of objective tests.

Current diagnostic approaches to fibromyalgia raise the need for more objective measures and biomarkers to improve the accuracy and efficiency of diagnosis. The reliance on subjective patient reports, along with a diagnostic process that places fibromyalgia as a diagnosis of exclusion (4), highlights the limitations in our understanding of this complex condition. Research efforts should be directed at identifying specific biomarkers or other techniques that can objectively diagnose FM, reduce reliance on exclusion of other possibilities, and provide a more accurate diagnosis in a shorter time.

4. Aims & Objectives

4.1. Main goal

Our main goal in the research is to diagnose FM in women using gut microbiome data to create a direct objective test of FM.

4.2. Sub-goals

In addition to the main goal, we tried to achieve several goals:

- To find the connection between microbiome and fibromyalgia, if it exists, in order to open new possibilities of treatment for this disease.
- To choose the smallest number of the most significant Operational Taxonomic Units (OTUs) out of 1620 OTUs, which may be able to reduce the cost of the test and optimize the treatment.
- Comparison of the different classification algorithms using a dataset that contains data on the gut microbiome of women with FM and a control group. This comparison will make it possible to get an idea of which algorithm is most suitable for fibromyalgia.

5. Related Work

In this section, we delve into the work related to diagnosing FM, examining potential advances, both clinical and technological, that may offer new perspectives and solutions for the problems arising from the lack of an objective diagnosis for this disease.

5.1. Fibromyalgia and Biomarkers

Biomarkers are measurable indicators of biological processes or disease states that can aid in diagnosis, prognosis, and monitoring response to treatment.

Fibromyalgia, despite its prevalence and impact on quality of life, remains challenging to diagnose and manage due to its subjective symptoms and the lack of specific biological markers. In the case of FM, the identification of reliable biomarkers may revolutionize clinical care by providing objective measures to support diagnosis and guide personalized treatment strategies (12).

Some academic studies have investigated the neural markers associated with FM using imaging techniques such as functional MRI (fMRI) and positron emission tomography (PET) to try and correlate brain activity in FM patients. Currently, although there may be reliable neural biomarkers, their clinical applicability is limited (35, 36), partly because fMRI is an expensive test.

Significant biomarkers that stood above others were features related to the metabolic process. especially metabolic profiles with indications regarding the presence of a specific gut microbiome (37, 38, 39). Most studies in this field indicate a correlation between the gut microbiome and FM. This line of research seems relevant since it is known that the gut-brain axis has an effect on several FM symptoms such as - pain and sensitivity (40), mood and cognitive function (41), sleep disorders (42) and IBS (14).

5.2.The Association Between Gut Microbiome and Fibromyalgia

Metabolomic analysis and Isotope Ratio Mass Spectrometry (IRMS) were employed to scrutinize the metabolic profiles of patients diagnosed with rheumatoid arthritis (RA), osteoarthritis (OA), and fibromyalgia (FM). Notably, a significant discovery emerged indicating distinct metabolic variations, particularly in the metabolism of tryptophan, among patients with fibromyalgia compared to those with RA and OA. Tryptophan, an essential amino acid, serves as a precursor for neurotransmitters like serotonin and melatonin. Serotonin contributes to mood regulation, sleep, and appetite control, while melatonin regulates the sleep-wake cycle. The altered metabolism of tryptophan observed in fibromyalgia suggests potential dysregulation in serotonin and melatonin pathways, implicating it as a potential biomarker for the condition (37).

Furthermore, this analytical approach offers several advantages over traditional methods, particularly in terms of speed, sample preparation, and cost-effectiveness. Compared to conventional neural markers, metabolomic analysis coupled with IRMS enables rapid examination of various sample types, necessitating minimal preparation and requiring only small sample volumes. Additionally, the implementation of

specialized software for pattern recognition has demonstrated remarkable accuracy in distinguishing fibromyalgia from RA and OA (34).

5.3. Microbiome-based machine-learning identification of fibromyalgia patients

Differentiating one disease from another was one step forward in terms of associating metabolic features with FM. Another key step was the use of ML algorithms to find significant differences in the composition of the microbiome, especially in several bacterial taxa, when comparing patients with FM to a control group.

A groundbreaking study by Amir Minrabi et al., presented the change in the composition of the microbiome in people with FM. In this study, data were rigorously collected from both FM patients and healthy female controls, using stool samples and comprehensive questionnaires. Through whole-genome sequencing, researchers aimed to unveil microbial signatures linked to FM, pinpointing significant differences in bacterial profiles compared to control participants (11).

The researchers used machine learning algorithms to investigate the diagnostic utility of microbiome composition alone. They used DESeq2 to reduce the number of OTUs from 1620 to 72 and then used LASSO and SVM to achieve a ROC-AUC classification accuracy of 87.8%, effectively discriminating between FM patients and controls. In addition, this study presented the differences between the OTUs of the FM sand group versus the control group.

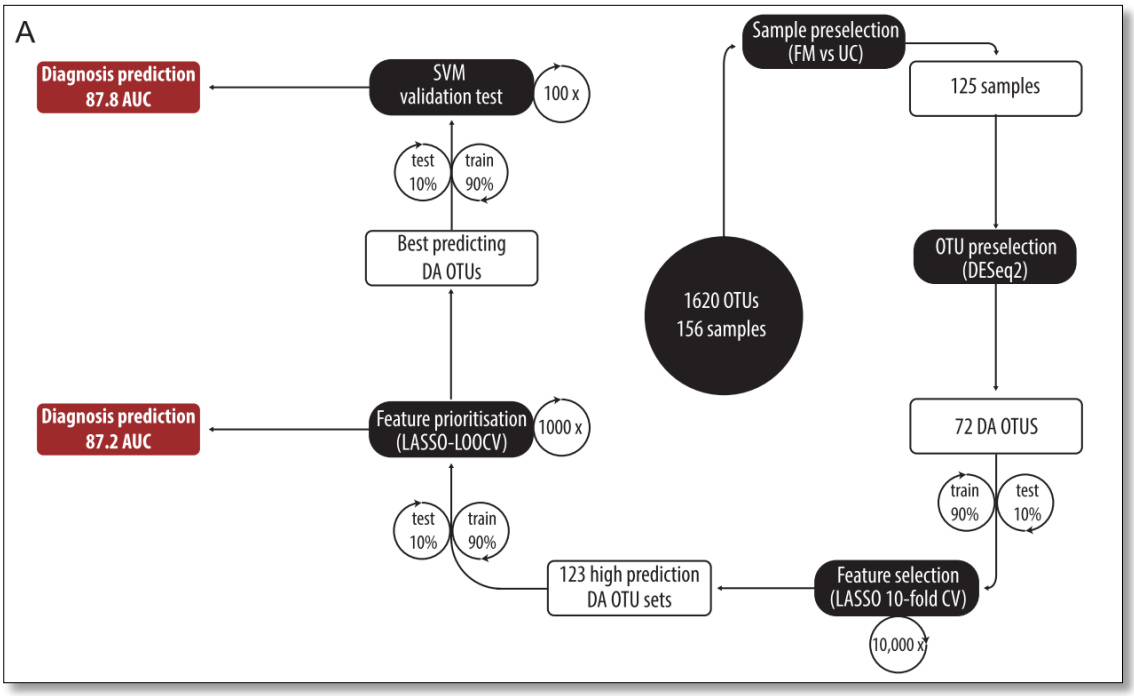


Figure 8. Description of the research process that includes the selection of OTUs, classification and results. The best prediction accuracy in LASSO was used in Support Vector Machine (SVM) (Source: 'Altered Microbiome Composition in Individuals with Fibromyalgia [11])

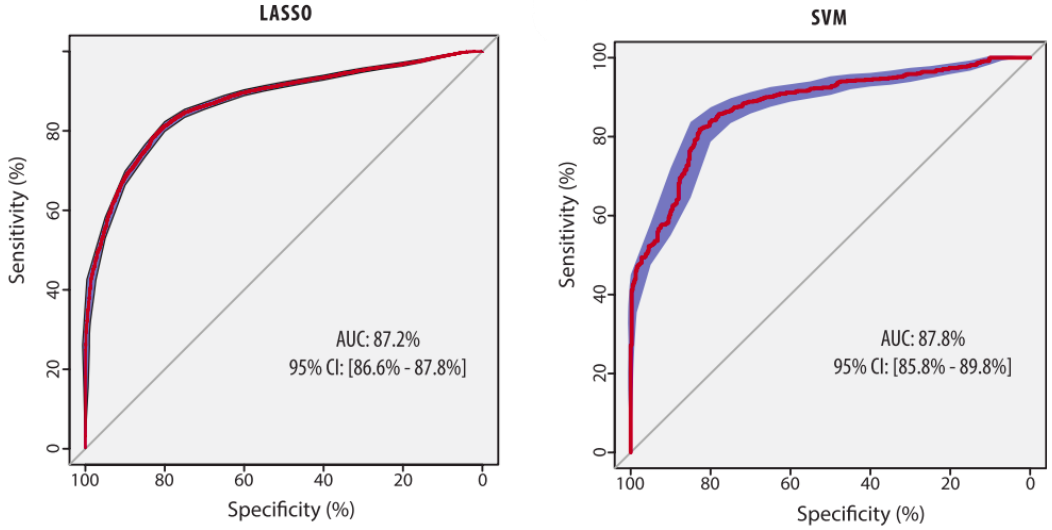


Figure 9. ROC-AUC of the LASSO and SVM results (Source: 'Altered Microbiome Composition in Individuals with Fibromyalgia' [11])

These findings not only contribute to our understanding of the pathophysiology of FM but also lay the foundation for future investigations. The potential development of diagnostic aids and novel treatment methods holds promise for enhancing the management of FM, offering new perspectives on personalized medicine tailored to the individual's gut microbiome profile (11).

6. Methodology

In order to achieve our goals, we use tools from the world of machine learning. In this chapter we will expand on the process of processing the dataset, the use of the different feature selection and classification algorithms, and strategies for optimizing the results.

6.1. Dataset Description

The dataset we use in this study was collected and used in a previous study by Amir Minrabi et al. (11). This dataset was generated by analyzing 156 questionnaires and stool samples from 77 FM patients and 79 controls (48 healthy women, 20 men and 11 women with a family history of fibromyalgia). A total of 1620 operational taxonomic units (OTU) were identified and then reorganized with different taxonomic levels – domain, phylum, class, order, family, genus and species.

The data set consisted of 2 tables. One table contained only gut microbiome data with 1620 rows, each of which is a specific OTU, labeled for the different taxonomic levels described earlier and the number of units from each OTU for each sample. The second table contained the results of the questionnaires and the diagnosis of FM, where 1 is for a positive diagnosis of FM and 0 means that there was no indication of FM.

In order to use classification algorithms, we merged these 2 tables into one containing the samples as the rows and the columns as the microbiome features, the last column

being of course the diagnostic column itself. In addition, in order to make the results more accurate, we only used 125 samples that included the 48 healthy women and the 77 women with FM only.

At the end of the process, the processed table contained data of 1620 OUTs and diagnosis for 125 samples.

6.2. Algorithms and Tools

In our research we've used a variety of algorithms and tools, including – KNN, SVM, Extra Trees Classifier, Logistic Regression (sklearn), XGboost and CatBoost. Our goal was to find the best solution for FM diagnosis based on ML. To do so, we tried to fine-tune the algorithms parameters and the dataset's features.

6.2.1. Parameters Optimization

Each algorithm has its own parameters. In order to find the best ones, we ran a set of tests with all of parameters' combinations.

Algorithm	Parameters	Tested Values	Meaning
KNN	K	1 – 5	The number of neighbors
	Threshold	0.1 – 1	
SVM	Kernel	Polynomial (degree 1-5), Linear, Sigmoid, rbf	Set of mathematical functions
	regularization parameter (C)	0.1, 1, 10, 20, 30, 50, 100	Control error tolerance and smoothness
	Gamma	Scale, auto	
Extra Trees Classifier (28)	n_estimators	10, 50, 100, 150, 200	Number of trees in the forest
	Criteria	gini, entropy, log_loss	The function to measure the quality of a split
Logistic Regression (29)	Penalty	none, l1, l2	Specify the norm of the penalty
	Solver	lbfgs, liblinear, newton-cg, sag, saga	Algorithm to use in the optimization problem
	C	0.1, 0.5, 1, 10, 20, 25, 30, 50, 100	Inverse of regularization strength
XGboost (43)	Random state	42	
	tree method	Approx.	Gradient Boosting for efficient ensemble learning
	Booster	dart	Type of model
	Alpha	1-10	the L1 regularization term on weights
	Eta	0.01-1	Learning rate
	Gamma	0.01-1	Minimum loss reduction required to make a further

			partition on a leaf node of the tree.
CatBoost (27)	Iteration	1 - 1000	Number of iterations
	Depth	1-10	he maximum depth of the trees in the gradient boosting model
	learning rate	0.01-1	Learning rate used for training
	loss function	logloss	

6.2.2. Feature Selection

Another step in the FM prediction optimization process was the use of the SelectKBest feature selection algorithm. We used it to better refine the data that is more influential and may have a stronger relationship with FM.

We used SelectKBest with different parameters, including K and different scoring functions. Scoring functions are statistical functions like chi-squared and f-statistic. The K parameter determines the number of top features. In our research we tested the different algorithms both with the full dataset and with the top K selected features.

The parameters we used were:

Parameters	Values
K	1-72
Scoring functions	chi2, f_classif, mutual_info_classif from the python sklearn (scikit-learn) ML library

The select k best algorithm uses a verity of statistical tests from which the k features with the highest scores will be chosen. In our research we used chi-square (chi^2, or χ^2) as the statistical measure.

chi-square is a common statistical test for feature selection, especially in the context of categorical data and classification tasks. The chi-square test is used to determine whether there is a significant relationship between two categorical variables. It is based on the difference between the expected and observed frequencies in a contingency table.

In the context of feature selection, each feature is treated as a categorical variable, and the classes of the target variable are used to create a contingency table.

6.3.ROC-AUC Influence

In our research, we utilized the ROC-AUC output to both optimize the classification algorithms results and to compare the suitability of the different algorithms to the problem we are trying to facing with. The optimization was done by choosing the threshold that gives us the best results. The threshold that will give us the best results is the one the is defined by the spot on the scheme with the farthest length above the diagonal.

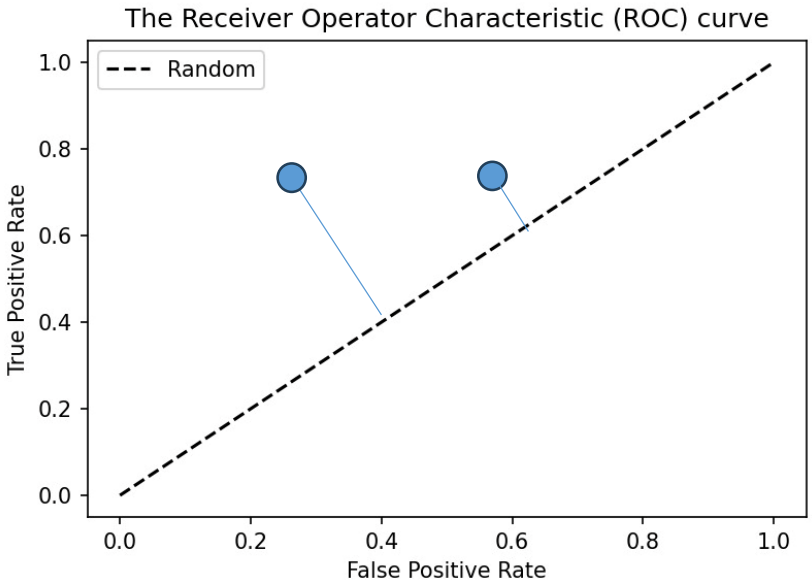


Figure 10. The Receiver Operator Characteristics (ROC) curve

The second and final use of ROC-AUC in our research is as a scoring tool to suggest the best set of algorithm-parameters-features to solve the FM diagnostic problem.

6.4. Accuracy

Accuracy is a fundamental metric in classification algorithms, measuring the model's ability to correctly predict the class labels of the data instances. It represents the exact percentage that the classification algorithm is right on the data set it receives as a test.

In fact, accuracy quantifies the effectiveness of the model in making correct predictions in all classes. High accuracy indicates that the model makes accurate predictions most of the time, while lower accuracy indicates that the model has difficulty correctly classifying cases.

To maximize the accuracy for each classification algorithm, we used all the tools detailed in this chapter.

6.5. Dataset Taxonomic Levels Handling

Other than the use of feature selection algorithms, we also tried to find an association of FM with specific gut-microbiome taxonomic levels. To do so, we tried to aggregate data by its taxonomic level with 2 methods – OTU enumeration and OTU summarization. By doing this, we strived to find a strong association between a specific gut-microbiome taxonomic level to FM patients.

7. Results

Working on the different research pathways led us to a variety of results. In this section we will put the focus on the most significant results.

7.1. Select 12 Best OTUs

By using Select K best with *k=12* we receive this list of the top 12 Operational Taxonomic Units (OTUs)

1	Prevotella_copri_1	5	Alistipes_finegoldii_2	9	Alloprevotella_1
2	Prevotella_12	6	Bacteroides_MS_3	10	Bacteroides_3
3	Bacteroides_uniformis_1	7	Parabacteroides_merdae_3	11	Prevotella_4
4	Bacteroides_dorei_1	8	Ruminococcaceae_MG_1	12	Akkermansia_muciniphila_1

Using the KNN classification algorithm on these 12 OTUs, we achieved accuracy of 100% on a train-test split 90% - 10%. For comparison with different algorithms:

Algorithm	Parameters	Accuracy
KNN	k=2 Threshold=0.5	100.00 %
Polynomial SVM	C=20 Degree=3	84.62 %
RBF SVM	C=50 Gamma = 'scale'	84.62 %
XGboost	eta = 0.1 gamma = 1 tree_method = "approx" booster = "dart"	76.92 %
Extra Trees Classifier	Criterion = "gini"	76.92 %
CatBoost	iterations=2 depth=2 learning_rate=0.1 loss_function='Logloss'	76.92 %
Logistic Regression	Penalty = 'l2' Solver='saga' C=0.5	69.23 %

ROC and AUC:

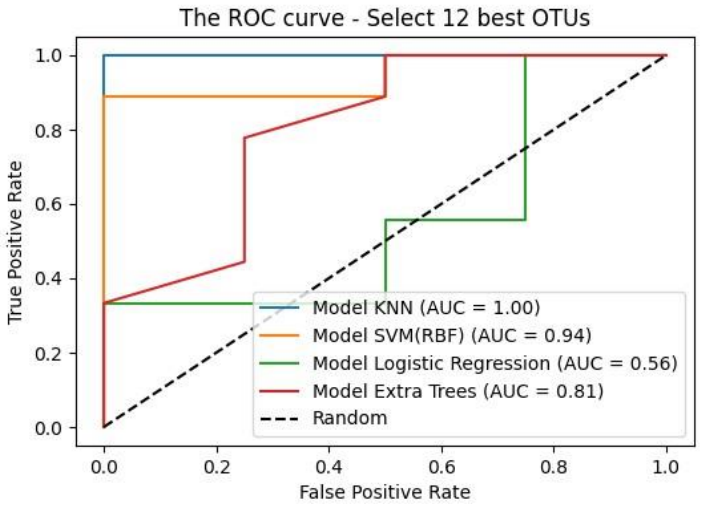


Figure 11. Graphic representation of the ROC curve and the respective AUC values of 'Select K Best' where k=12 OTUs predicting FM using different classification algorithms different classification algorithms

7.2. T-Test for the selected 12 best OTUs

To determine if there is a significant difference between the means of the two groups (FM patients and the control group) and how they are related, we used a T-test and FDR correction for each of the top 12 OTUs and found that the Bacteroides_uniformis_1 OUT is statistically significant using FDR of 0.05.

Rank	Out	Original P Value	Critical Value	Benjamini-Hochberg Adjusted P value	Significant using an FDR of 0.05
1	Prevotella_copri_1	0.00716	0.0041667	0.04476	No
2	Bacteroides_uniformis_1	0.00746	0.0083333	0.04476	Yes
3	Prevotella_12	0.02138	0.0125	0.08552	No
4	Akkermansia_muciniphila_1	0.03601	0.0166667	0.09108	No
5	Parabacteroides_merdae_3	0.03795	0.0208333	0.09108	No
6	Prevotella_4	0.08475	0.025	0.1695	No
7	Alistipes_finegoldii_2	0.11364	0.0291667	0.1948114	No
8	Alloprevotella_1	0.20311	0.0333333	0.2994436	No
9	Ruminococcaceae_MG_1	0.25784	0.0375	0.2994436	No
10	Bacteroides_MS_3	0.26613	0.0416667	0.2994436	No
11	Bacteroides_dorei_1	0.27449	0.0458333	0.2994436	No
12	Bacteroides_3	0.43008	0.05	0.43008	No

7.3.Select 3 Best OTUs

Motivated to find a tighter group of the most significant OTUs associated with FM we used select K best with *k=3* and received these 3 top features:

1	Prevotella_copri_1
2	Prevotella_12
3	Bacteroides_uniformis_1

Algorithm comparison in this case is as follows:

Algorithm	Parameters	Accuracy
KNN	k=2 Threshold=1	84.62 %
SVM (RBF kernel)	Threshold=0.59368 C=50	84.62 %
SVM (Polynomial)	Degree = 3 C = 30	76.92 %
XGboost	eta = 0.1 gamma = 1 tree_method = "approx" booster = "dart"	76.92 %
Logistic Regression	solver = 'sag' C=0.1	69.23 %
ExtraTreesClassifier	Criterion = "gini"	61.54 %
CatBoost	iterations=2 depth=2 learning_rate=0.1 loss_function='Logloss'	69.23 %

ROC and AUC:

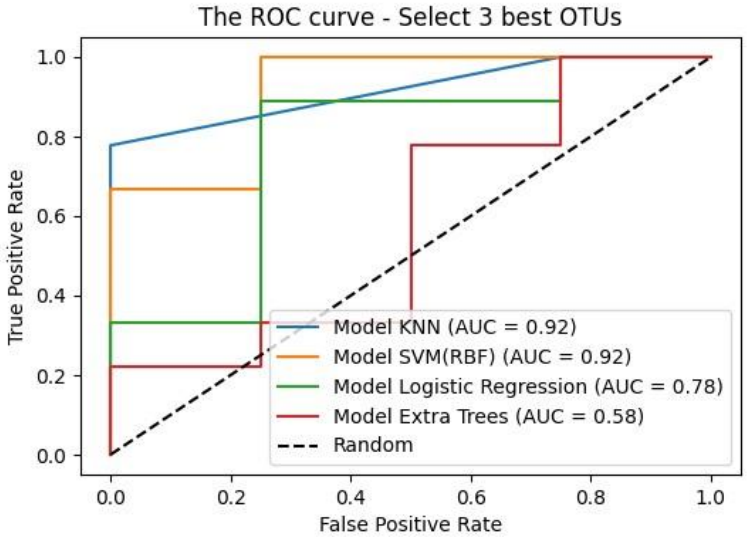


Figure 12. Graphic representation of the ROC curve and the respective AUC values of 'Select K Best' where k=3 OTU predicting FM using different classification algorithms different classification algorithms

7.4. Isolation And Examination of Each of The Top 3 OTUs

To better understand the association between the top 3 OTUs and FM, each one on its own, we tried to use the individually to predict FM.

7.4.1. Prevotella Copri 1

The best accuracy Prevotella Copri 1 achieved was 76.92% by using the polynomial SVM algorithm.

7.4.2. Prevotella 12

The best accuracy Prevotella 12 achieved was 69.23% by using the SVM algorithm with RBF kernel.

7.4.3. Bacteroides Uniformis 1

While the 2 OTUs Prevotella Copri 1 and Prevotella 12, when tested individually did not achieve impressive results, the best accuracy achieved by Bacteroides_uniformis_1 was 84.62% by using KNN algorithm (k=2, threshold=0.5), with ROC/AUC:

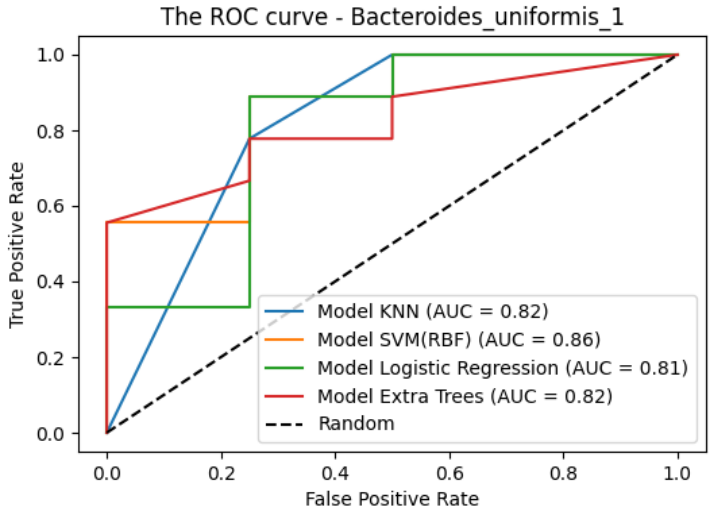


Figure 13. Graphic representation of the ROC curve and the respective AUC values of the Bacteroides_uniformis_1 OTU predicting FM using different classification algorithms

7.4.4. Prevotella copri 1 and Bacteroides uniformis 1

By using both Prevotella copri 1 and Bacteroides uniformis 1 with KNN (k=2, threshold=1) and SVM RBF (C=50), we get accuracy of 84.62%. Just like Bacteroides Uniformis 1 by its own. Yet, we elevate the ROC/AUC by significant magnitude:

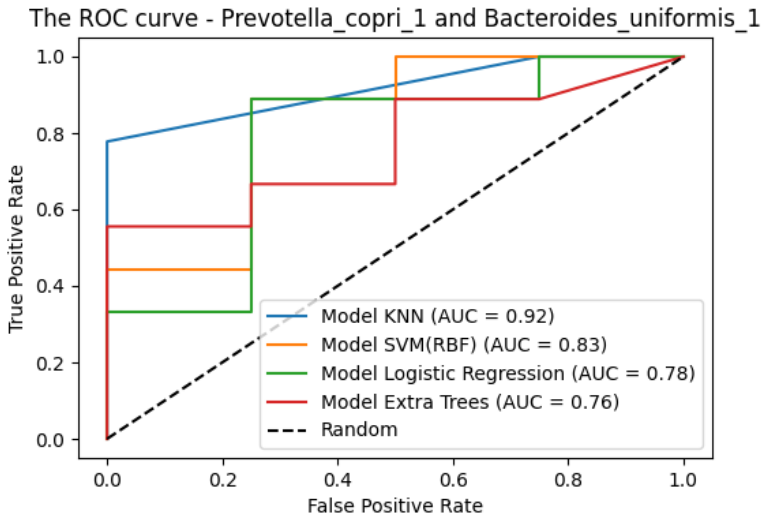


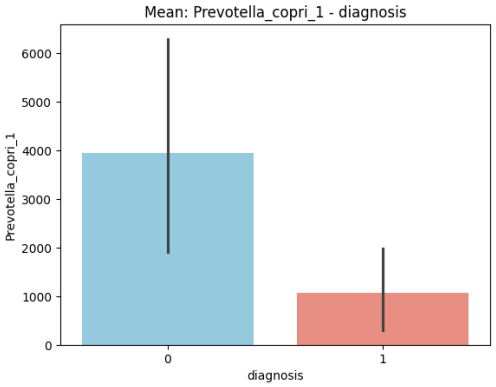
Figure 14. Graphic representation of the ROC curve and the respective AUC values of the Prevotella_copri_1 and Bacteroides_uniformis_1 OTUs predicting FM using different classification algorithms

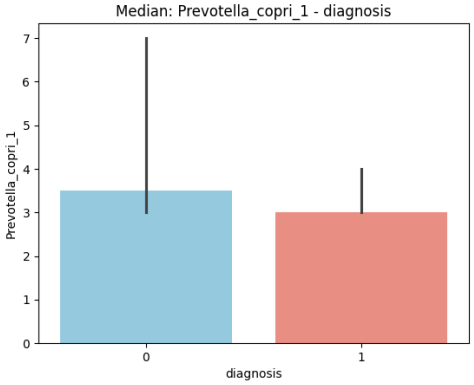
7.5. Bar graphs for the selected 3 best OTUs and Bacteroides uniformis 3

Below are presented bar graphs that represent a statistical estimate (mean on the left and median on the right) for each of the top three selected OTUs and for Bacteroides uniformis 3 found to be abundant in the control group (11). The red column represents the FM patients, and the blue column represents the control group.

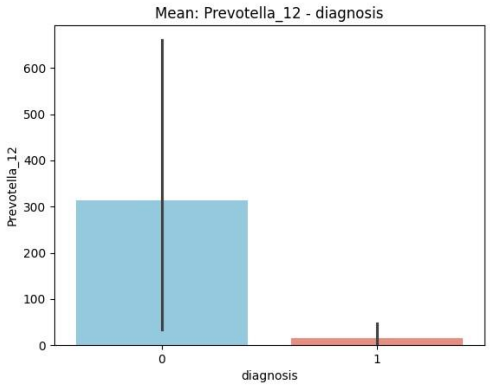
The black error bar indicates the uncertainty around this estimate.

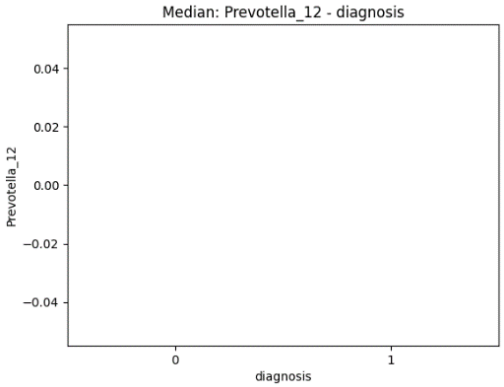
7.5.1. Prevotella copri 1



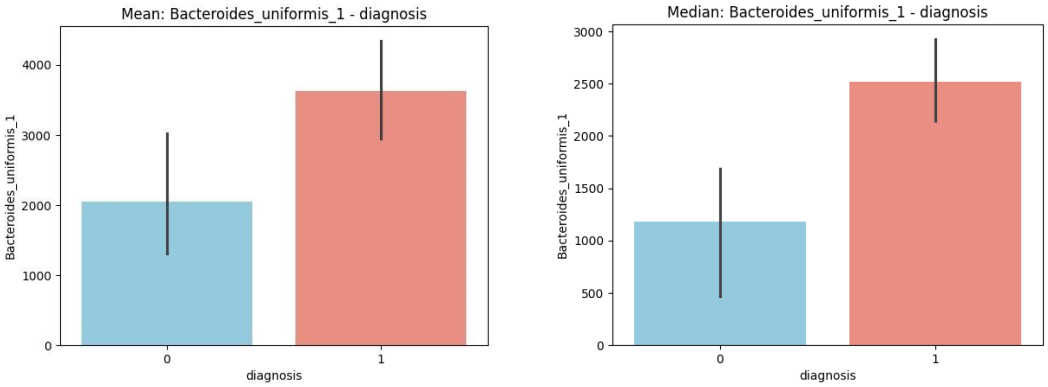


7.5.2. Prevotella 12





7.5.3. Bacteroides uniformis 1



7.5.4. Bacteroides uniformis 3

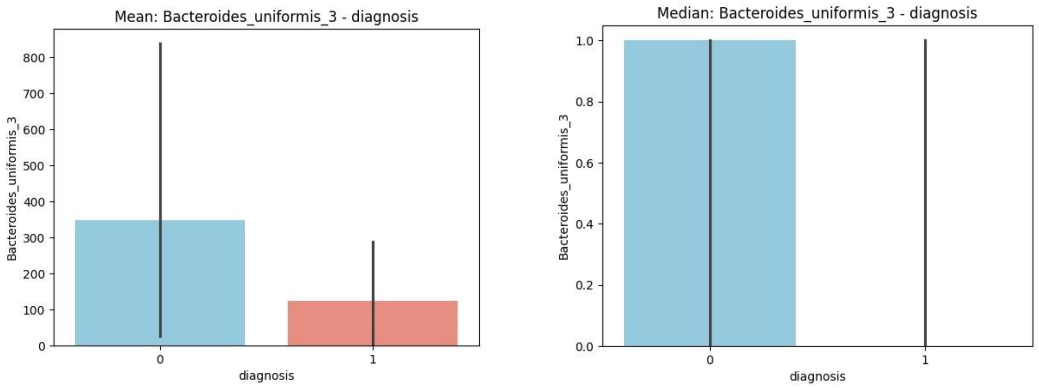


Figure 15. Statistical estimations for each of the top three selected OTUs

7.6. Bacteroides Uniformis Species

Bacteroides such as Bacteroides uniformis may play a role in alleviating obesity. A low amount of B. uniformis found in the gut of formula-fed infants was associated with a high risk of obesity (44). In addition, B. uniformis has been reported to improve immunological dysfunction and metabolic disorders, associated with intestinal dysbiosis in obese mice. Acute administration of this strain did not show any adverse effects (45).

Bacteroides Uniformis Species contains 3 OTUs: Bacteroides_uniformis_1, Bacteroides_uniformis_2 and Bacteroides_uniformis_3.

7.6.1. Bacteroides_uniformis_1, Bacteroides_uniformis_2 and Bacteroides_uniformis_3

The best accuracy achieved from diagnosis FM using only all three OTUs was 84.62% by using KNN algorithm (k=2, threshold=1), with ROC/AUC:

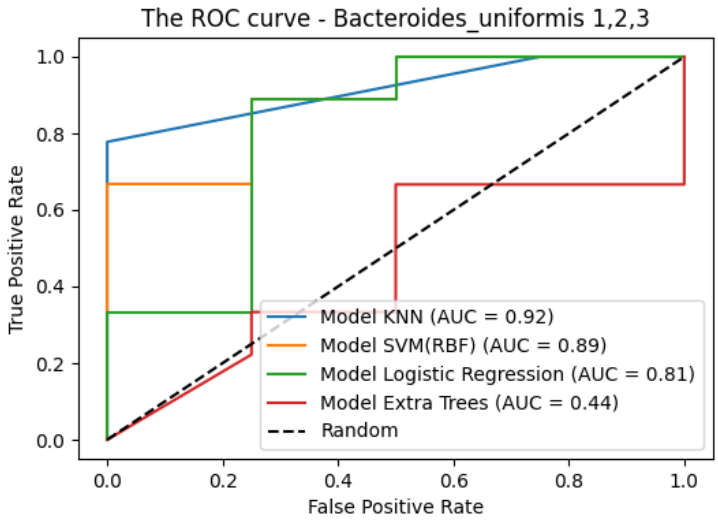


Figure 16. Graphic representation of the ROC curve and the respective AUC values of the Bacteroides_uniformis_1, Bacteroides_uniformis_2 and Bacteroides_uniformis_3 OTUs predicting FM using different classification algorithms

7.6.2. The sum of all OUTs from the species Bacteroides Uniformis

Diagnosis FM from one column created from the sum of all OUTs from the species Bacteroides Uniformis achieved best accuracy 84.62 % using CatBoost, and ROC AUC:

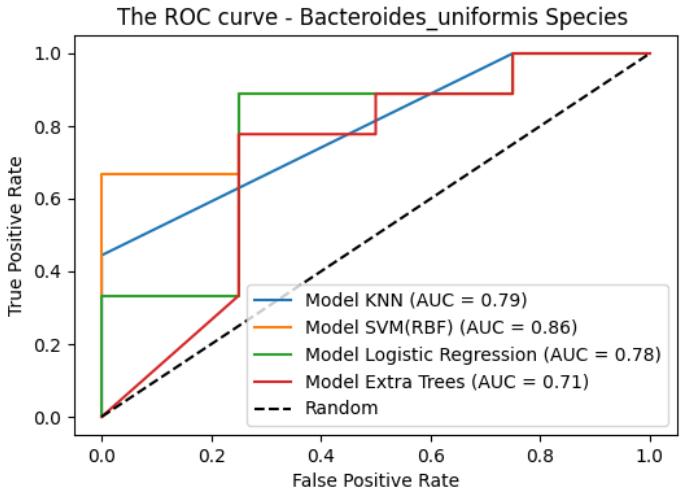


Figure 17. Graphic representation of the ROC curve and the respective AUC values of the Bacteroides_uniformis species as the sum of all its OTUs predicting FM using different classification algorithms

7.6.3. Bacteriodes_uniformis_3

In the previous study, it was reported that Bacteriodes_uniformis_3 was abundant in the control group compared to FM patients (11), and we also saw this using the bar graphs above. When diagnosed using it alone, the accuracy is only 69.23% using XGboost, CatBoost and SVM. And the ROC AUC results are:

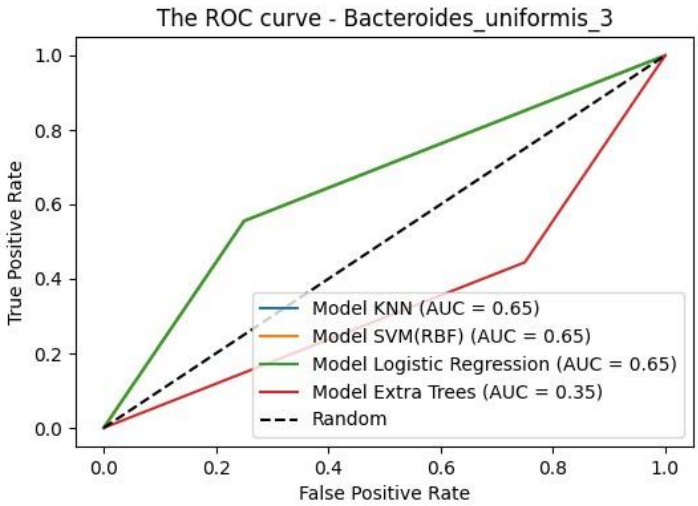


Figure 18. Graphic representation of the ROC curve and the respective AUC values of the Bacteroides_uniformis_3 OTU predicting FM using different classification algorithms

7.6.4. Bacteriodes_uniformis_1 and Bacteriodes_uniformis_3

While Bacteriodes_uniformis_3 did not give us impressive results compared to the interesting results we got from Bacteriodes_uniformis_1 in trying to diagnose FM, their combination gave significant results. The accuracy obtained is 84.62% using KNN (k=2, threshold =0.8). ROC AUC results:

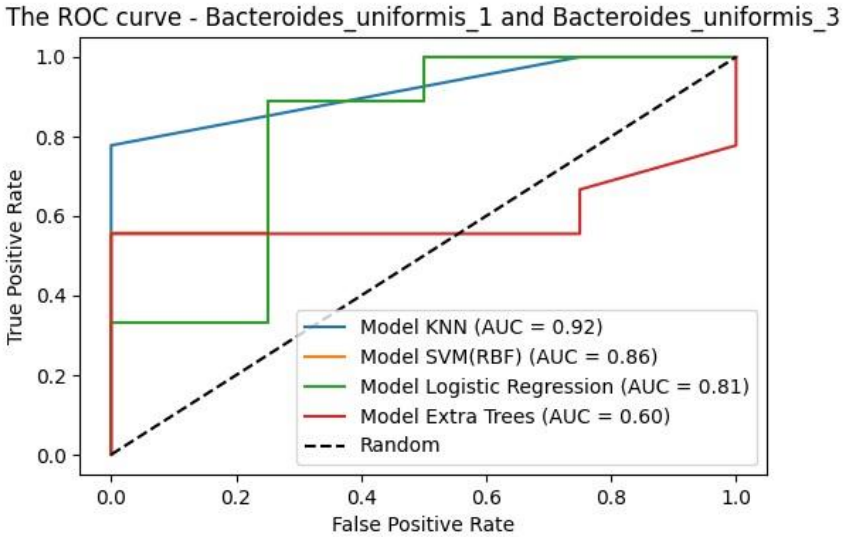


Figure 19. Graphic representation of the ROC curve and the respective AUC values of the Bacteroides_uniformis_1 and Bacteroides_uniformis_3 OTUs predicting FM using different classification algorithms

7.7. Various Renderings of The Dataset

In this section we will share the results accepted by emphasizing the different taxonomic levels. For each sample we summarized the OTUs by the respected taxonomic level and counted how many OTUs there are from this taxonomic level. Then, we used 'select K best' with *k=40* and tested the rendered dataset with the different classification algorithms.

The most significant results for each taxonomic level are:

Taxonomic Level	Algorithm	Accuracy
Species	SVM	84.62%
Genus	SVM	76.92%
Family	SVM	76.92%

8. Discussion

8.1. Comparing the results with a previous article

In the following table we will see a comparison of the tools and results between the previous study that created the dataset and studied it for the first time (11) and our study:

	Altered microbiome composition in individuals with fibromyalgia	**Our research**		
Feature selection algorithm	DESeq2	Select K best		Deductive choice of OTU 'Bacteroides Uniformis 1'
Number of picked features	72	12	3	1
Best accuracy	N/A	100% (using KNN)	84.62% (using both KNN and SVM)	84.62% (using KNN)
Best AUC using SVM	87.8%	94%	92%	82%
Best AUC using KNN	N/A	100%	92%	86%

8.2. KNN as a diagnostic tool using gut microbiome

In this study, we saw that the highest accuracy per day and the best AUC value were obtained using the KNN classification algorithm, therefore it is the most suitable for diagnosing fibromyalgia using the gut microbiome. It is possible that this result can also give an idea about using the KNN algorithm for trying to diagnose other diseases using gut microbiome.

8.3. Bacteriodes_uniformis_1

Bacteriodes_uniformis_1 itself gave us 84.62% accuracy using KNN and high AUC values of 0.86 using SVM.

Bacteriodes_uniformis_1 is abundant in fibromyalgia patients compared to the control group. In the case of Bacteriodes_uniformis_3 the situation is the opposite, it is abundant in the control group compared to the patients.

It is interesting to see that the best results for the diagnosis attempt by Bacteriodes_uniformis_1 and Bacteriodes_uniformis_3 are obtained by the KNN algorithm, the accuracy is 84.62% and the AUC is 0.92, just like the use of the three

OUTs of this species and just like the use of Bacteroides uniformis 1 together with Prevotella copri 1 which also It is more abundant in the control group.

8.4. The selected 12 OTUs and their taxonomic levels

The search for diagnostic methods of FM is one that has been going on for an extended period, in which various biomarkers have been explored. Our findings strongly indicate the possible relationship between gut microbiome and FM diagnosis. Furthermore, it appears that employing feature selection methods and narrowing down to specific OTUs significantly enhances the accuracy of the diagnosis.

Focusing on the selected OTUs and their taxonomic levels might also raise interest regarding specific microbiome families. Showcasing the 12 OTUs selected by the select K best features algorithm and their taxonomic levels bring to light that there are 2 families that stand out more than others – Prevotellaceae and Bacteroidaceae. They occupy two-thirds of the total number of OUT families. The selection of the top 12 OTUs is made in order of importance, which means that the top 4 are divided half and half between the 2 significant families.

OTU	Family	Genus	Species
Prevotella_copri_1	Prevotellaceae	Prevotella	Prevotella_copri
Prevotella_12	Prevotellaceae	Prevotella	Prevotella
Bacteroides_uniformis_1	Bacteroidaceae	Bacteroides	Bacteroides_uniformis
Bacteroides_dorei_1	Bacteroidaceae	Bacteroides	Bacteroides_dorei
Alistipes_finegoldii_2	Rikenellaceae	Alistipes	Alistipes_finegoldii
Bacteroides_MS_3	Bacteroidaceae	Bacteroides	Bacteroides_MS
Parabacteroides_merdae_3	Tannerellaceae	Parabacteroides	Parabacteroides_merdae
Ruminococcaceae_MG_1	Ruminococcaceae	Ruminococcaceae_MG	Ruminococcaceae_MG
Alloprevotella_1	Prevotellaceae	Alloprevotella	Alloprevotella
Bacteroides_3	Bacteroidaceae	Bacteroides	Bacteroides
Prevotella_4	Prevotellaceae	Prevotella	Prevotella
Akkermansia_muciniphila_1	Akkermansiaceae	Akkermansia	Akkermansia_muciniphila

In summary, examining the association between the gut-microbiome and FM indicates clear evidence of potential use for FM diagnostic. In addition, we believe that the relationship between specific OTUs/families raises the probable possibility of different treatment approaches, deepening in the biological mechanisms relevant to the symptoms of FM might lead to a cure alongside better diagnosis.

9. Conclusion and Future Work

9.1. Conclusion

Fibromyalgia (FM) is a common medical condition that is more prevalent in women, defined primarily by the presence of a chronic pain disorder. FM has a considerable impact on people who suffer from symptoms which, in addition to chronic pain, include fatigue and sleep disturbances (1). As of today, and although the prevalence of the disease is between 2% and 8% of the population (2), there is no objective way to determine if a person suffers from FM.

The diagnostic process of FM includes identifying symptoms, ruling out similar diseases, and applying the American College of Rheumatology (ACR) classification criteria, which is a subjective questionnaire (46).

A previous groundbreaking study on the relationship between gut microbiome and FM collected a dataset including gut microbiome samples from women (FM patients and controls), and by using the SVM machine learning algorithm and selecting 72 specific Operational Taxonomic Units (OTUs) achieved an AUC of 87.8% (11).

In our study, we processed the aforementioned data set, used different types of classification algorithms and other machine learning tools in order to achieve our main goal - to try to improve the ability to distinguish between FM patients and the control group.

In addition to this main goal, we tried to achieve several goals:

1. To compare the results between the different classification algorithms in order to conclude which is the most appropriate algorithm for diagnosing fibromyalgia using the gut microbiome.
2. To find the connection between microbiome and fibromyalgia.
3. Select the smallest number of most significant OTUs out of 1620 OTUs.

Finally, the results exceeded our expectations, by using the Select K Best algorithm to identify the 12 most influential OTUs associated with FM. It contributed greatly to our research, achieving 100% accuracy and 100% Area Under the Curve (AUC) with KNN algorithm and 92% AUC with SVM.

Using only two OTUs, Prevotella copri 1 and Bacteroides uniformis 1, we obtained an AUC of 92% using KNN and a score of 84.62%. We also found that using two OTUs, Bacteroides_uniformis_1 and Bacteroides_uniformis_3, both from Bacteroides uniformis species, obtained the same percentages.

In addition, we discovered that one OTU, Bacteroides_uniformis_1, is very significant in the diagnosis of fibromyalgia, and we saw this in 2 different ways - using ROC AUC and using a t-test with a p-value that shows a significant difference between two groups (FM vs Control).

These results offer promising avenues for understanding the pathophysiology of FM, developing diagnostic aids, and investigating new treatment methods.

9.2.Future Work

This study is the tip of the iceberg in researching the relationship between the microbiome and fibromyalgia.

Following on from this research, we offer future research possibilities:

1. We suggest that the medical community consider examining the relationship between the 12 OTUs selected for FM in more depth and in particular the relationship of the disease with Bacteroides_uniformis_1, since it presents unequivocal results for diagnosis and enables this through simple laboratory tests (stool samples) that are significantly cheaper than other solutions presented, such as scans fMRI.
2. Investigate what is affected by Bacteroides_uniformis_1 and not only as part of the Bacteroides Uniformis species.
3. In this study we used a dataset that contains 125 samples, we suggest trying to reproduce the results on a larger dataset.
4. It is worthwhile to check correlations and a relationship between the 12 OTUs we found that provide an accurate diagnosis for FM and blood tests (for example iron), BMI and more.
   This way we can try to understand if it is possible to improve their health status through diet.
5. Try to cure fibromyalgia by changing the microbiome based on our results. Change of 12 OTUs, 3 OTUs and 1 OTU.

10. Reference

(1) Kodner, Charles. 'Common Questions about the Diagnosis and Management of Fibromyalgia'. American Family Physician, vol. 91, no. 7, Apr. 2015, pp. 472–78. https://pubmed.ncbi.nlm.nih.gov/25884747/

(2) Clauw, Daniel J. 'Fibromyalgia: A Clinical Review'. JAMA, vol. 311, no. 15, Apr. 2014, pp. 1547–55. PubMed, https://doi.org/10.1001/jama.2014.3266.

(3) Bhargava, Juhi, and John A. Hurley. 'Fibromyalgia'. StatPearls, StatPearls Publishing, 2024. PubMed, http://www.ncbi.nlm.nih.gov/books/NBK540974/.

(4) Wolfe, Frederick, and Winfried Häuser. 'Fibromyalgia Diagnosis and Diagnostic Criteria'. Annals of Medicine, vol. 43, no. 7, Nov. 2011, pp. 495–502. DOI.org (Crossref), https://doi.org/10.3109/07853890.2011.595734.

(5) Wolfe, F., et al. 'The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee'. Arthritis and Rheumatism, vol. 33, no. 2, Feb. 1990, pp. 160–72. PubMed, https://doi.org/10.1002/art.1780330203.

(6) Wolfe, Frederick, et al. 'The American College of Rheumatology Preliminary Diagnostic Criteria for Fibromyalgia and Measurement of Symptom Severity'. Arthritis Care & Research, vol. 62, no. 5, May 2010, pp. 600–10. DOI.org (Crossref), https://doi.org/10.1002/acr.20140 .

(7) Ablin, Jacob N., et al. 'Mechanisms of Disease: Genetics of Fibromyalgia'. Nature Clinical Practice. Rheumatology, vol. 2, no. 12, Dec. 2006, pp. 671–78. PubMed, https://doi.org/10.1038/ncprheum0349 .

(8) Boyle Wheeler, Regina, 'How is Fibromyalgia Diagnosed?' (September 09, 2023), Medically Reviewed by Sabrina Felson, MD, WebMD, https://www.webmd.com/fibromyalgia/fibromyalgia-diagnosis-and-misdiagnosis

(9) Rossy, Lynn A., et al. 'A Meta-Analysis of Fibromyalgia Treatment Interventions'. Annals of Behavioral Medicine, vol. 21, no. 2, June 1999, pp. 180–91. Springer Link, https://doi.org/10.1007/BF02908299

(10) Thursby, Elizabeth, and Nathalie Juge. 'Introduction to the Human Gut Microbiota'. Biochemical Journal, vol. 474, no. 11, June 2017, pp. 1823–36. PubMed Central, https://doi.org/10.1042/BCJ20160510.

(11) Minerbi, Amir, et al. 'Altered Microbiome Composition in Individuals with Fibromyalgia'. Pain, vol. 160, no. 11, Nov. 2019, pp. 2589–602. PubMed, https://pubmed.ncbi.nlm.nih.gov/31219947/

(12) Rojo, David, et al. 'Exploring the Human Microbiome from Multiple Perspectives: Factors Altering Its Composition and Function'. FEMS Microbiology Reviews, vol. 41, no. 4, July 2017, pp. 453–78. DOI.org (Crossref), https://doi.org/10.1093/femsre/fuw046

(13) Clapp, Megan, et al. 'Gut Microbiota's Effect on Mental Health: The Gut-Brain Axis'. Clinics and Practice, vol. 7, no. 4, Sept. 2017, p. 987. www.mdpi.com, https://doi.org/10.4081/cp.2017.987.

(14) Cryan, John F., et al. 'The Microbiota-Gut-Brain Axis'. Physiological Reviews, vol. 99, no. 4, Oct. 2019, pp. 1877–2013. DOI.org (Crossref), https://doi.org/10.1152/physrev.00018.2018.

(15) Shreiner, Andrew B., et al. 'The Gut Microbiome in Health and in Disease': Current Opinion in Gastroenterology, vol. 31, no. 1, Jan. 2015, pp. 69–75. DOI.org (Crossref), https://doi.org/10.1097/MOG.0000000000000139.
(16) Vijay, Amrita, and Ana M. Valdes. 'Role of the Gut Microbiome in Chronic Diseases: A Narrative Review'. European Journal of Clinical Nutrition, vol. 76, no. 4, Apr. 2022, pp. 489–501. www.nature.com, https://doi.org/10.1038/s41430-021-00991-6 .
(17) Ogunrinola, Grace A., et al. 'The Human Microbiome and Its Impacts on Health'. International Journal of Microbiology, vol. 2020, June 2020, p. 8045646. PubMed Central, https://doi.org/10.1155/2020/8045646
(18) 'Educative Answers - Trusted Answers to Developer Questions'. Educative, https://www.educative.io/answers/supervised-vs-unsupervised-vs-reinforcement-learning.
(19) Toh, Christopher, and James P. Brody. 'Applications of Machine Learning in Healthcare'. Smart Manufacturing - When Artificial Intelligence Meets the Internet of Things, IntechOpen, 2021. www.intechopen.com, https://doi.org/10.5772/intechopen.92297.
(20) 'Using Machine Learning for Healthcare Challenges and Opportunities'. Informatics in Medicine Unlocked, vol. 30, Jan. 2022, p. 100924. www.sciencedirect.com, https://doi.org/10.1016/j.imu.2022.100924
(21) Habehh, Hafsa, and Suril Gohel. 'Machine Learning in Healthcare'. Current Genomics, vol. 22, no. 4, Dec. 2021, pp. 291–300. PubMed Central, https://doi.org/10.2174/1389202922666210705124359.
(22) H, Roshna S. 'K-Nearest Neighbors Algorithm'. Intuitive Tutorials, 7 Apr. 2023, https://intuitivetutorial.com/2023/04/07/k-nearest-neighbors-algorithm/ .
(23) Zhang, Shichao, et al. 'Learning k for kNN Classification'. ACM Transactions on Intelligent Systems and Technology, vol. 8, no. 3, May 2017, pp. 1–19. DOI.org (Crossref), https://doi.org/10.1145/2990508 .
(24) Jakkula, Vikramaditya. "Tutorial on support vector machine (svm)." School of EECS, Washington State University 37.2.5 (2006): 3.
(25) 'Support Vector Machine'. Wikipedia, 19 Dec. 2023. Wikipedia, https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=1190739318
(26) Chen, Tianqi, and Carlos Guestrin. 'XGBoost: A Scalable Tree Boosting System.' Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, pp. 785–94. arXiv.org, doi:10.1145/2939672.2939785. (https://arxiv.org/abs/1603.02754)
(27) CatBoost - State-of-the-Art Open-Source Gradient Boosting Library with Categorical Features Support. https://catboost.ai .
(28) 'Sklearn.Ensemble.ExtraTreesClassifier'. Scikit-Learn, https://scikit-learn/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html.
(29) 'Sklearn.Linear_model.LogisticRegression'. Scikit-Learn, https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html .
(30) Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. Data classification: Algorithms and applications, 37.

(31) D, Kavya. ‘Optimizing Performance: SelectKBest for Efficient Feature Selection in Machine Learning’. Medium, 16 Feb. 2023, https://medium.com/@Kavya2099/optimizing-performance-selectkbest-for-efficient-feature-selection-in-machine-learning-3b635905ed48.
(32) ‘Classification: ROC Curve and AUC | Machine Learning’. Google for Developers, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
(33) ‘Classification: ROC Curve and AUC | Machine Learning’. Google for Developers, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc. Accessed 3 Jan. 2024.
(34) Hackshaw, Kevin V. ‘The Search for Biomarkers in Fibromyalgia’. Diagnostics, vol. 11, no. 2, Feb. 2021, p. 156. www.mdpi.com, https://doi.org/10.3390/diagnostics11020156.
(35) Uygur-Kucukseymen, Elif, et al. ‘Decreased Neural Inhibitory State in Fibromyalgia Pain: A Cross-Sectional Study’. Neurophysiologie Clinique, vol. 50, no. 4, Sept. 2020, pp. 279–88. ScienceDirect, https://doi.org/10.1016/j.neucli.2020.06.002
(36) Napadow, Vitaly, and Richard E. Harris. ‘What Has Functional Connectivity and Chemical Neuroimaging in Fibromyalgia Taught Us about the Mechanisms and Management of `centralized’ Pain?’ Arthritis Research & Therapy, vol. 16, no. 4, Oct. 2014, p. 425. DOI.org (Crossref), https://doi.org/10.1186/s13075-014-0425-0.
(37) Hackshaw, Kevin V., et al. ‘A Bloodspot-Based Diagnostic Test for Fibromyalgia Syndrome and Related Disorders’. Analyst, vol. 138, no. 16, July 2013, pp. 4453–62. pubs.rsc.org, https://doi.org/10.1039/C3AN36615D
(38) Erdrich, Sharon, et al. ‘Determining the Association between Fibromyalgia, the Gut Microbiome and Its Biomarkers: A Systematic Review’. BMC Musculoskeletal Disorders, vol. 21, no. 1, Mar. 2020, p. 181. Springer Link, https://doi.org/10.1186/s12891-020-03201-9 .
(39) Clos-Garcia, Marc, et al. ‘Gut Microbiome and Serum Metabolome Analyses Identify Molecular Biomarkers and Altered Glutamate Metabolism in Fibromyalgia’. EBioMedicine, vol. 46, Aug. 2019, pp. 499–511. DOI.org (Crossref), https://doi.org/10.1016/j.ebiom.2019.07.031
(40) Moloney, Rachel D., et al. ‘Stress and the Microbiota–Gut–Brain Axis in Visceral Pain: Relevance to Irritable Bowel Syndrome’. CNS Neuroscience & Therapeutics, vol. 22, no. 2, Feb. 2016, pp. 102–17. DOI.org (Crossref), https://doi.org/10.1111/cns.12490.
(41) Foster, Jane A., and Karen-Anne McVey Neufeld. ‘Gut-Brain Axis: How the Microbiome Influences Anxiety and Depression’. Trends in Neurosciences, vol. 36, no. 5, May 2013, pp. 305–12. PubMed, https://doi.org/10.1016/j.tins.2013.01.005.
(42) Wang, Zhe, et al. ‘The Microbiota-Gut-Brain Axis in Sleep Disorders’. Sleep Medicine Reviews, vol. 65, Oct. 2022, p. 101691. ScienceDirect, https://doi.org/10.1016/j.smrv.2022.101691.
(43) XGBoost Parameters — Xgboost 2.0.3 Documentation. https://xgboost.readthedocs.io/en/stable/parameter.html
(44) Bacteroides’. Wikipedia, 23 Jan. 2024. Wikipedia, https://en.wikipedia.org/w/index.php?title=Bacteroides&oldid=1198255416.

(45) Bacteroides Uniformis - an Overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/immunology-and-microbiology/bacteroides-uniformis#:~:text=phagocytosis%20%5B52%5D.-,B.,not%20show%20any%20adverse%20effects. Accessed 16 Feb. 2024

(46) Goldenberg, Don L. 'Diagnosis and Differential Diagnosis of Fibromyalgia'. The American Journal of Medicine, vol. 122, no. 12, Supplement, Dec. 2009, pp. S14–21. ScienceDirect, https://doi.org/10.1016/j.amjmed.2009.09.007 .

תקציר

פיברומיאלגיה (FM), בעיה רפואית נרחבת המאופיינת בכאב כרוני, קשיים קוגניטיביים, עייפות והפרעות שינה. מאתגר מאוד לאבחן פיברומיאלגיה וקשה לטפל בה. תהליך האבחון של FM כרוך בזיהוי תסמינים, שלילת מחלות דומות ומענה על שאלון סובייקטיבי. נכון להיום, אין בדיקה ישירה ואובייקטיבית.

במחקר שלנו, ניסינו למצוא את הקשר בין מיקרוביום המעי ופיברומיאלגיה על ידי שימוש בסוגים שונים של אלגוריתמי סיווג וכלי למידת מכונה אחרים על מערך נתונים הכולל דגימות מיקרוביום מעי מנשים עם FM וקבוצת ביקורת.

השתמשנו באלגוריתם Select K Best כדי לזהות את 12 חיידקי המיקרוביום המשפיעים ביותר בהקשר של פיברומיאלגיה. צימצום זה תרם רבות למחקר שלנו, ובאמצעות אלגוריתם KNN הצלחנו להשיג 100% דיוק ו-100% שטח מתחת לעקומה (AUC), ועם SVM השגנו 92% AUC.

בנוסף, מצאנו כי Bacteroides_uniformis_1 הוא חיידק מיקרוביום משמעותי ביותר באבחון של פיברומיאלגיה.

תוצאות מחקר זה מציעות דרכים מבטיחות להבנת הפתופיזיולוגיה של FM, פיתוח עזרי אבחון וחקירת שיטות טיפול חדשות.

המכללה האקדמית תל-אביב

בית הספר למדעי החשב

חיזוי פיברומיאלגיה ממיקרוביום המעי באמצעות למידת מכונה

חיבור זה הוגש כחלק מהדרישות לקבלת התואר "מוסמך " – M.Sc.

במכללה האקדמית תל-אביב

על ידי

מור מרים צ'קו

העבודה הוכנה בהדרכתם של

פרופ' עדי שרייבמן ודר' דורית שוויקי